Injury & Disability Schemes Seminar
Resilience and Recovery

17 – 19 October 2021

Lost cause: Getting at causation in our datasets

Prepared by Hugh Miller

Presented to the Actuaries Institute
2021 Injury and Disability Schemes Seminar
17-19 October 2021

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Abstract

We live in a world where we often have incredibly good data, but limited ability to use it to directly answer the questions that we most care about. Can we find the impact of legal representation if more severe claims are more likely to be represented? What is the impact of an education support payment if it is targeted at disadvantaged students? What is the fiscal saving from an employment placement if some people would have found a job anyway? Answering such questions requires a deeper understanding of the data than straight descriptive statistics.

Traditional predictive modelling will assign effects, but many of these will be correlative rather than causative. However, significant progress has been made in types of causal modelling which aim to get at actual effects. While formal experiments such as randomised controlled trials remain the gold standard for many applications, other types of causal estimates can be made from quasi-experimental designs such as regression discontinuity, instrumental variable, stepped wedge designs or propensity scoring. This paper introduces quasi-experimental designs and provides some examples of how we have applied them in actuarial contexts. Examples will draw on quasi-experimental evidence in injury schemes, welfare policy and housing.

Being able to estimate the actual impacts of underlying factors significantly improves actuaries' ability to provide good advice to questions of most interest to decision-makers.

1 Introduction

1.1 Background

Many actuaries and statisticians have the lesson 'correlation is not causation' drilled into them throughout their education. In many cases this is obviously true. The spurious correlations website[1] takes delight identifying high-correlation patterns that clearly have no causal relation. Ayoub et al. (2021) found a strong positive correlation of FIFA soccer ranking and COVID-19 incidence rates and highlighted as a reminder to take care when considering other published research suggesting COVID-19 risk factors such as blood type, vitamin D levels and the BCG (tuberculosis) vaccine. More seriously, the recent lesson around Hydroxychloroquine as a COVID-19 treatment is instructive (see for example Jorge, 2021). A treatment with a course of hydroxychloroquine seemed positive based on early observational studies but was found to be ineffective through more careful subsequent work.

Actuaries have taken this lesson to heart. When we build statistical case estimates, we are careful not to assign a causal interpretation to factors. Just because we predict that claims with legal representation will have higher ultimate cost does not mean that lawyers are the 'cause' of high claims. Large actuarial models predicting how long-term welfare costs vary with education level stress that effect sizes are not necessarily causal (Greenfield et al., 2017).

And yet, causal questions are fundamental to many of the programs and schemes that actuaries advise on. We are asked how much liabilities will be affected by behavioural changes, or the degree to which a concurrent mental health affects recovery speed. These questions go beyond predictive modelling because it asks us to estimate a counterfactual (what would happen if…). An ability to address causation meaningfully is of great value in delivering advice.

More broadly, there appears to be increased appetite to talk about causation. For example, economics is one discipline that has seen significant growth in causal claims. In some cases this is through increased use of experimental and quasi-experimental techniques. However, sometimes it merely reflects increased confidence that there is something meaningful to be gained from observational data. To pick a recent example, Hope and Limberg (2020) look at the impact of major tax cuts for the rich across 18 OECD countries, finding strong evidence for increased income inequality but no evidence of higher growth or lower unemployment. While there are some quasi-experimental elements to the work (e.g. matching across countries), what is striking is the inclusion of causal claims from data that is ultimately observational. The increased willingness to address causal questions carries risks but is ultimately more satisfying when successful.

1.2 Purpose of this paper and further reading

The purpose of the paper is threefold:

1. To explore some general principals about causality and how we can draw causal conclusions from data.
2. To introduce some key quasi-experimental methods in causal modelling. These are techniques that attempt to estimate a treatment effect even when the underlying data was not designed as a formal randomised experiment.
3. To cover some recent examples of quasi-experimental work performed by the author in an actuarial context.

Section 2 focuses on the first item, articulating some key principles in causal thinking. Section 3 covers the second and third item in the list. Section 3.6 offers some final thoughts on the topic.

There is substantial literature on the topic of causation – our paper only scratches the surface of many topics. We point to some key books here for interested readers:

[1] https://www.tylervigen.com/spurious-correlations

- Imbens & Rubin (2015) and Rosenbaum (2010) are well-regarded textbooks with a statistical emphasis, from leaders in the field; in fact, the 'Rubin causal model' is one of the most common frameworks for thinking about causality through the lens of potential outcomes.
- For a greater emphasis on practical applications both Hernán & Robins (2020) and Morgan & Winship (2015) offer good introductions from an epidemiological and social science perspective respectively.
- Pearl & Mackenzie (2018) is an accessible book exploring the topic of causation. If focuses on much of Judea Pearl's personal research work in the area, with a strong emphasis on working with causal diagrams and the language of causality, illustrated with important examples through history (including two of those used in Section 2). Pearl also wrote a more technical introduction to causal inference and structural causal models (Pearl et al., 2016).

For a brief overview of quasi-experimental methods, of the types explored in Section 3, the paper by Kim & Steiner (2016) is also recommended.

1.3 Acknowledgements

We would like to acknowledge those that funded the work presented in our examples and gave permission for them to be used in this paper:

- The **NSW State Insurance Regulatory Authority** (SIRA), who commissioned and funded our work exploring outcomes for Minor Injury claims (see section 3.4)
- The **NSW Department of Communities of Justice** (DCJ), and FACSIAR within it, who commissioned and funded our evaluation work on student scholarships (see section 3.5)
- The **NZ Ministry of Social Development** (MSD), who commissioned and funded our annual reviews that included analysis on the 3k to Work program (see section 3.3).

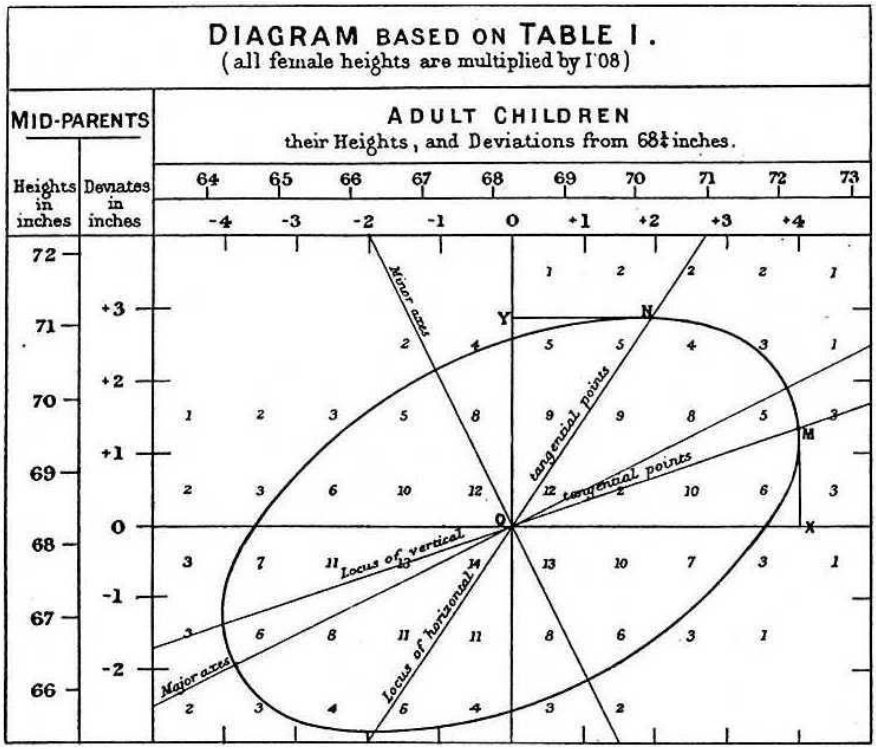
2 Some useful principles in causation

2.1 Principle 1: Causation requires a model that extends beyond the data

Except in special circumstances, inference around causality does not directly fall out of the data – the data must be interpreted with reference to an external causal model

We start with an example. Sir Francis Galton was a prodigious thinker in 19th Century England who was interested in questions around inheritance and the statistics to support it. One of his most famous exercises was comparing the heights of parents to their (adult) children. Galton found strong effects, albeit with a regression slope less than one (every inch of parental heights adds roughly 0.4 inches to the child's height).

Figure 1 – An extract of Galton's work comparing parent-child heights, including regression lines



While this study is often used to describe the discovery of regression to the mean, there is another causation-related insight here. Galton was originally interested in the causal, hereditary nature of height. Regression shows a clear relationship, however Galton observed that *a similar relationship is found if parents heights are estimated as a function of child's height*; a similarly-sized regression effect is found (every inch of child heights adds roughly 0.4 inches to the parent's height).

This demonstrates a deeper truth regarding much statistical modelling work. Correlations and regressions do not automatically give a direction of causation. In the example it is easy to argue that the direction of causation is from parent to child; basic common sense, plus a rudimentary understanding of genetics provides this. But the principle we are flagging is that this argument must be made outside the statistical analysis. A broader 'model of causation' is needed.

With the possible exception of randomised experimentation (Principle 2 below), all our work requires some form of broader model, explicit or implicit, for how causality is reflected in the data. This may appear daunting and subjective. But there is also significant value in being able to articulate a causal mechanism and sensibly discuss issues such as potential confounders.

A broader model needs to consider questions such as whether the data will be subject to other effects that impact the observed impact. Even in the example above more work is needed to isolate the impact of genetics on height. For instance, nutrition is a potential confounder (a parent with poor childhood nutrition is more likely to be shorter and also more likely to have a child with poor nutrition growing up).

Observations like this can drive further research, or enable better controls to be applied to the regression.

There are different ways that a causal model can be articulated. Pearl & Mackenzie (2018) argue for the use of 'causal diagrams'. These are directed graphs – graphs of measures with joining arrows indicating causation. This is a simple and intuitive way of thinking about the causal mechanisms, particularly when there are multiple measures and confounders to considered. Many economists still lay out causal assumptions as equations; but since equations can be rearranged, the reader must be aware of the implicit assumptions about which terms are causal to others.

2.2 Principle 2: Always randomise when practical

Inferring cause in the absence of a randomised experiment is possible, as later parts of this paper demonstrate. However, randomisation, and specifically the randomised controlled trials (RCT), is the most powerful tool in the statistical toolbox for isolating cause. In many fields, such as drug testing, RCTs remain the gold standard of evidence needed for approving a new intervention.

In practice this means when a new intervention is piloted, some explicit randomisation step is needed. A (literal or virtual) coin could be flipped before someone is streamed into a new claims management program, or new prevention program.

One objection to RCTs is fairness – why should some people miss out on a new service due to luck? However, this argument seems shallow considering that the most common area for RCTs currently are medical interventions; in some trials people will die because of 'bad luck' of being in the wrong treatment group, but the downsides are more than offset by future lives saved by using a proven treatment (or avoiding an ineffective one). Equally, it may be 'unfair' to subject everyone to a new program or treatment if it delivers no benefits.

A second objection is the difficulty of embedding randomisation in day-to-day operations. Again, this seems to run counter to some examples. One prominent recent example is the RECOVERY Trial, led by Prof. Peter Horby and Prof. Martin Landray in the UK. This study produced some vital results early in the pandemic, such as the inefficacy of hydroxychloroquine (RECOVERY Collaborative Group, 2020), and the efficacy of dexamethasone (Horby et al., 2021); the latter discovery is estimated to have saved over a million lives. Patients could voluntarily enrol in the trial at hospital admission with a COVID-19 diagnosis and would then be allocated to 'standard treatment' or 'standard treatment with drug'. Four drugs were added initially and more added subsequently. What is notable is that this experiment was quickly rolled out across over 100 hospitals (enrolling 10,000 people within 2 months) despite the demand on the healthcare system due to COVID-19. A simple design combined with an online platform for easy registration and randomisation was enough to quickly build an evidence base.

In practice, there are some nuances that means RCTs do not instantly solve all causal issues:

- **Randomisation works best when there is a single intervention.** In medical research this is often the use of a specific drug, or treatment. This creates an easy link between impact and treatment. In broader context, such as an injury scheme, *multiple changes are often made at once*. A new triage model might be applied with a new case management structure and new training or IT. If we use an RCT for such an intervention the measured effect represents the combined impact of several changes, potentially without the ability to separate out contributions. Further, if there is no effect, it might be because one part of the changes was ineffective, even if other parts have value. A deeper evaluation (for example, a mixed methods approach that assesses the implementation of individual steps) might be more relevant, and the randomisation less crucial.
- **RCT results do not generalise easily.** By simplifying and controlling the causal estimation process, RCTs define impact under specific conditions which will not always map cleanly to the real world. A medical trial applied to a specific group of people (for example, adults without an existing disease or disorder, such as the AstraZeneca AZD1222 vaccine Phase III trial with a four-week interval) will not provide evidence beyond the testing regime (children, those with existing conditions, longer gap between doses, etc). To cover such heterogeneity an RCT would have to be prohibitively large.

- **RCTs often use narrow outcome measures.** RCTs also tend to have a very clear metric (for example, has a person returned to work at 4 weeks?) to enable comparability. It will not generally use more detailed analysis to estimate tailored improvement (for example, change in RTW for people with specific conditions associated with longer time off work).
- **Ethical considerations.** There are many subject areas that are not feasible for RCTs. For instance, an RCT that tests the impact of smoking on health by requiring people to smoke is clearly unethical.

2.3 Principle 2b: No one actually does randomised controlled trials

Despite the optimism of Principle 2, randomised designs still represent a minority in the development of new programs outside medical research. There certainly are some contemporary Australian examples; to pick a few:

- **Justice** – The Youth on Track program in NSW is subject to a RCT managed by BOCSAR (Trimboli, 2019)
- **Child protection** – the Resilient Families Service run by the Benevolent Society was subject to RCT (Leahy et al., 2020)
- **Road safety** – the role of telematic feedback for young drivers was tested by RCT (SIRA, 2019)

While we believe RCTs can and should be used more frequently outside medicine, we accept the reality that often impact must be gauged in real-world settings where an RCT is not used. This means there is still an important role in observational studies and quasi-experimental methods.

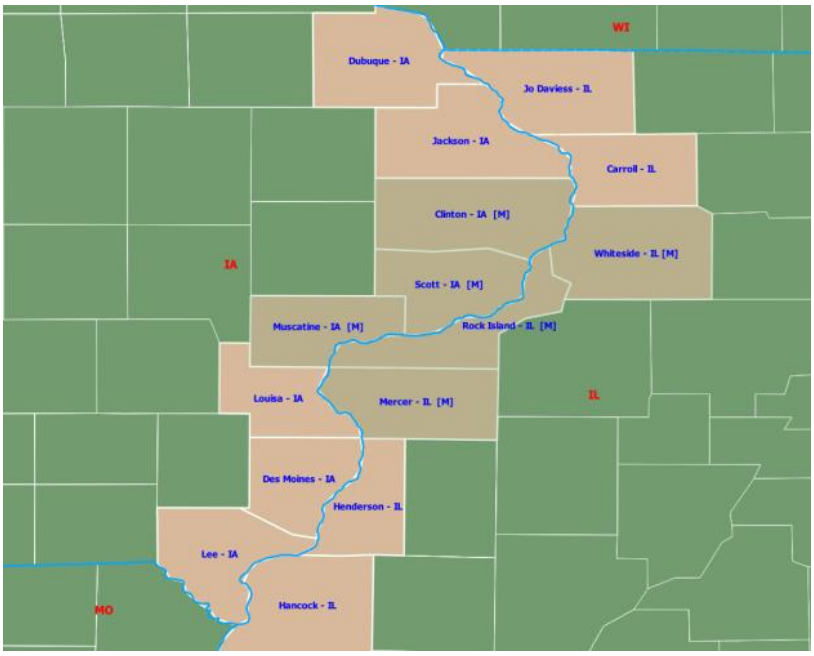
2.4 Principle 3: Natural experiments do occur

In the absence of formal randomised trials, identifying natural experiments, where treatment and control groups are created by specific sets of circumstances, is often a viable alternative.

The most important lesson around natural experiments is learning to recognise when they occur. Identification allows a specific analysis and inference to be made.

As an example, Haynes et al. (2020) recognised that differences in American state-level COVID restrictions allow testing of the impact of stay-at-home orders. Early in the pandemic, they identified neighbouring counties for three state borders that could be considered natural experiments and then compared case number growth rates. While there are undoubtably limitations (for example, different testing rates across states) and represents just one of many studies done on the effect of COVID-19 interventions, it gives explicit estimates of the impact of shelter-in-place on lowering transmission.

Figure 2 – Neighbouring counties on the Iowa-Illinois border used by Haynes et al. (2020). Illinois introduced a shelter-in-place order in March 2020, whereas Iowa did not.



2.5 Principle 4: You can (probably) infer something from straight observational data

Often large observational datasets show significant effects, and it is natural to ask whether it is reasonable to ascribe a cause. The set of 'Bradford Hill criteria' is one common reference point for answering this question. These were set out by the statistician Sir Austin Bradford Hill in 1965, and reflects his thinking on prior work, including the first work done linking smoking to lung cancer in 1950.

The criteria are not without controversy, and certainly debateable how many criterion need to be met before a causal interpretation is plausible. In many cases best practice is to confirm the results with prospective observational trials. We have reproduced the version of the criteria developed by Jeremy Howick and colleagues (see Howick et al., 2009, or Spiegelhalter, 2019).

Table 1 – The Bradford Hill criterion for inferring cause from observational data

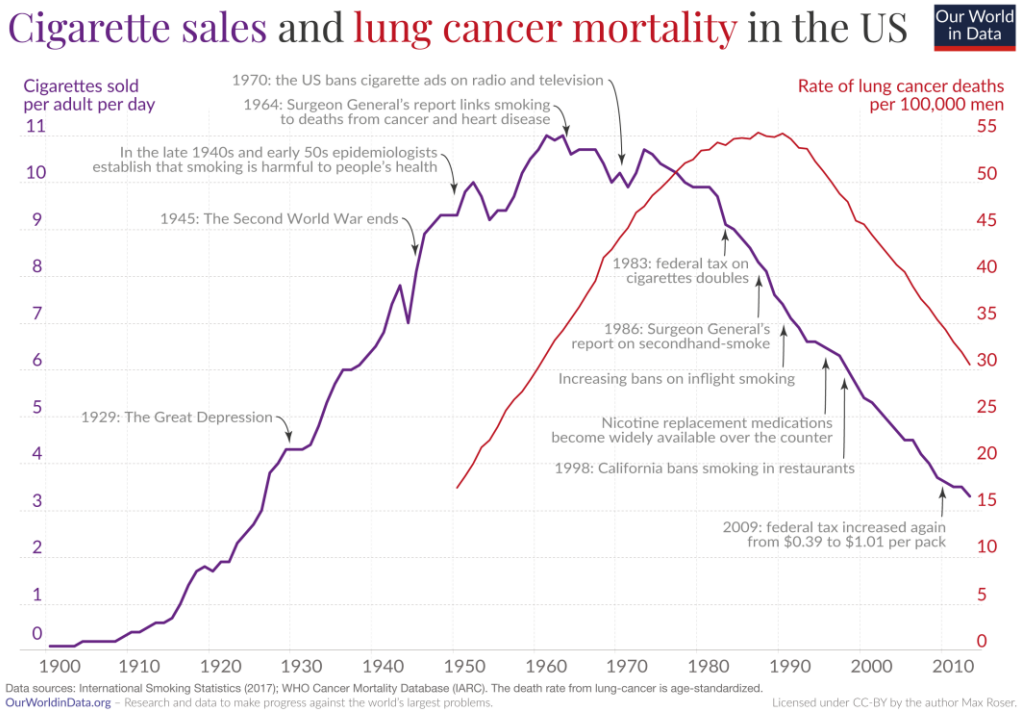
Direct evidence	1. The size of the effect is so large that it cannot be explained by plausible confounding. 2. There is appropriate temporal and/or spatial proximity, in that cause precedes effect and effect occurs after a plausible interval, and/or cause occurs at the same site as the effect. 3. Dose responsiveness and reversibility: the effect increases as the exposure increases, and the evidence is even stronger if the effect reduces upon reduction of the dose.
Mechanistic evidence	4. There is a plausible mechanism of action, which could be biological, chemical, or mechanical, with external evidence for a 'causal chain'.
Parallel evidence	5. The effect fits with what is known already. 6. The effect is found when the study is replicated. 7. The effect is found in similar, but not identical, studies

The linking of smoking to lung cancer in the mid-20th century is a key example of such observational work (covered in both Spiegelhalter, 2019 and Pearl & Mackenzie 2018). Since cancer occurrence is not

immediate (often years later) and not all smokers get cancer, the observational evidence only emerged gradually over time. Randomised controlled trials were not practical for ethical reasons.

However, once evidence was assembled the effect size was overwhelmingly large; lung cancers had increased from historical (low) rates by an order of magnitude, meaning that something had definitely changed. Trends followed that for smoking rates, albeit with a lag (see Figure 3). Smoking had natural mechanistic evidence also, since it smoke has direct contact with the lungs.

Figure 3 – Neighbouring counties on the Iowa-Illinois border used by Haynes et al. (2020). Illinois introduced a shelter-in-place order in March 2020, whereas Iowa did not.



Source: Our world in data, https://ourworldindata.org/smoking-big-problem-in-brief

Even with this substantial evidence there remained controversy. Ronald Fisher, one of the giants of early 20th century statistics, including experiment design, did not accept the results as they emerged in the 1950s; he argued that this did not rule out confounding, such as the existence of a 'smoking gene' that both increased the propensity to smoke and increased the risk of lung cancer.

More generally, when working with observational data, the method of allocation between 'treatment' and comparison groups is one of the most important parts in gauging feasibility of estimating impact. If we know there's an element of randomisation than this can be leveraged. If we know there are formal rules for allocation that this can be used. If we have a dataset where we have all the relevant criteria affecting allocation, then estimation is still possible. However, if there are subjective allocation effects that are invisible to the analyst, then this will create difficulties.

3 Quasi-experimental methods for causal inference

3.1 Introduction

While Section 2 focuses on more general observations around causal inference, it is important to recognise that there are a collection of specific techniques that are often used to identify causal effects in specific circumstances. They are often termed 'quasi-experimental' design, in that they are mimic the types of estimates you might achieve with a formal experiment, even if that data is not generated from one. We aim to give the intuition behind each approach; interested readers are referred to the references in Section **Error! Reference source not found.**.

We use some mathematical notation throughout to be concrete about our setup. We use the terminology 'treatment' for the effect that we want to understand causally. This is denoted T; it will often be a binary treatment (the impact of $T = 1$ versus no treatment $T = 0$), but could be continuous to represent an amount of treatment. The 'response', Y,the variable that is measured for change. Other variables $X = (X_1, \dots X_p)$ are known about observations and may be needed for control purposes.

We use the term control group as a cohort that is used for comparison in measuring the treatment effect. The term is used for familiarity, although most authors acknowledge that 'quasi-control' or 'comparison' group is more accurate and avoids potential confusion that a formal experiment has taken place.

3.2 Regression that controls for confounders

Description

Perhaps the simplest approach (albeit one with theoretical challenges to consider) is to model the treatment along with other variables X and to estimate the effect size of the treatment in the model (e.g. as a regression parameter). A predictive model is built targeting the response:

$$y \approx f(X, T)$$

Then the average treatment effect can be estimated directly over the dataset by computing the average value of $f(X, 1) - f(X, 0)$ for each observation.

The biggest drawback of the approach is that it assumes that the observed treatment is handled as an experimental variable. In many cases selection effects and other confounding factors will make this a heroic assumption. Thus, it requires a strong assumption that requires judgement on how the data is generated.

In particular, the setup assumes that controlling for X by incorporating them in the model is enough to control for all the major confounders that may affect a naïve relationship between treatment and response.

Example – airline competitor price elasticity

As a brief example, consider estimation of the 'competitor price effect' in the airline industry. This is the degree to which a growing price gap between your price and a competitor price affects demand for your tickets (as people switch to the cheaper fare).

The potential information to estimate this effect is large; a flight is typically on sale for almost a year before departure, so every day we can record the price offered, the competitor premium as well as the number of ticket sales. This will often vary over time for a flight as airlines 'step up' their price as a plane fills up. Extending this to all flights on a route over a period creates hundreds of thousands of observations for estimation. Furthermore, a range of control variables can be added:

- Number of days till departure
- Time of flight and day of week

- Seasonal factors
- Prices of other flights on the same route (self-competition).

While the data is primarily observational, the volume of data points combined with enough price variation over time means that it is probably reasonable to regard this as a natural experiment, with some confidence of robust price elasticity effects.

Potential confounders remain though; for instance, specific spikes in demand (like for a popular football match) will appear in the data as incidences where demand remains strong despite the price premium offered.

Expansion to heterogenous effects using machine learning

In situations where the dataset is relatively large and firm effect sizes can be estimated, a natural extension is exploring heterogenous impacts; which groups see the largest or smallest treatment responses. Machine learning approaches offer an attractive solution as they can simultaneously model both control variables and variations in the treatment effect. For example, causal random forests[2] have a similar structure to classical random forests except that instead of using decision tree splits to maximise the difference in the response variable, it maximises the difference in treatment effect via a gradient function. Thus it can get at the question of 'what works for who'.

Such approaches are common for online digital applications (for example digital advertising effectiveness), where randomisation is possible and sample sizes tend to be large.

See, for instance, Knaus et al. (2021) for further discussion and a comparison of performance of different machine learning models for causal learning.

3.3 Matching, including propensity matching

Description

While the approach in section 3.2 controls for the effects of X through regression, it is also possible to control through matching techniques. The idea is intuitive; if X is judged a reasonable set of control variables we have a set of treatment observations ($T = 1$) and a larger set of untreated (control) observations ($T = 0$), then choose a subset of the untreated observations so that the distribution of control variables match the treatment group. For example, choose a subset so that age distribution in the control subgroup matches that of the treated subgroup.

This process of matching becomes more difficult as the number of control variables grow. Not only are there more variables to keep track of, but also their potential correlations. **Propensity matching** (Rosenbaum & Rubin, 1983) offers an elegant alternative in this situation. The first step is to build a model predicting whether an observation is in the treatment group as a function of the other control variables:

$$P(obs\ i\ gets\ treatment) \approx g(X_i).$$

These are termed the propensity scores, and the 'matched' control observations are then selected based on the closest propensity value to each of the treatment observation (subject to some minimum level of agreement, or 'calliper'). One-to-one matching is common, but in situations of large control groups many-to-one matching can also stabilise results.

The assumptions of propensity matching are largely the same as regression-based approaches. If there is an important confounding variable not included in X, or if there are other selection effects attached to the treatment that also correlate to the response, then the effect size estimated through matching may be wrong. More formally Rosenbaum & Rubin (1983) specify the key assumptions as:

[2] As implemented in packages such as `grf` for R

- The treatment assignment is independent of the response, conditional on the control variables. This means there are no invisible selection effects, such as subjective judgment allocating to a treatment.
- All included observations have a nonzero chance of being in treatment or control. So if there are observations that had no chance of receiving treatment (or 100% chance), these should be excluded.

Under these assumptions the key theorems establish that propensity matching is a balancing score (recovers the same distribution of covariates in treatment and matched control, and that derived statistics (such as treatment effect) will similarly be unbiased.

While the underlying assumptions are ultimately strong, the procedure allows an observational dataset to be considered carefully from a causal view. Having an explicit control group can be advantageous for looking at variations in outcomes (such as how treatment effect varies across subgroups).

Example – 3k to work

We look at an example taken from the New Zealand welfare system; the 3K to Work grant from Greenfield et al. (2017). This was first introduced as 3K to Christchurch in July 2014 and expanded to be nationwide in December 2015. The grant was paid to people who were offered employment in a different region as both incentive and payment to assist with the costs of moving. The grant was focused on clients who were younger and had been in receipt of main benefits for over 6 months.

For the analysis we looked at 1,300 grants paid up to 30 June 2016 for people who were receiving the main work-ready jobseeker benefit. Our outcome of interest was benefit status four quarters later.

A propensity score was fit, using a tree-based gradient boosting machine for whether a person took up the program. Many variables were found to be significant in affecting the rate of take-up, as shown in Figure 4.

Figure 4 – Relative variable important in the propensity model for 3K to Work analysis

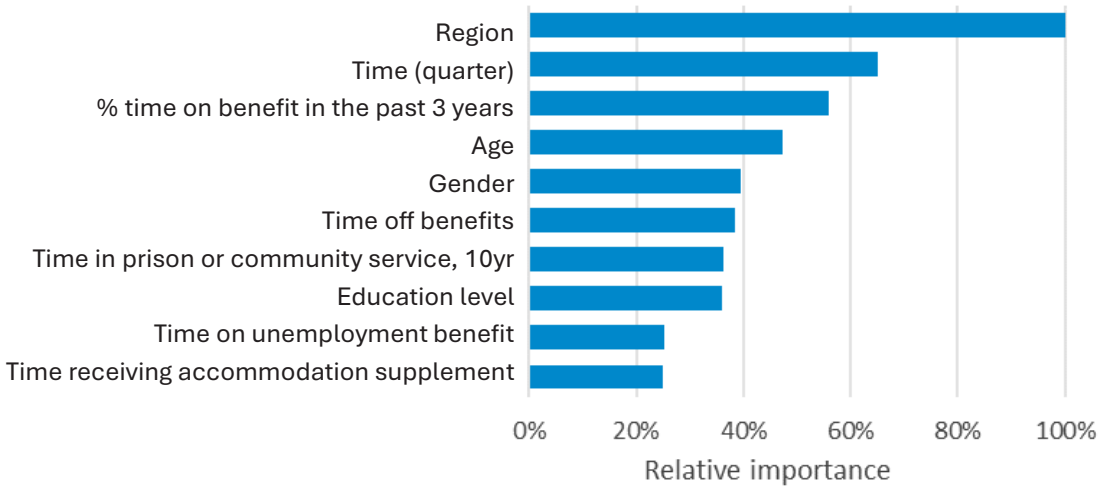
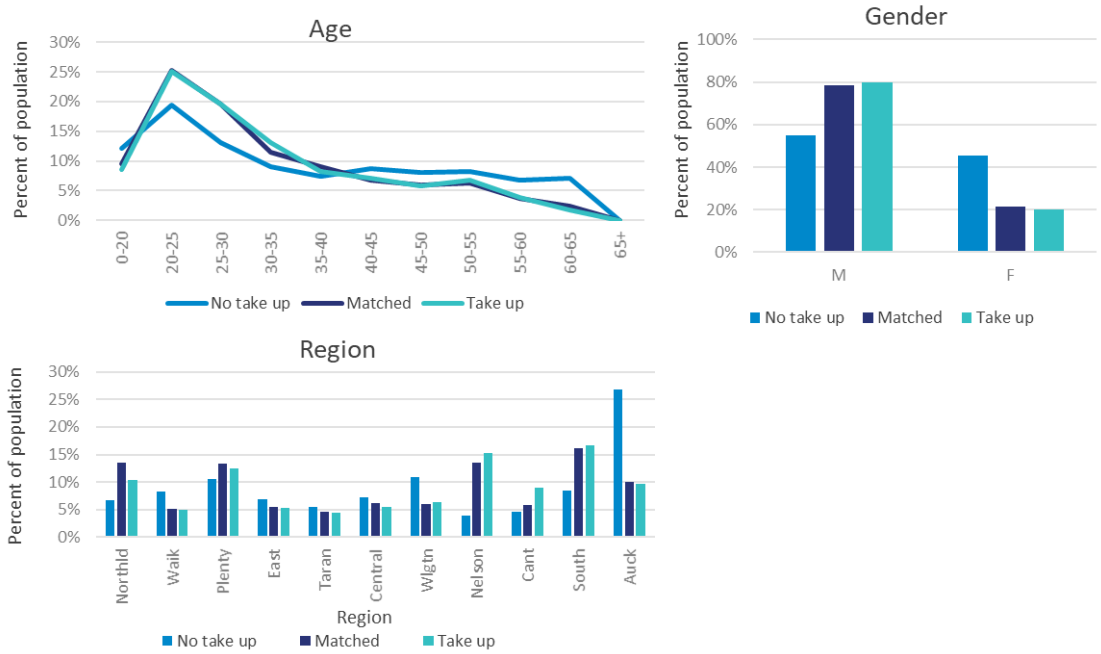


Figure 5 shows the distribution for the treatment group (those taking up the grant), broader welfare population and our matched distribution. People were more likely to be South Island (reflecting the original 3K to Christchurch premise) and much less likely to be moving out of Auckland. Young males were much more likely to be taking up the program.

Figure 5 – Variable Importance in the propensity model of JS-WR clients participating in 3k to Work

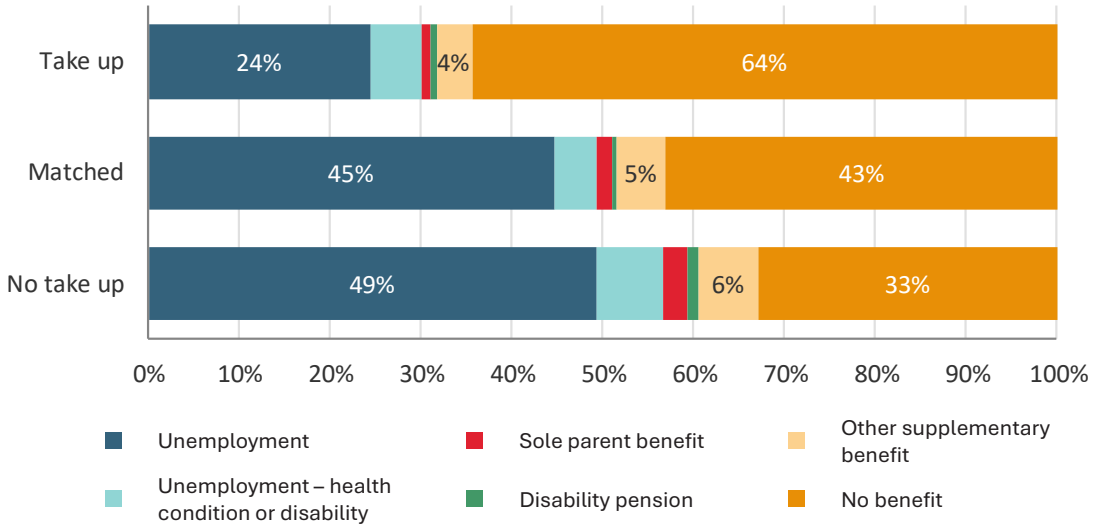


We looked at two outcomes in particular:

1. **Benefit exits** – Are grant recipients more likely to be off main benefits a year later?
2. **Sustained exits** – Are grant recipients who exit in quarter one more likely to remain off for four quarters?

On the first question, Figure 6 shows the benefit status one year later. It shows that 68% of people in the 3k take-up group ('treatment') are off main benefits (no benefit or supplementary only), compared to 39% for people who do not take up the program. A large gap is unsurprising, since it involves comparing people who have a job offer to those that do not; there will be significant selection effects. However, the propensity match is still important, since it shows us that the correct comparison rate should be 48%; the group taking up the grant tend to have better employment outcomes anyway. This means that the upper bound on impact is 68% – 48% = 20%, of which a portion may included selection effects (people more likely to find a job in any case, including people who would have moved in any case).

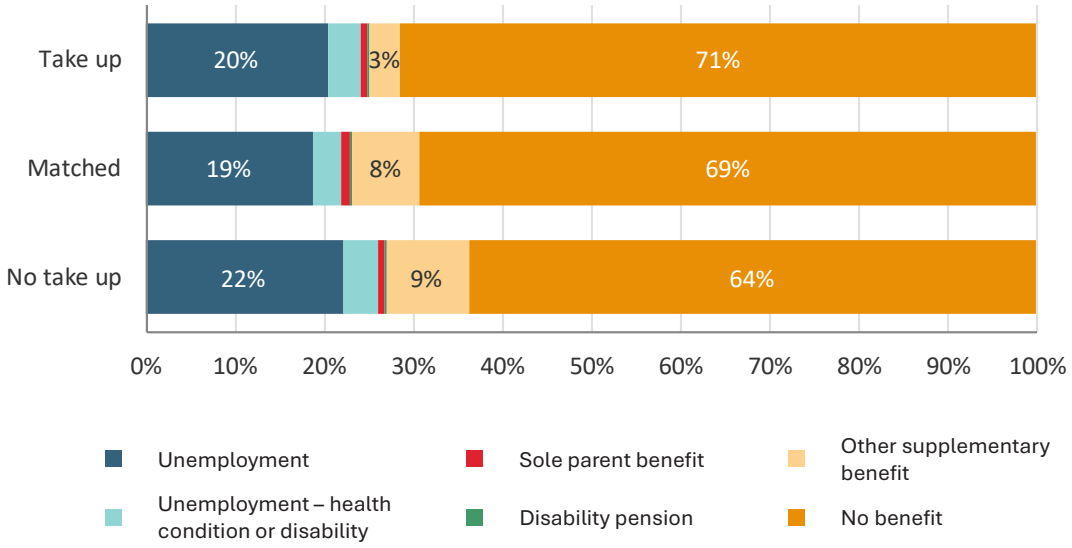
Figure 6 – Benefit status one year later



On the second question, Figure 7 shows a comparison of what happens to people who move off benefits in each of the groups. For the take-up group, 74% of people do not return to main benefits over

the next year after exit. For the broader no take-up group 73% of people do not move back and for the matched group 77% do not (although with a higher fraction of supplementary benefits). These numbers are broadly similar; it suggests that exits from the 3k to work program are sustained over the first year just as well as other types of benefit exits. This is useful, since one potential dynamic is people moving back to their home regions and re-entering benefits.

Figure 7 – Conditional on moving off benefit in first quarter, what proportion moved back onto benefits within a year



3.4 Instrumental variables

Description

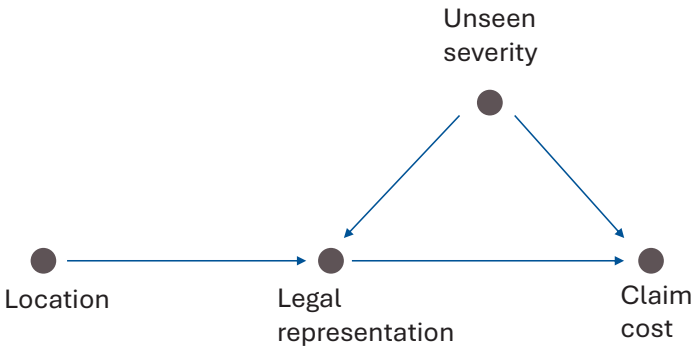
We often have the situation where the treatment variable of interest is correlated with the error term of a model (or commonly, correlated with a variable that is not visible to the model and affects the response).

For example, suppose we are trying to estimate the impact of legal representation on claims cost. A common confounder is injury severity, to the extent not captured in the data – see the figure below. A more severe injury is more likely to lead to legal representation and higher claim cost, so straight estimation of the legal representation effect overstates the impact of legal representation on claim size.

An instrumental variable (IV) is one that is correlated to the treatment but not to the confounder. Suppose location affects the rate of legal representation but not severity[3]. In this case it can act as a instrumental variable.

[3] There is evidence that regional accidents are typically higher speed and more severe. However, our example sees highest legal representation rates in certain parts of Sydney, so regional effects are likely secondary.

Figure 8 – Potential confounding of legal representation due to injury severity not visible on recorded data, and the role of location as an instrumental variable

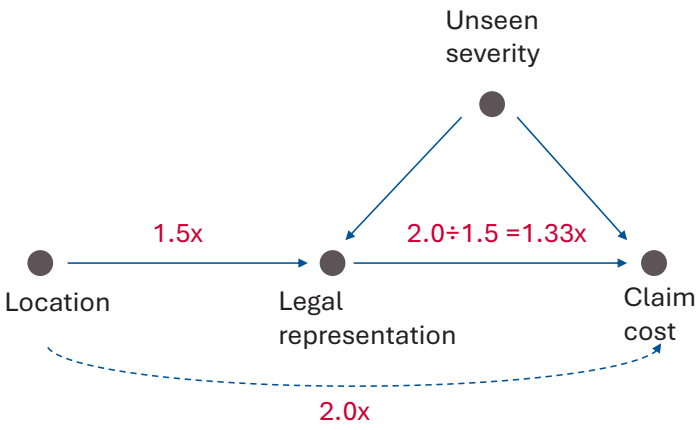


In such a case we can use the relationships to estimate our key effect (the effect of legal representation). The basic intuition, also shown in Figure 9, is relatively straightforward:

- Suppose we regress claim cost against location and found the claim cost in region B is 2.0x that of region A
- Suppose we regress legal representation against location and found that rates are 1.5x in region B
- Dividing 2.0 ÷ 1.5 = 1.33, this is the factor that must correspond to the impact of legal representation on claims cost.

The key observation is that neither regression will be confounded by the unseen severity factor. The resulting factor can be compared to the direct estimate to see how large the confounding effect is.

Figure 9 – Using an instrumental variable to indirectly estimate the impact of legal representation



The strength of the instrument is how well it predicts the treatment; in our examples if location was only a weak predictor of legal representation, then it would produce a very uncertain estimate of the treatment effect.

In practice when there are other predictor variables, more formal model structures are used. Two-stage least squares regression is a common approach where the treatment variable is first estimated using the instrumental variable and other controls; then in the main regression the treatment variable is replaced by the predicted values from the first step.

To describe two-stage least squares regression more formally, suppose we are interested in the following regression formula:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \delta T + u$$

With $E(u) = 0$ and $Cov(X_j, u) = 0$ for each j. If the residual effect u is also uncorrelated to T, then the equation could be estimated as normal and an unbiased estimate of treatment effect δ derived. However, if T is correlated with the residual effects u (e.g. legal representation correlated with unseen

aspects of severity), then standard regression will **not** return consistent parameter estimates. If the correlation is positive, then δ will be too large since it captures some of the covariation with u.

However, suppose an instrumental variable Z exists for T and we can estimate

$$T = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \theta Z + \epsilon$$

With θ not equal to zero and ϵ uncorrelated to the X_j and Z. Then create a set of predicted values for the treatment variable $\hat{T} = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \theta Z$ and use this in estimation of Y:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \delta \hat{T} + u$$

This **will** give consistent estimates for the parameters, and we can calculate corrected standard errors associated with the procedure.

The procedure can be extended naturally to situations with multiple treatment variables or even multiple instrumental variables.

Example – The impact of legal representation in CTP

As part of its legislated review of minor injury claims in 2019, SIRA commissioned Taylor Fry to undertake some analysis of patterns in minor injuries claims[4]. This new class of claims was introduced with the reforms in the Motor Accident Injuries Act 2017 for people with less severe injury that are expected to recover quickly. The definition of 'minor injury' includes:

- A soft tissue injury, or
- A minor psychological or psychiatric injury.

Our analysis spanned descriptive analysis of trends, claim characteristics and whether claims moved from minor to more general claims. One aspect of the analysis was looking at statistics related to legal representation and incidence of psychological injury[5]. Rates of legal representation were substantial, despite the fact that access to some forms of compensation such as common law was restricted.

Both legal representation and a declared psychological injury are strongly associated with higher claims cost; however, it is not immediately obvious whether these two effects have a strong causal link versus or whether much of the effect is confounded by injury severity, as per the earlier discussion.

The analysis was performed on 5,800 post-reform minor injury claims that were judged mature (at least 8 months development). Variables available covered a wide variety of claimant and payment characteristics.

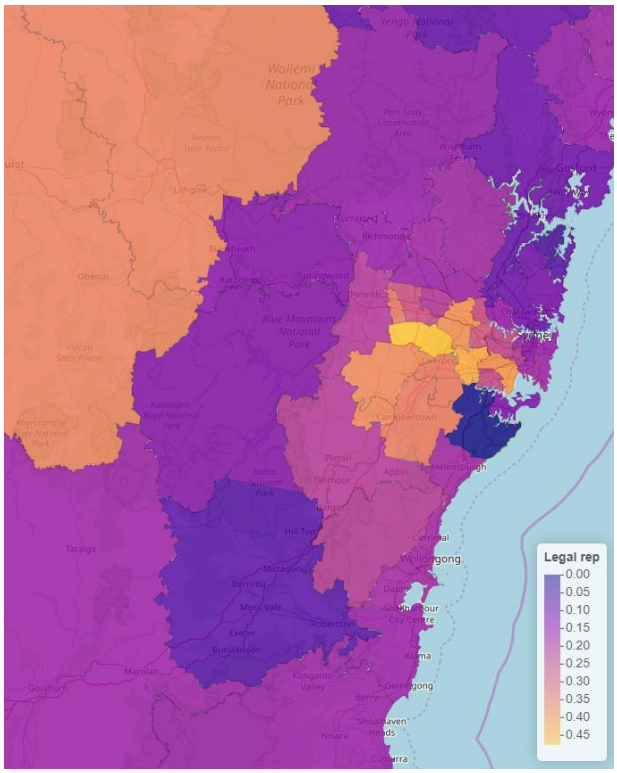
However, we also observe that rates of legal representation and psychological injury vary markedly by region. We modelled at an ABS statistical area 3 (SA3) level, which corresponds to districts of roughly 30k-130k people. The variability is shown in Figure 10. For legal representation some regions have rates 80% below the average and others 100% above. For psychological injury some regions have rates 70% below average and others 60% above. We do not believe regional differences in claim severity explain the difference; most is behavioural.

[4] See https://www.sira.nsw.gov.au/consultations/sira-review-of-the-minor-injury-definition for further information

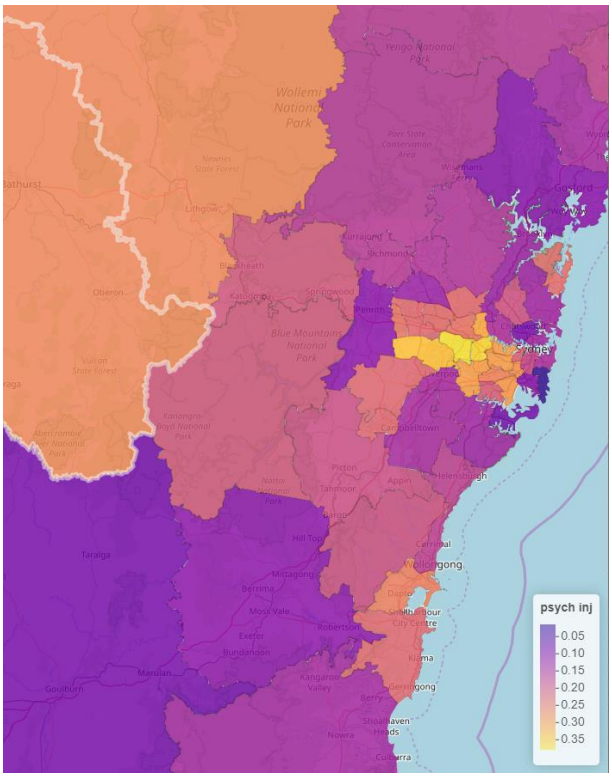
[5] We are actually assessing declared psychological injury; there is some evidence that not all psychological injuries are declared and recorded in claims data.

Figure 10 – Regional variability of for mature minor injury claims, October 2019

Minor injury legal representation rates



Psych injury rates



To understand how these variables relate to claims cost, we fit a two-stage least squares model with:

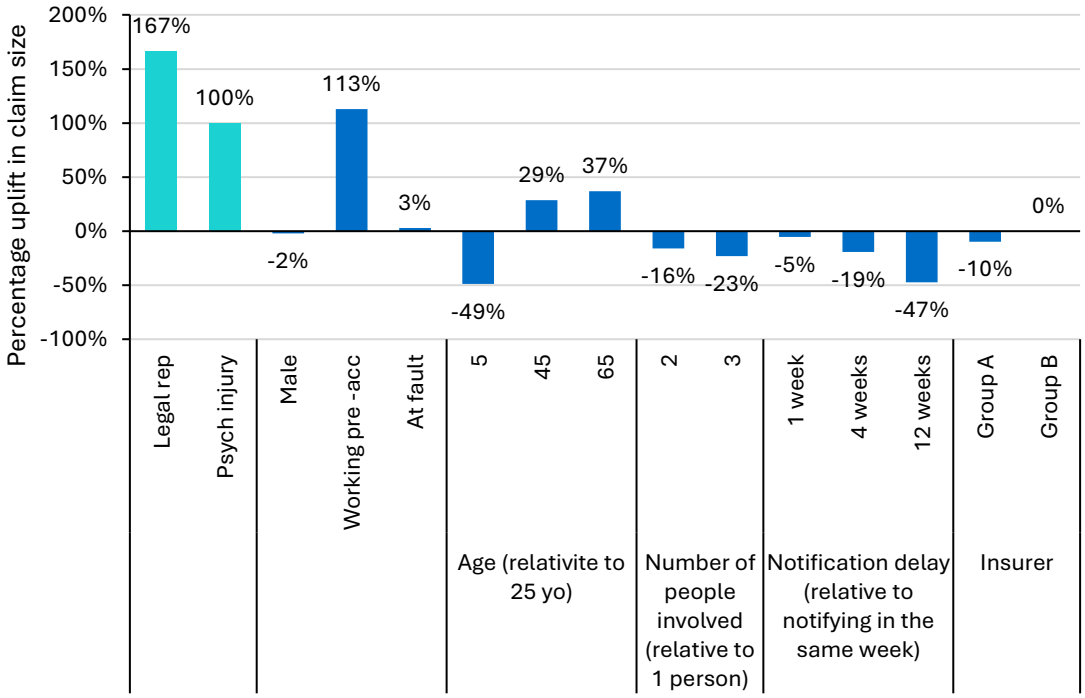
- Target the log of claims cost, defined as treatment + rehab + loss of income
- Two treatment variables to be estimated by IV – legal representation and psychological injury
- A range of other early claim characteristics used as predictor (control) variables, including age, gender, pre-injury work status, at-fault indicator, number of people involved, notification delate and insurer.
- SA3 region (as a categorical variable) was used as an IV for the two treatments.

After fitting the model we obtained results as shown in Figure 11. Most importantly:

- **Obtaining legal representation more than doubles the expected claim size** (increase of 167%, ± 78%)
- **Choosing to register a psychological injury doubles the expected claim size** (increase of 100%, ± 70%)
- These effects compound in many regions where both rates are high
- There is significant uncertainty in the causal terms, due to the high underlying variability in claim size as a function of early claim characteristics.

We regard this as very strong evidence that behavioural choices are driving higher rates of legal representation, psych injuries and claim costs.

Figure 11 – Claim size relativities for two-stage least squares model. Legal representation and psychological injury effects are estimated by IV.



There are obviously imperfections. For instance, there may be regional factors affecting claim cost that are not mediated by legal representation and psychological injury but correlated by them. To the extent this is true, the effect size measured might contain some degree of broader behaviour impacts.

We also performed an analysis where the rate of psychological injury was the target, and legal representation the treatment. A similar set of control variables were used. Region again was used as an IV variable. We found that **legal representation adds 51 percentage points (±8 pts) to the chance of a psych injury being recorded**.

3.5 Regression discontinuity

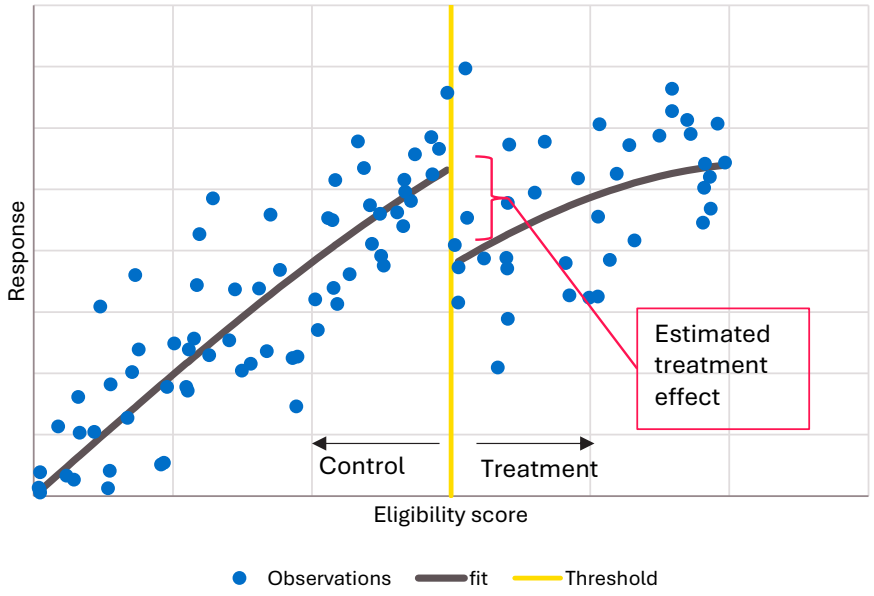
Description

Often a program or treatment uses an eligibility threshold to determine whether a person should be included. This creates a selection effect; a straight comparison of people who are or are not in the treatment group will not be a fair one. Often the people in the treatment group (above the threshold) will be those at right of worse outcomes and so a direct comparison will often see the treatment group do worse.

However, we can leverage the threshold as a source of signal. If there is a lot of data, we can restrict attention to the subset of observations near the threshold and do a straight comparison. If there's less data, fitting a curve to the entire dataset with a parameter to allow for a step change at the threshold can given an estimate of treatment effect.

This process of estimation is show schematically in Figure 12. The effect of interest, the impact of treatment on the response, is illustrated as the size of the discontinuity at the threshold.

Figure 12 – Schematic of estimation of effect size using a regression discontinuity fit



Example – The impact of scholarship supports for disadvantaged youth

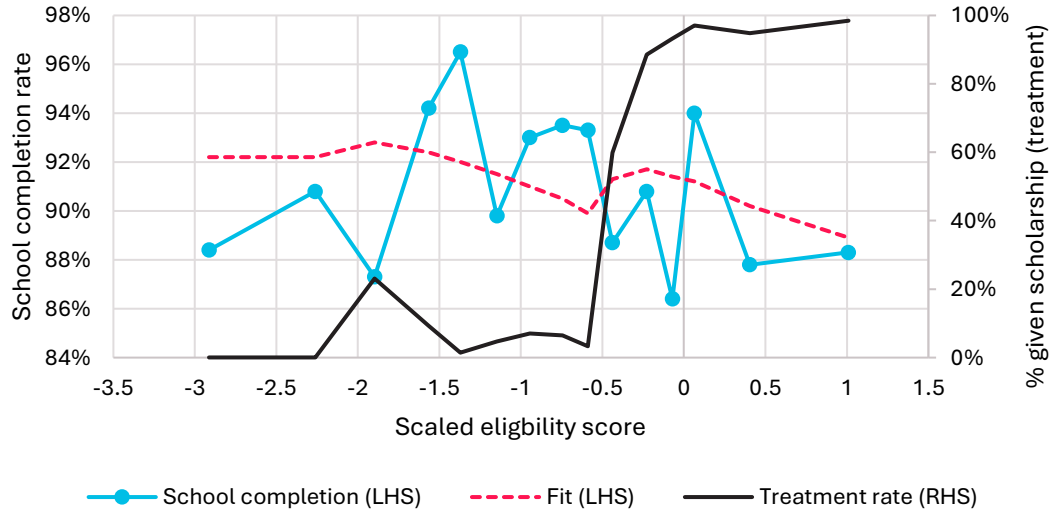
The NSW Department of Communities and Justice (DCJ) run a scholarships program for disadvantaged youth, primarily for those form social housing or in out-of-home care. This comprises $1,000, typically spent on education equipment such as laptops. A scoring rule using a range of variables is used to assign priority to those applying for the program. This is not a pure threshold; the cut-off points have changed over the years depending on demand and can also vary by region. The analysis of this type of threshold is often term a 'fuzzy' regression discontinuity, since a hard threshold effect (modelled as an indicator function) is replaced by a graduated term (reflecting the increasing probability of receiving the treatment with score).

A natural question is whether these scholarships have an impact on various outcomes. Using data linkage, we are able to measure annual school completion rates for those who applied for the program (both those awarded the scholarship and those who missed out).

Using a dataset of 1,822 student-years across three years, we can test a fuzzy regression discontinuity using scaled eligibility score (adjusted so the threshold from year to year is aligned). The results are illustrated in Figure 13. There remains volatility in school completion rates which makes conclusive findings difficult. However, there is good evidence of a downward trend in completion with eligibility score, and weak evidence of an upwards bump in completions associated with receiving a scholarship.

We tested the result a variety of ways, including other outcomes such as HSC attainment, and reached the same overall conclusion – any improvement is not strong enough to be judged statistically reliable. The regression discontinuity result is still useful in establishing that we avoid concluding the opposite; outcomes overall are poorer for those in the treatment group (since the program targets higher need), which is controlled for in the procedure.

Figure 13 – School completion rates, actual and modelled, for scholarships program



3.6 Stepped wedge regression

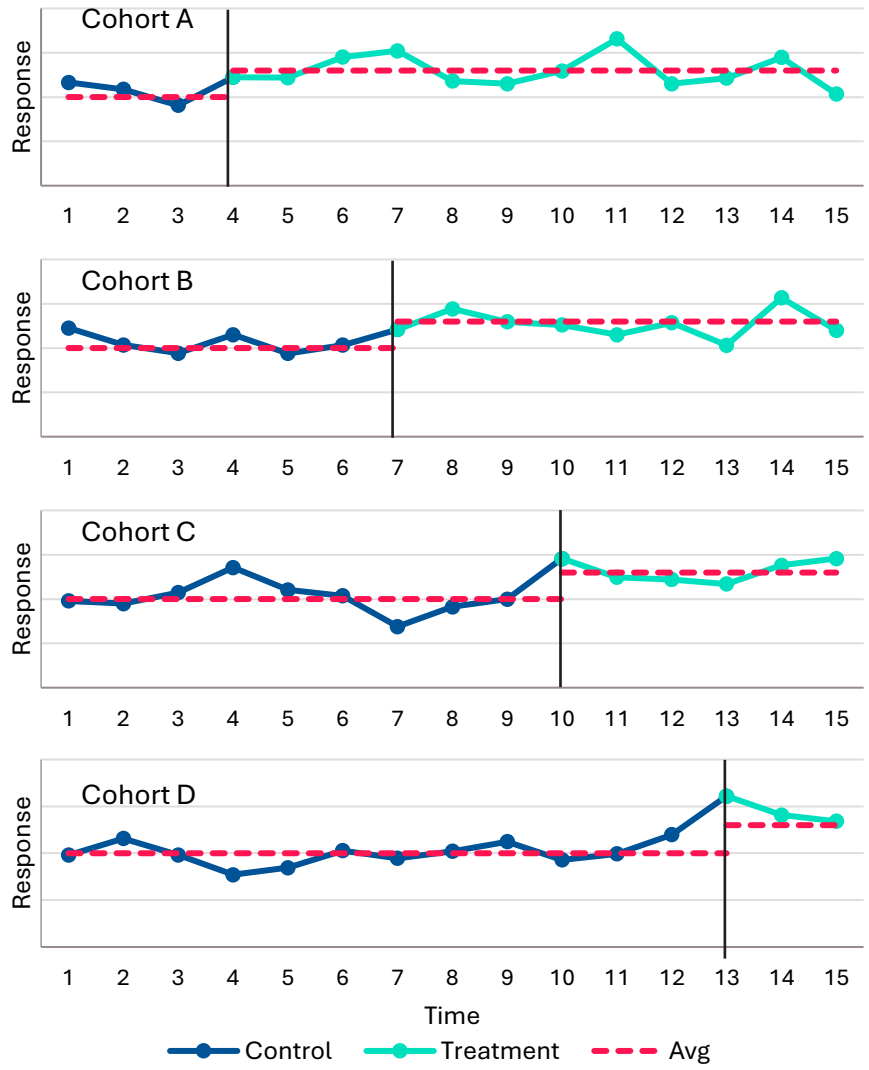
Description

In many situations a decision is made to apply a new intervention to all people, if there is a strong prior belief that it will deliver benefits. In such cases defining an in-time control group is not possible, but there are still ways to measure effects.

An interrupted time series analysis is a technique uses the same group as both treatment and control but compares the response before and after the introduction of the treatment. One weakness of such an approach is that other concurrent events may also be affecting the response, so isolating the impact of the treatment is difficult.

A stepped wedge design can alleviate this issue in situations where the timing of the treatment differs for different cohorts (or 'clusters'). In this case it is less likely that single external events will distort the measurement.

Figure 14 – Schematic of a stepped wedge design



Analysis should also recognise that covariation may occur within clusters. A mixed (or random-effects) model is often suitable to account for this.

There are natural drawbacks to stepped-wedge designs. Longer-term counterfactual observations (what would have happened had there been no intervention) are difficult. Temporal trends remain as potential confounders. And selection effects are still relevant if certain cohorts receive treatment first because they are judged most suitable.

That said, the design is useful, since it often provides a practical approach to measurement. Many programs are rolled out in stages; either because it is first piloted, or different business units (e.g. regions) apply changes on their own timetables. In such cases the variation in timing becomes a source of strength.

4 Conclusion

The continued digitisation of programs and improved data collection will multiply the opportunity for analysis. If we are careful about how the datasets are generated and how we analyse them, then we can usefully address questions of causation. As actuaries, we are often the people asked to look at the data and derive insight. Quasi-experimental methods represent a useful set of tools that can answer the questions people are most interested in.

5 References

- Ayoub, F., Sato, T., & Sakuraba, A. (2021). Football and COVID-19 risk: Correlation is not causation. Clinical Microbiology and Infection, 27(2), 291-292.
- Greenfield A., Miller, H., McGuire, G., & Dixie, L. (2017) *Annual Report of the Benefit System for Working-Age Adults*. New Zealand Ministry of Social Development. Accessed September 2021 at https://www.msd.govt.nz/about-msd-and-our-work/publications-resources/evaluation/annual-report-of-the-benefit-system-for-working-age-adults.html
- Haynes, K. E., Kulkarni, R., Li, M. H., & Siddique, A. B. (2020). The Impact of Differing COVID-19 Mitigation Policies: Three Natural Experiments. *Available at SSRN 3702606*.
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: what if*. Boca Raton: Chpan & Hall / CRC.
- Hope, D., & Limberg, J. (2020). The economic consequences of major tax cuts for the rich. Longdon School of Economics Working paper, accessed September 2020 at
- Horby, P., Lim, W. S., Emberson, J., Mafham, M., Bell, J. L., Linsell, L., ... & Adams, R. (2021). Dexamethasone in hospitalized patients with Covid-19. *New England Journal of Medicine*, *384*(8), 693-704.
- Howick, J., Glasziou, P., & Aronson, J. K. (2009). The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *Journal of the Royal Society of Medicine*, *102*(5), 186-194.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Jorge, A. (2021). Hydroxychloroquine in the prevention of COVID-19 mortality. *The Lancet Rheumatology*, *3*(1), e2-e3.
- Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational psychologist*, *51*(3-4), 395-405.
- Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), 134-161.
- Leahy, S., Leahy Gatfield, R., Quail, S., Eastman, C., Christian F., and Williams, I. (2020). Evaluation of the Resilient Families service. Commissioned by the NSW Office of Social Impact Investment. Accessed September 2021 at https://www.osii.nsw.gov.au/assets/office-of-social-impact-investment/Evaluation-of-the-Resilient-Families-Service-Final-Evaluation-April-2020.pdf.
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.
- RECOVERY Collaborative Group (2020). Effect of hydroxychloroquine in hospitalized patients with Covid-19. *New England Journal of Medicine*, *383*(21), 2030-2040.
- Rosenbaum, P. R., & Briskman. (2010). *Design of observational studies* (Vol. 10). New York: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.
- SIRA, (2019). *NSW Young Drivers Telematics Trial: Findings, implications and lessons learnt*. Accessed September 2021 at https://www.sira.nsw.gov.au/__data/assets/pdf_file/0010/556264/NSW-Young-Drivers-Telematics-Trial.pdf.
- Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Penguin UK.
- Trimboli, L. (2019). Youth on Track randomised controlled trial: Process evaluation. Accessed September 2021 at https://www.bocsar.nsw.gov.au/Publications/BB/2019-Report-Youth-on-track-randomised-controlled-trial-BB141.pdf.