Ex- post evaluation of transport interventions using causal inference methods

September 2017

A Schiff (Schiff Consulting)
L Wright and T Denne (Resource Economics)

Resource Economics was contracted by the NZ Transport Agency in 2016 to carry out this research.

This publication is copyright © NZ Transport Agency. This copyright work is licensed under the Creative Commons Attribution 4.0 International licence. You are free to copy, distribute and adapt this work, as long as you attribute the work to the NZ Transport Agency and abide by the other licence terms. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

While you are free to copy, distribute and adapt this work, we would appreciate you notifying us that you have done so. Notifications and enquiries about this work should be made to the Manager National Programmes, Investment Team, NZ Transport Agency, at research@nzta.govt.nz.

**Keywords:** causal inference, cross-sectional regression, difference-in-differences, empirical Bayes, ex-post analysis, instrumental variables (IV) regression, interrupted time series (ITS), panel regression, propensity score matching (PSM), regression discontinuity (RD)
An important note for the reader

The NZ Transport Agency is a Crown entity established under the Land Transport Management Act 2003. The objective of the Agency is to undertake its functions in a way that contributes to an efficient, effective and safe land transport system in the public interest. Each year, the NZ Transport Agency funds innovative and relevant research that contributes to this objective.

The views expressed in research reports are the outcomes of the independent research, and should not be regarded as being the opinion or responsibility of the NZ Transport Agency. The material contained in the reports should not be construed in any way as policy adopted by the NZ Transport Agency or indeed any agency of the NZ Government. The reports may, however, be used by NZ Government agencies as a reference in the development of policy.

While research reports are believed to be correct at the time of their preparation, the NZ Transport Agency and agents involved in their preparation and publication do not accept any liability for use of the research. People using the research, whether directly or indirectly, should apply and rely on their own skill and judgement. They should not rely on the contents of the research reports in isolation from other sources of advice and information. If necessary, they should seek appropriate legal or other expert advice.
Acknowledgements

We are grateful to the many people who have assisted with this project, particularly our steering group (Ernest Alberquerque, Siobhan Bakker, Antony Flux and Haobo Wang) and our peer reviewers: Professor Dan Graham (Imperial College, London) and Dr Adolf Stroombergen (Infometrics).
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive summary</td>
<td>7</td>
</tr>
<tr>
<td>Abstract</td>
<td>10</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>11</td>
</tr>
<tr>
<td>1.1 Objectives and scope of this report</td>
<td>12</td>
</tr>
<tr>
<td>1.2 Case studies and software</td>
<td>13</td>
</tr>
<tr>
<td>1.3 Organisation of this report</td>
<td>13</td>
</tr>
<tr>
<td>2 Basic issues and concepts</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Key issues for estimating causal effects</td>
<td>14</td>
</tr>
<tr>
<td>2.2 The importance of counterfactuals</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Policy context for causal inference analysis</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Measuring treatment effects and the fundamental problem of causal inference</td>
<td>17</td>
</tr>
<tr>
<td>2.5 Randomised experiments</td>
<td>18</td>
</tr>
<tr>
<td>2.6 Summary of practical issues in ex-post analysis</td>
<td>21</td>
</tr>
<tr>
<td>2.7 Predictive versus causal inference</td>
<td>22</td>
</tr>
<tr>
<td>2.8 Multi-valued and continuous treatments</td>
<td>22</td>
</tr>
<tr>
<td>3 Ex-post evaluation in practice</td>
<td>23</td>
</tr>
<tr>
<td>3.1 New Zealand</td>
<td>23</td>
</tr>
<tr>
<td>3.2 International examples</td>
<td>25</td>
</tr>
<tr>
<td>3.3 Summary</td>
<td>27</td>
</tr>
<tr>
<td>4 Overview of key techniques and literature review of empirical applications</td>
<td>28</td>
</tr>
<tr>
<td>4.1 Causal inference with simple regression models</td>
<td>28</td>
</tr>
<tr>
<td>4.2 Propensity score matching</td>
<td>32</td>
</tr>
<tr>
<td>4.3 Difference-in-differences</td>
<td>37</td>
</tr>
<tr>
<td>4.4 Instrumental variables</td>
<td>43</td>
</tr>
<tr>
<td>4.5 Regression discontinuity</td>
<td>47</td>
</tr>
<tr>
<td>4.6 Interrupted time series analysis</td>
<td>49</td>
</tr>
<tr>
<td>4.7 Bayesian methods</td>
<td>53</td>
</tr>
<tr>
<td>4.8 Multi-valued and continuous treatment methods</td>
<td>57</td>
</tr>
<tr>
<td>4.9 Summary</td>
<td>61</td>
</tr>
<tr>
<td>5 Evaluation and selection of techniques</td>
<td>62</td>
</tr>
<tr>
<td>5.1 Characteristics of methods</td>
<td>62</td>
</tr>
<tr>
<td>5.2 Choosing a method based on available data</td>
<td>63</td>
</tr>
<tr>
<td>5.3 Data collection to enable causal inference analysis</td>
<td>65</td>
</tr>
<tr>
<td>5.4 Methods for evaluating different types of intervention</td>
<td>65</td>
</tr>
<tr>
<td>5.5 Conclusions on methods</td>
<td>66</td>
</tr>
<tr>
<td>6 Method application guidance</td>
<td>67</td>
</tr>
<tr>
<td>6.1 Hamilton city safer speed areas case study</td>
<td>67</td>
</tr>
<tr>
<td>6.2 Simple cross-sectional regression models</td>
<td>69</td>
</tr>
<tr>
<td>6.3 Difference-in-differences methods</td>
<td>75</td>
</tr>
<tr>
<td>6.4 Panel fixed-effects models</td>
<td>80</td>
</tr>
<tr>
<td>6.5 Propensity score matching</td>
<td>82</td>
</tr>
</tbody>
</table>
Executive summary

This report gives guidance for empirical ex-post evaluation of projects, policies, and other interventions in the transport sector. We focus on statistical methods that are designed to estimate the effects caused by an intervention, known as methods of ‘causal inference’. There are many causal inference methods and we provide guidance for selecting a method depending on the data that is available and the type of intervention being evaluated.

We focus on causal inference techniques for binary interventions, ie cases where the only relevant comparisons are between outcomes with and without the intervention. This is consistent with most of the causal inference literature to date. However, in some cases the effects of non-binary interventions may also be of interest, eg if an intervention can be applied at different ‘levels’ (or ‘doses’) then it may be useful to analyse how outcomes vary with the design of the intervention.

We also illustrate the application of causal inference methods to two New Zealand case studies: the implementation of ‘safer speed areas’ in Hamilton City, and Auckland’s Northern Busway. Both case studies are intended to illustrate the use of causal inference techniques for ex-post evaluation, and are not intended to be a full evaluation of the effects of the two interventions.

Counterfactuals and ex-post evaluation

The effects caused by an intervention must be estimated by comparing post-intervention outcomes against a counterfactual scenario. The counterfactual represents the state of the world without the intervention, and the effects of the intervention are estimated from the differences between actual and counterfactual outcomes. Without a counterfactual, analysis of outcomes can only be descriptive, not causal.

It is not possible to observe outcomes for a given place or group of people or businesses with and without an intervention at the same time. The counterfactual must therefore be estimated using data and statistical models, and some assumptions are required when comparing the counterfactual with post-intervention outcomes. The assumptions required, and therefore the robustness of the estimated effects of an intervention, depend crucially on the type of data available for analysis.

Estimating causal effects

The ‘gold standard’ for estimating causal effects is a randomised experiment where many units (eg people, places, businesses) are randomly assigned to ‘treatment’ and ‘control’ groups. If done properly, random assignment cancels out the effects of all other factors aside from treatment (ie the intervention) on outcomes, and the effects of the treatment can then be inferred from differences in outcomes between the two groups.

However, in the transport sector, random assignment is almost always impossible and evaluators must rely on observational rather than experimental data. Even if treatment and control groups of units can be established, the decision to apply an intervention to a unit may be based on factors that cannot be perfectly observed. Such ‘confounding’ can make it difficult to separate the effects of an intervention from the effects of other factors. In some cases, interventions apply everywhere at once, so a control group cannot be established and an alternative approach must be used to generate a counterfactual.

The estimated effects of an intervention from observational data are likely to be biased if the counterfactual is not appropriately defined and if the influence of other factors aside from the intervention, including potentially unobserved factors, on outcomes are not accounted for in the analysis using techniques that we review in this report. Simple before and after comparisons, where ‘before’
outcomes are used as the counterfactual, are unlikely to give good estimates of causal effects. Simple comparisons of outcomes between a treatment group and a control group where treatment is not randomly assigned can also run into difficulties due to the likelihood that unobserved factors affected selection of the treatment group and outcomes.

Causal inference techniques applied to observational data use statistical models to define counterfactuals and separate the effects of an intervention from other observed factors. If suitable data is available, it is also possible to distinguish from the effects of certain unobserved factors.

Evaluators should be aware of the assumptions and limitations of causal effects estimated from observational data. In general terms, the extent to which the results depend on assumptions is inversely related to the richness of the analysed data. If possible, it is important to consider the data that will be needed for robust ex-post evaluation before an intervention is implemented, and set up processes to collect the required data.

Causal inference methods

We review and evaluate several causal inference methods widely used in the applied causal inference literature, including for evaluation of transport interventions. Our review focuses on cross-sectional regression, difference-in-differences (DiD), propensity score matching (PSM), fixed effects (FE), instrumental variables (IV), regression discontinuity (RD), and interrupted time series analysis (ITSA).

In chapter 6 we review the key features of each technique, summarise its assumptions and limitations, and discuss some selected examples of it in the transport evaluation literature. In chapter 8 we give practical guidance for the application of each technique, including worked examples in Excel, R and STATA using the Hamilton safer speed areas case study. In chapter 9 we apply some of the techniques to ex-post evaluation of the effects of Auckland’s Northern Busway.

Evaluation of methods and guidance for choosing a method

Methods that exploit both temporal and cross-sectional variation, such as DiD and fixed effects models, are more likely to give estimates of causal effects that are free from bias, compared with methods that rely on a single source of variation, such as cross-sectional regression models, RD and ITSA. The latter methods are only able to distinguish the effects of an intervention from other factors that have been observed and can be included in the analysis, whereas DiD and fixed effects methods can also eliminate bias caused by certain types of unobserved factors (see section 6.3). IV methods can also correct for such bias, but can be difficult to apply in practice as suitable ‘instruments’ can be hard to find (see section 6.4).

This suggests using DiD or FE models for ex-post evaluation where possible. Such methods can be further improved by the application of PSM to increase the comparability of the treatment and control group. The combination of PSM and DiD is particularly powerful yet relatively straightforward to implement if suitable data is available. However, in practice, data limitations may make it impossible to use sophisticated techniques and evaluators may need to fall back on simpler methods, most commonly cross-sectional regression models and ITSA. In such cases, it is important to be aware of the additional assumptions necessary to estimate causal effects (see sections 6.1 and 6.6).

Our recommendation is the use of DiD or fixed-effects models (with or without PSM), if this is possible. In our view, ITSA methods are less reliable, but may be the only feasible option in some cases.

Certain types of intervention and certain types of outcomes lend themselves to particular techniques for ex-post evaluation, or make others infeasible:
Executive summary

- One-off local projects may be difficult to evaluate via PSM or other purely cross-sectional methods, as these require multiple control and treated units, but a suitable control group may be difficult to find for a one-off project and it may not be possible to define multiple treated units for the project.

- Projects or policies applied in specific geographic areas are commonly evaluated using DiD or panel data methods as other (non-treated) areas can provide acceptable counterfactuals if data on pre- and post-intervention outcomes is available.

- Homogenous local or regional projects selected by the same cost–benefit analysis (e.g. multiple bridge widening projects that have undergone the same cost–benefit analysis) favour an RD design as those ‘nearly treated’ units may become appropriate counterfactuals. This is because homogenous projects with comparable net benefit prospects are likely to have similar characteristics and thus similar counterfactuals.

- National policies applied to the whole population or all geographic areas are often evaluated via ITSA or empirical Bayes methods, whereby pre-intervention outcomes become the ‘control’ units in absence of a proper control group.

Data collection to facilitate ex-post evaluation

A key theme of this report is that robust ex-post evaluation via causal inference techniques requires a suitably rich dataset. If possible, data on outcomes and characteristics of treatment and control units should be collected ex-ante as well as ex-post to enable one of the more robust methods outlined above to be implemented. Ideally the data collected for evaluation will:

- be a panel dataset containing observations of pre- and post-intervention outcomes of the treatment and a control group
- include data on other factors that are expected to influence outcomes of either group, so these can be used as explanatory variables
- if random assignment is not possible, be assigned based on observed characteristics
- have broad overlap in the characteristics of units in the treatment and control groups.

Evaluators also need to carefully consider:

- What is the counterfactual scenario? Is it a continuation of the status quo, or will some kind of intervention occur anyway?
- What units are a sufficient control group? Ideally, these units will have similar attributes to the treatment but will not be directly or indirectly affected by the intervention. In transport applications, this may require the treatment and control groups to be physically distant from each other.
- How and why outcomes of the treated units may differ from outcomes for the control units, in absence of an intervention? If the two groups are systematically different in ways unrelated to the intervention, these differences need to be measured so they can be controlled.
- If the results of the evaluation need to be generalised beyond the treatment group, there must be sufficient overlap in the characteristics of the treatment and control units.

Recommendations

Based on the research summarised in this report, we make the following recommendations:
Ex-post evaluation of transport interventions using causal inference methods

- All attempts at ex-post estimation of the effects caused by transport interventions must carefully consider how to define and measure an appropriate counterfactual for the intervention.

- Simple before-and-after comparisons of outcomes, or comparisons of average outcomes for a treatment and control group, are unlikely to give reliable estimates of causal effects, and these methods should not be used to draw conclusions in ex-post analysis. However, naïve comparisons can be a useful benchmark against which results from other methods can be compared.

- If the only data available for ex-post analysis is observations of a single outcome over time, an interrupted time series regression model should be used to estimate the effect of an intervention that occurred at a certain point in the time series. Where possible, the regression model should include other variables aside from the intervention that may have affected the outcome of interest over time. However, this method should not be used if the data is available to use other more reliable methods as it is difficult to be sure all other factors have been controlled.

- If the only data available for ex-post analysis is outcomes for treatment and control groups at a single point in time, cross-sectional regression analysis that includes variables measuring observed characteristics of the two groups should be used. However, the results of such analysis could be biased by the effects of unobserved factors that cannot be controlled.

- Whenever possible, a difference-in-differences or panel fixed-effects model should be used for estimating causal effects. These methods require pre- and post-intervention data on outcomes for treatment and control groups, and are more likely to produce reliable estimates of causal effects than the other methods described.

- If possible, reporting systems should be set up so suitable data for difference-in-differences analysis is available to evaluate interventions ex-post. This requires suitable treatment and control groups to be established, and observing pre-intervention characteristics of both groups.

Abstract

This report reviews, summarises and gives guidance for the choice and application of causal inference techniques to ex-post evaluation of the impacts of interventions in the transport sector. Such techniques seek to establish robust counterfactual outcomes against which post-intervention outcomes can be compared, while accounting for the effects of other factors aside from the intervention itself. We illustrate the application of key techniques using two New Zealand case studies, and give recommendations to improve the robustness of ex-post evaluations.
1 Introduction

In the transport sector, the impacts of decisions are subject to more analysis before they are made than afterwards. This is particularly true where decisions require the commitment of significant resources that cannot be re-allocated. Ex-ante (from before) analysis is used to choose among options in the face of resource constraints. Ideally ex-ante analysis is based on cost–benefit analysis (CBA), such as set out in the Economic evaluation manual (EEM) for transport projects (NZ Transport Agency 2016).

In contrast, ex-post (from after) analysis is done after a decision has been made and implemented. For example, post-implementation reviews (PIRs) are conducted by the NZ Transport Agency (the Transport Agency) for a selection of projects. Ideally, an ex-post analysis enables decision makers to estimate the effects caused by an intervention, rather than simply check if there are differences in outcomes before and after the intervention. It can also be used to understand whether the impacts of the intervention were as expected, and may shed light on the reasons for any variation from expectations.

Over time, the results of ex-post analysis can feed back into ex-ante decision-making processes, such as by updating the parameters representing causal effects that are used in CBA models. This should lead to better decision making and better allocation of scarce resources. Ex-post analysis is therefore a useful complement to ex-ante analysis, rather than a substitute for it.

Simplistic ex-post analysis is often descriptive, eg it describes:

- observed differences in variables of interest before and after an intervention
- observed differences between where an intervention was applied and where it was not applied
- differences between actual and expected outcomes.

While such descriptive analysis can show there is a difference in outcomes associated with an intervention, in most cases it cannot prove these differences were caused by the intervention and not something else. In this report we discuss techniques that can be used to perform more rigorous ex-post analysis and establish causal relationships.

For example, consider the decision to widen various roads with the objective of reducing traffic congestion (appropriately measured) on these roads. A simplistic ex-post analysis might compare measured congestion on each road immediately before and after the widening and assume any difference was caused by the intervention. However, among other possible problems, this does not account for other factors that affect traffic volumes and may have changed over the same time period. Estimating the change in congestion caused by the widening requires isolating and removing any effects on congestion caused by other factors.

Similarly, an estimate of the net benefits caused by an investment such as extending a motorway requires controlling in some way for all other factors that could have contributed to the observed differences, to isolate the impact of the investment itself. Most of this report is concerned with the ways this ‘controlling’ can be done, depending on the data that is available, to produce reliable estimates of causal effects. Such analysis may also need to deal with other technical issues, which we will discuss in this report.

A variety of empirical techniques exist for ex-post data analysis to estimate effects caused by an intervention. Such techniques are referred to as analysis of ‘treatment effects’, or methods for ‘causal
Ex-post evaluation of transport interventions using causal inference methods

Inference’. We use both terms interchangeably in this report. We also sometimes refer to the intervention or decision that is being analysed as a ‘treatment’.

An alternative approach to ex-post analysis is to use a predetermined theoretical model, such as a cost-benefit model, with parameters based on ex-post observations. This is typically less data intensive than a purely empirical approach, but it assumes the theoretical model connecting outcomes to the intervention is correct. In contrast, the empirical methods that are the focus of this report make fewer assumptions about the mechanism by which the intervention causes any effects, and instead concentrate on accurately estimating the magnitude of these effects.

The key to all causal inference is a counterfactual scenario against which an intervention is compared (Khandker et al 2010). The counterfactual represents what would have happened if the intervention had not been applied, and a robust counterfactual is crucial for robust estimation of the effects caused by the treatment. It is crucial that evaluators consider what an appropriate counterfactual scenario is, and estimate it as accurately as possible.

Depending on the available data, outcomes in the counterfactual may be estimated directly from observed data, or indirectly using predictions from a statistical model. There are various methods for doing this, and the choice of method will depend on how the intervention was implemented and what data is available. Not all techniques are equally robust, and they can be distinguished by their data requirements and the assumptions that are used to estimate a causal effect. Techniques that require stronger assumptions to estimate a causal effect are generally less robust, particularly where these assumptions cannot be tested.

1.1 Objectives and scope of this report

The objectives of this report are to:

- Review the main empirical techniques for ex-post evaluation using causal inference techniques and explain how they can be applied and interpreted, the data requirements and limitations.
- Review the literature on applications of causal inference techniques, with a focus on evaluation of interventions and policies in the transport sector.
- Develop practical guidance and tools for applying causal inference techniques to perform ex-post evaluation in the transport sector.
- Demonstrate the issues through a New Zealand case study.

Our emphasis is on clearly explaining the conceptual issues and developing practical guidance for the use of causal inference techniques in ex-post analysis. A comprehensive review of the statistical theory underlying causal inference is beyond the scope of this report, but we give references to where more detail can be found if required.

---

1 In statistics, ‘inference’ refers to the process of estimating unknown properties of an assumed statistical distribution using data. For example, the average height of men in a population can be estimated (‘inferred’) by calculating the average of a sample of men’s heights. Statistical inference can also be used to estimate parameters that reflect relationships, such as the correlation between a man’s height and his income. However, such inference does not necessarily imply a causal relationship. As we will explain, causal inference has more onerous data requirements and involves additional assumptions compared to non-causal inference.
1 Introduction

1.2 Case studies and software

This report includes two case studies illustrating the use and implementation of the empirical methods for ex-post analysis. These are the implementation of ‘safer speed areas’ in Hamilton city (see chapter 6) and Auckland’s Northern Busway (see chapter 7).

We have provided examples of ex-post analysis for the case studies using Excel (where possible) and R and STATA code. We note that while it is possible to implement some of the techniques in Excel or other similar software, spreadsheet analysis is limited by the built-in functions and it is not possible to do extensive diagnostic testing compared with what can be done with statistical software. For this reason, we recommend the use of statistical software such as R or STATA where possible.

1.3 Organisation of this report

The report:

- starts by discussing key issues and concepts relevant to causal inference analysis (chapter 2)
- reviews ex-post evaluation in practice in New Zealand and elsewhere (chapter 3)
- reviews the main techniques used in causal inference analysis (chapter 4)
- discusses how to select the best method depending on the data available and/or the type of intervention to be analysed (chapter 5)
- provides guidance on the practical implementation of the different techniques (chapter 6)
- applies the techniques to a detailed case study application – Auckland’s Northern Busway (chapter 7)
- provides conclusions and recommendations on the use of the techniques (chapter 8)
2 Basic issues and concepts

This section discusses key conceptual and practical issues for estimating causal effects from observational data. We focus on the case of binary interventions, i.e., where ‘observational units’ (defined below) either receive a treatment or not, and the treatment is the same for each unit that receives it.

2.1 Key issues for estimating causal effects

The following issues must be addressed in any empirical analysis of the effects caused by an intervention or treatment:

- **The treatment itself must be clearly identified and defined.** This includes determining the point in time at which the treatment occurred, and describing how the treatment was intended to work and the effects it was expected to have.

- **Definition of the observational units across which outcomes will be measured and the analysis of treatment effects will be performed.** These could be people, businesses, geographic areas, physical assets, points in time, or other well-defined units.

- **Definition and measurement of one or more outcome(s) the treatment was expected to affect.** These will be used to measure the causal effect and must be recorded for each observational unit being studied.

- **Definition of a counterfactual or ‘control’ against which the effects of the treatment will be compared.** As we will discuss, there are various ways a counterfactual can be defined or estimated, depending on the type of data available. In all cases, however, it is critical to have a counterfactual.

- **Estimation of the effect on the outcome(s) of the treatment relative to the counterfactual.** This involves comparing outcomes where the treatment was applied versus outcomes under the counterfactual, and assessing whether any differences can be distinguished from random noise in the data. Such comparisons may be straightforward or may need to address various technical issues, depending on the type of data available to do the analysis.

It is critical that measurements of the outcome(s) of interest are available for a sufficiently large number of observational units to allow any treatment effects to be distinguished from noise (i.e., unexplained variation) in the data. In addition, the following questions will shape the analysis that is possible:

- Is there data on outcomes for a well-defined control group of observational units for which the treatment was not applied? If so, do outcomes for this group accurately represent the counterfactual scenario for the units that received treatment? If not, can a robust synthetic counterfactual be estimated using a statistical model?

- Could the probability of receiving treatment and/or the outcome have been affected by any observed or unobserved\(^2\) characteristics of the units?

- Are observations of the outcomes for each unit available both before and after treatment, or only after treatment?

- Are the characteristics of units in the treatment and control groups sufficiently similar or very different?

- Aside from the outcomes, what data is available to measure characteristics of the observational units?

---

\(^2\) Unobserved here means that a variable is not captured in the data, because it is unmeasured or unmeasurable.
We return to these practical issues throughout the report and discuss how they affect the analysis of treatment effects. The issue of counterfactuals is of primary importance so we discuss this further in section 2.2 below. We then review the overarching framework for empirical estimation of treatment effects, known as the ‘potential outcomes’ model, and briefly review estimation of treatment effects in the ideal case where the data has been generated by a randomised, controlled experiment. Well-designed experiments use random treatment assignment to create a robust counterfactual and allow causal effects to be estimated relatively easily. While it may not be possible to carry out a randomised experiment for practical reasons, this gives a benchmark against which other methods for causal inference with non-experimental data can be compared.

2.2 The importance of counterfactuals

The effects caused by an intervention cannot be estimated without a counterfactual against which post-intervention outcomes are compared. The counterfactual represents the state of the world without the intervention, and the effects caused by the intervention are estimated using differences between actual and counterfactual outcomes. Without a counterfactual, the best we can do is to describe the state of the world after an intervention, but it is not possible to reach any conclusions about the effects of the intervention.

The difficulty is that a counterfactual cannot be observed directly – it is not possible to observe outcomes for a given place or group with and without an intervention at the same time. The counterfactual must therefore be estimated using data and statistical models, and some assumptions are required when comparing the counterfactual with post-intervention outcomes.

As mentioned in the introduction, it can be tempting to rely on simplistic counterfactuals, eg:

- Use observed outcomes where the intervention is applied prior to its application as outcomes under the counterfactual, ie conduct a ‘before and after’ comparison, or
- Use observed outcomes where the intervention is not applied after its application as outcomes under the counterfactual, ie a simple post-intervention comparison of a ‘treatment’ group and a ‘control’ group.

Such simplistic counterfactuals may be reasonable if conditions are relatively static, ie if other factors aside from the intervention are expected to have minimal impacts on outcomes over time, or if we are confident the characteristics of the treatment and control groups are similar. Aside from these special cases, or unless special care is taken in the analysis, both approaches can run into problems that lead to over- or under-estimation of the impacts of the intervention. This report is largely concerned with techniques that can be used to overcome such problems.

Having said this, naïve comparisons are often used as a starting point for causal analysis, as it is useful to see how the results change when more sophisticated methods are used.

2.2.1 Issues with ‘before and after’ comparisons

The first type of counterfactual described above is used in ‘before and after’ comparisons, where it is assumed any changes in outcomes pre- and post-intervention are due to the intervention. This is problematic for two main reasons:

1. It does not allow for other factors, unrelated to the intervention, which may have affected outcomes over the same period. Unless the effects of such factors can be identified and removed, the effects of the intervention could be significantly over- or under-estimated.
2. It is difficult to be sure the effects of the intervention are not specific to where it was applied, i.e., the results of a before and after comparison are difficult to generalise because the observed impacts of an intervention may depend on the characteristics of the places or things to which it was applied.

The first issue is a problem for all before and after comparisons. The effects of other factors can be controlled for in the analysis to the extent these factors can be observed. However, it is often not possible to observe all factors that may have affected outcomes of interest. Overcoming such problems requires more sophisticated analysis, discussed later in this report (see chapters 4, 5 and 6).

The second issue is only problematic if the results of a before and after comparison need to be generalised. For example, if a road safety intervention involving reduced speed limits was trialled in certain areas, it would not be reliable to use a before and after comparison for those areas to estimate the effects of applying the same intervention in other areas. However, if we are only interested in the impacts of an intervention specific to where it was applied, this is not a concern. For example, if we want to estimate the effects of a motorway extension, and the analysis will not be generalised to any other extensions, then a before and after comparison may be useful, provided the impacts of other factors can be sufficiently controlled for as discussed above.

2.2.2 Issues with simple ‘treatment and control’ comparisons

Another simple approach to constructing a counterfactual involves using observed outcomes for a ‘control’ group of units (e.g., places, or groups of people) where the intervention was not applied. The effect of the intervention is estimated by assuming outcomes for the control group are a good estimate of outcomes in the treatment group if the intervention had not been applied.

Such comparisons can also be problematic. There may be systematic differences in characteristics of the treatment and control groups that do not make outcomes for the control group a very good counterfactual for the treatment group. The treatment group’s characteristics may also amplify or dampen the effects of the intervention, so the estimated impact of the intervention cannot be generalised.

These problems can be overcome if relevant characteristics of the treatment and control groups can be observed and if the treatment and control groups are sufficiently similar in their characteristics. However, it may be the case that the treatment group was selected using characteristics that cannot be observed, and/or the compositions of the treatment and control groups are quite different. In such cases, a simple comparison of post-intervention outcomes between the treatment and control groups can lead to over- or under-estimating the effects of the intervention.

2.3 Policy context for causal inference analysis

Causal inference analysis seeks to identify the causality of effects. This is of interest to policy makers asking whether regulations, policies or projects: (1) have been effective, i.e., have produced the results expected; and (2) have had net positive impacts on wellbeing. Although simple before-after analysis is of interest (policy makers are concerned ultimately with improvements in wellbeing regardless of how it happened), it is also useful to know that government interventions have contributed to wellbeing improvements.

2.3.1 Regulatory and policy analysis

The NZ Productivity Commission (2014) in its review of regulatory practices in New Zealand remarked that ‘New Zealand does not have strong processes for reviewing regulatory regimes, leading frequently to a ‘set and forget’ mindset to regulation’. The New Zealand Government had agreed to a set of expectations for regulatory stewardship in 2013, including how departments should monitor and review regulations to ensure
they remain fit for purpose (NZ Treasury 2013). Such reviews include questions on whether regulations are still justified, eg because of possible changes in the nature of the initial problem being addressed, in addition to an examination of their effectiveness and the benefits they have produced.

Denne and Wright (2016) reviewed the use of ex-post analysis in regulatory analysis internationally. An OECD analysis of the use of ex-post analysis of regulations in member countries found that 17 countries plus the European Union (EU) had a mandatory requirement for ex-post evaluation of at least some existing primary laws, but New Zealand was among another group of 17 countries that had no requirement for ex-post evaluation (OECD 2015). For the countries that undertake ex-post regulatory analysis, the analytical requirements are set out largely in process terms and include requirements for analysis of effectiveness, costs and benefits. There is no evidence of particularly sophisticated analytical tools.

The techniques discussed in this report include those which would enable evaluation of regulations. For example, in chapter 6 we use a selection of causal inference techniques to evaluate a case study intervention: safer speed areas in which speed limits were reduced from 50km/h to 40km/h on 10 streets in eight different residential areas in Hamilton city in 2011. The limits were changed with the objective of reducing the number of serious crashes.

The analysis of this intervention was more straightforward because it applied to some streets only. It is more difficult to use the more sophisticated techniques where the regulation applies everywhere, eg if speed limits were reduced to 40km/hour in all residential streets throughout New Zealand. In these circumstances, it is more difficult to establish a counterfactual against which the effects are compared and also more difficult to identify causal effects.

In comparison to before-after analysis for which time series data is required, for these other techniques the key characteristic for analysis is that a counterfactual can be produced in the current time period, ie that there are some streets, vehicles or people to which the intervention applies and some with similar characteristics to which it does not.

2.3.2 Project appraisal

The other context for the use of causal inference techniques is in the appraisal of projects, eg in the analysis of the effectiveness or net benefits of new roads, public transport services or other projects. As with regulatory analysis, the key requirements for developing a counterfactual are the identification of comparator sites or services. These do not have to be the same in all respects but they do have to have some similar characteristics.

2.4 Measuring treatment effects and the fundamental problem of causal inference

The ‘potential outcomes’ model is the foundation of the empirical causal inference techniques we discuss in this report. This model formalises how we measure treatment effects and captures the so-called ‘fundamental problem’ of establishing causality.

Suppose we are interested in an outcome $y$ which is measured on observational units indexed by $i$, eg crash rates ($y$) on particular sections of state highways ($i$). Suppose that each unit is either subject to a treatment $T$ or not, eg reduced speed limits, ie $T_i = 1$ for units receiving the treatment and $T_i = 0$ otherwise.

---

3 The potential outcomes model is also known as the Rubin causal model, referring to the work of Donald Rubin in the 1970s to formalise this approach to causal inference (Rubin 1974).
For unit $i$ we can define potential outcomes $y_{i0}$ and $y_{i1}$ under the control and treatment respectively, and the treatment effect for a particular unit is $y_{i1} - y_{i0}$. The average treatment effect (ATE) is the expected value of the treatment effect across all units, ie $\text{ATE} = E[y_{i1} - y_{i0}]$. It is also possible to calculate the average treatment effect on the treated units (ATET), ie the expected value of the treatment effect for units that do receive the treatment.

The difference between the ATE and ATET can be understood by thinking about outcomes for a randomly selected unit. The ATE is the expected difference in outcomes if a randomly selected unit could be subjected to both the treatment and control conditions. The ATET is the expected difference in outcomes if a randomly selected unit in the treatment group could be subjected to both the treatment and control conditions.

In many cases, the ATE is the estimate of the treatment effect we are interested in as we want to know the general impact of the treatment across all potential units that could be treated. However, ATET may be of interest if we are estimating the effects of specific interventions or programmes that will not be repeated. In such cases we may be mostly interested in the difference in outcomes for units that did actually receive the treatment.

The ATE or ATET may be quite different to the treatment effect experienced by any individual unit receiving the treatment. Predicting the treatment effect for a specific individual (prior to receiving treatment) is considerably more difficult than estimating the average effect. The techniques we review in this report all focus on average effects.

The so-called fundamental problem of causal inference is that for any unit $i$, we can only observe $y_{i1}$ or $y_{i0}$ but not both. This means that causal effects for an individual unit, or on average across units, can never be measured directly because we cannot calculate $y_{i1} - y_{i0}$ for any unit $i$. In other words, for any given unit we can observe the actual outcome given whether the unit received the treatment or not, but we cannot observe the potential outcome in the other case, even if the data was obtained from a randomised experiment. Instead we must make assumptions that allow us to estimate treatment effects from what can be observed about each unit. These assumptions enable us to define a counterfactual and estimate a treatment effect.

While it is not a focus of this report, we also note the potential outcomes model can be extended to include non-binary treatments. In such cases, we may be interested in the average potential outcome (APO) of units for a given ‘dose’ level of an intervention. The APO will generally be a function of the level or ‘strength’ of the dose, and causal inference in such cases seeks to estimate that function. We briefly review methods for estimating APOs in section 4.8.

2.5 Randomised experiments

A well-designed randomised experiment is generally considered to be the ideal way to solve the fundamental problem of causal inference and produce robust estimates of treatment effects. In a randomised experiment, observational units are randomly assigned to receive the treatment or to be part of the control group. Outcomes observed in the two groups are compared, and outcomes for the control group are assumed to establish a counterfactual for the treatment group.

The basic idea of a randomised experiment is that it allows comparison of outcomes for ‘similar’ units in the treatment and control groups. The outcomes for units in the control group are assumed to represent outcomes that would have occurred if units in the treatment group had not received the treatment. Thus the control group establishes the counterfactual.
2.5.1 Estimating treatment effects

If the assignment of units to the treatment and control is truly random, any other factors aside from the treatment that also affect outcomes will ‘cancel out’ across the two groups, provided the groups are sufficiently large. If there are no other complications, the average treatment effect can be estimated as

\[ ATE = \text{avg}(y_i^1) - \text{avg}(y_i^0), \]

i.e. the difference between the average outcome calculated across units receiving the treatment and the average outcome calculated across units in the control group.

In addition:

- If there are other observed factors aside from the treatment that affect outcomes and/or the probability that a unit is selected for treatment, then these factors need to be accounted for in the analysis to estimate the treatment effect accurately. This can be done by estimating the average treatment effect using a simple regression model that includes other factors such as explanatory variables (see section 4.1 below).

- However, if there are other unobserved factors that affect both the probability of treatment and the outcome, estimation of the treatment effect is no longer straightforward. Such factors cause bias and can lead to incorrect estimates of treatment effects. This should not occur in a truly randomised experiment, but is a common problem in observational studies.

These cases are illustrated in figure 2.1, and we will return to them later in this report (see chapters 4, 5 and 6) when we discuss techniques for estimation of treatment effects in detail.

Factors that affect both the probability of treatment and the outcome of interest are called confounding covariates because they are related to the treatment (i.e. they co-vary with treatment status) and because their effect on outcomes is mixed up (confounded) with the effect of the treatment. All causal inference techniques, whether applied to data from randomised experiments or observational data, rely on an assumption that there are no unobserved confounding covariates. This is equivalent to the assumption of conditional independence defined below.

Figure 2.1 Factors to consider in causal inference analysis

2.5.2 Assumptions for estimating treatment effects from experimental data

More formally, the following assumptions must be satisfied to estimate an average treatment effect of a binary treatment from a straightforward comparison of outcomes across treatment and control groups:

1. **Conditional independence.** Conditioned on the observed characteristics, $X_i$, whether or not unit $i$ receives treatment and its potential outcomes under both the treatment and control are independent. Formally, $y_{i0}, y_{i1} \perp T_i | X_i$ for all $i$, where $\perp$ indicates statistical independence. This will be satisfied if treatment is truly randomly assigned or is assigned based only on observed characteristics of the units that can be controlled for in subsequent analysis (ie variables included in $X$).\(^5\)

2. **Common support.** For any given set of observed characteristics, $X_i$, there is some non-zero chance that unit $i$ would be assigned to the treatment or to the control group. That is, $0 < pr(T_i = 1 | X_i = x) < 1$ for all $x$. In other words, the ranges of characteristics encompassed by units in the treatment and control groups overlap. Outside the region of overlap, additional assumptions are required to estimate causal effects.

3. **No interference between units.** Assignment of one unit to the treatment or control does not affect the outcome of any other unit. In other words, there are no interaction or ‘general equilibrium’ effects between units. This assumption may need careful consideration in transport applications where units may be physically close to each other and/or connected in a network, so treatment effects may spill over to other groups.

2.5.3 Practical limitations of experiments

In many cases of transport interventions (or social policy more generally), it is not possible to conduct a purely randomised, controlled experiment. There may be ethical, legal, or practical barriers to random assignment. In addition, policymakers may prefer to target interventions only to areas where they are needed most (eg road safety improvements targeted at high crash-rate areas), so assignment cannot be truly random.

To some extent these issues can be dealt with in analysis, by making additional assumptions, and much of the rest of this report focuses on causal inference in such situations. However it is important to recognise that these additional assumptions reduce the confidence we can have in the estimates of treatment effects, relative to a well-designed randomised experiment. This is especially true where the assumptions are untestable.

2.5.4 Using experimental design to guide analysis

Although a truly randomised experiment may not be possible to analyse a particular intervention, it is often useful to imagine the ideal experiment that would estimate the causal effect of interest. By comparing the actual data that is available with what would be generated from the ideal experiment, we are forced to consider issues such as:

---

4. This is also called the ‘unconfoundedness’ assumption, the ‘ignorability’ assumption, the ‘selection on observables’ assumption, and the ‘exogeneity’ assumption in the literature.

5. If treatment is truly randomly assigned, outcomes for any individual unit will be unconditionally independent.

6. This is also called the ‘overlap’ assumption in the literature. The combination of the common support and conditional independence assumptions is also known as the ‘strong ignorability’ assumption.

7. This is also called the ‘independent and identically distributed’ sampling assumption or the ‘stable unit treatment value assumption’ (SUTVA) in the literature.
• What exactly is the control condition? If it is a continuation of the status quo, what does that entail? Is a ‘do nothing’ scenario a reasonable counterfactual, or will some default intervention occur anyway? This affects the interpretation of any estimated treatment effects, ie, what these effects are measured relative to.

• What is the treatment we would apply to estimate the causal effect of interest and how similar is the actual intervention to this ideal treatment?

For example, if we were investigating the effect of building urban protected cycle lanes on cyclist safety, what would be the optimal experimental design if no practical, legal, or moral issues existed? We would want to randomly assign the construction of new cycle lanes to a number of urban roads throughout the country, and randomly select a similar number of non-intervention roads to act as controls. Other factors influencing road cyclist safety would have to be held constant, thus we may want to halt any area-specific road safety campaigns, ensure every cyclist within the studied areas wears the same safety apparel, and check traffic flows and speed limits are, on average, similar for both treated and control roads. In reality such things may not be possible and we should think carefully about what the control condition actually is in data produced by an observational study.

It is also plausible that building cycle lanes influences general road safety behaviour. In other words, there may be more than one treatment at work, eg the physical separation of driver and cyclist, and the heightened awareness of other road users. Observational data probably cannot distinguish these treatments, and thus any estimated effect of cycle lanes on cyclist safety will be the combined effect of physical separation and other induced effects such as changes in awareness. All of these issues affect the interpretation of the results of analysis of observational data.

2.5.5 Quasi-experiments

If pure random assignment of units to the treatment and control groups is not possible, it may be possible to set up a quasi-experiment by careful choice of the way the intervention is implemented. For example, before a policy is applied to all geographic areas, it could be trialled in selected areas chosen randomly or based on characteristics that can be observed. This can mitigate concerns about random assignment while still defining treatment and control groups for estimating treatment effects.

2.6 Summary of practical issues in ex-post analysis

It is often the case that ex-post estimation of causal effects was not considered when policies and interventions were originally designed and/or implemented. This means empirical analysis of causal effects must be performed using the data that happens to be available. In transport applications, this may be problematic due to:

• **Non-random assignment**: Interventions may be targeted at locations or assets where current or future problems have been identified. This can be accommodated in analysis of treatment effects if appropriate data is available or if factors affecting the decision to intervene do not also affect the outcome of intervention. However, if unobserved factors affect both treatment and outcomes then estimating treatment effects is more difficult.

• **Inadequate control**: Non-random assignment or other practical issues may also mean there is little or no overlap in characteristics between units in the treatment and control groups. This means stronger assumptions are required to estimate causal effects, ie we must extrapolate beyond the range of observe characteristics included in the control and/or treatment groups. In general this reduces the robustness of causal inference.
• **Lack of a control group:** Some interventions (eg broad policy changes) may be applied to all potential observational units simultaneously, eg to all geographic areas, or all users of the transport network simultaneously. In such cases, it is not possible to define a control group, other than to use data on outcomes prior to the intervention. Causal inference in such cases requires making strong assumptions about the factors that may have affected outcomes over time.

### 2.7 Predictive versus causal inference

Finally, it is important to understand the differences between predictive and causal inference. Predictive inference seeks to find relationships between variables that have predictive power but these may or may not embody causal relationships. It is possible to find a statistically significant predictive relationship between two variables but where changes in one variable do not cause changes in the other. It is also possible to be unable to find a statistically significant predictive relationship between two variables when changes in one variable do cause changes in the other. Either of these cases occurs when there is a third variable that also plays a role in determining values of the two variables of interest. Distinguishing causal effects from such other factors makes causal inference more challenging than predictive inference.

Generally, questions that motivate ex-post analysis in transportation studies are causal in nature (Karwa et al 2011). Predictive inference is useful where we need to predict or forecast new or future observations, such as annual revenues raised from fuel taxes, or to predict the ex-ante benefit of a road safety promotion in a specific location (Shmueli 2010). In these cases, estimating an outcome that is as close as possible to the true outcome is the main priority, and whether any model used to make these predictions reflects true causal effects is not important. In contrast, in causal inference we are primarily interested in finding true causal relationships with a given outcome, eg how enhanced road lighting will change the average number of night-time crashes at locations where it is installed.

### 2.8 Multi-valued and continuous treatments

The above discussion and most of the rest of this report is concerned with the case of binary treatments. It is also possible to analyse treatments that have different discrete ‘levels’ or ‘doses’, or which vary continuously. The issues involved with causal inference in these cases are conceptually similar but the objective is to estimate treatment effects for each level of the dose (for multi-valued treatments) or to estimate a continuous function relating dosage to outcomes (for continuous treatments). This may be important in transport applications – for example we may wish to understand the impacts of varying levels of investment in additional transport capacity, or changes to public transport services.

With multi-valued and continuous treatments, potential outcomes $Y_i(d)$ are defined for each dose, $d$, of the treatment, where $d$ either takes on a fixed number of values or is a continuous variable. The control group receives $d = 0$ and thus the causal effect for an individual unit $i$ is $Y_i(d) - Y_i(0)$.

As with binary treatments, we usually aim to estimate the average treatment effect rather than the effect for any individual unit. With multi-valued or continuous treatments, the ATE is not constant and depends on $d$, ie $ATE(d) = E[Y_i(d)] - E[Y_i(0)]$. Causal inference seeks to estimate this relationship. A set of assumptions conceptually similar to those described in section 2.5.2 is required for valid causal inference in the case of multi-valued and continuous treatments (see Graham 2014 for a summary).

Some examples of causal inference for multi-valued and continuous treatments are discussed in section 4.8.
3 Ex-post evaluation in practice

This chapter examines ex-post evaluation schemes used by national transport authorities, including the Transport Agency in New Zealand and international examples (England, France, Norway and Australia). There is little evidence that robust causal inference methods are regularly used in the project evaluation schemes we have reviewed. Instead, transport authorities tend to focus on comparing predicted against actual impacts. Graham (2014) suggests this might reflect a greater interest in predicting how future investments will fare than in evaluating how well resources have been allocated in the past. Nevertheless, valuable insight can be gained through examining other elements of these schemes, such as project prerequisites for ex-post evaluation, evaluation frequency, methodology used and outcomes measured.

3.1 New Zealand

The Transport Agency conducts PIRs of a selection of projects each year. These aim to:\(^8\)

- Assess and explain how well projects and packages have achieved their main expected transport benefits (such as improved safety, increased capacity to meet traffic volume growth, or improved journey time and reliability) to give an overall assessment of the value for money of completed projects or packages reviewed.
- Explain any variation between actual results and expected benefits and costs.
- Identify lessons learned that can be used to make business improvements.

The analyses include comparison of pre- and post-project outcomes, eg crash rates, and they also include analysis of trends in control groups to correct for trends in the counterfactual. Generally, control groups are other routes not affected by the intervention, or outcomes relating to the wider region. However, the latter is generally not a very robust counterfactual as control outcomes must be fully independent of the intervention.

Statistical techniques are sometimes used in the evaluation process, for example, in the analysis of a recent project (the State Highway 2 – Dowse to Petone upgrade) to improve safety and reduce crash rates, the Transport Agency (2014) notes that crashes tend to vary randomly over time. Given this, it applies a Poisson distribution (mean = variance) to the data to define a change in crash rates that would be regarded as significant at the 90% confidence level.\(^9\)

In general, PIRs include the fundamental components of a robust ex-post analysis: consideration of a counterfactual (through use of control groups), and statistical tests to discern whether an outcome is due to random variation of the data or the intervention itself. However, PIRs give far less attention to other factors affecting the outcome than the causal inference examples discussed in the literature review in chapter 4. Still, a review of a Transport Agency enhanced PIR in section 3.1.1 provides evidence of best practice causal inference.

---


9 A Poisson distribution applies when: (1) the event is something that can be counted in whole numbers; (2) occurrences are independent, so that one occurrence neither diminishes nor increases the chance of another; (3) the average frequency of occurrence for the time period in question is known; and (4) it is possible to count how many events have occurred in a time period, but meaningless to ask how many such events have not occurred. Poisson distributions can be used in regression analysis, see De Paola et al (2010) in section 4.6.3.2 for an example.
3.1.1 NZ Transport Agency enhanced PIR - Northern Gateway toll road

In 2012 the Transport Agency contracted Infometrics and Azimuth to conduct an Enhanced Post-Implementation Review (EPIR) of the Northern Gateway toll road, otherwise known as the ALPURT (Albany-Puhoi Realignment) B2 road (Stroombergen and Barsanti 2012). Their analysis investigated the impacts of ALPURT B2 on traffic volumes, travel times, employment, property prices and crashes.

It is clear that Stroombergen and Barsanti carefully considered factors that could lead to incorrect estimation of the treatment effects. For example, they acknowledge that treatment effects estimated via a time series analysis could include the effect of changes in other factors that occur simultaneously (e.g. economic downturn), and treatment effects derived through cross-sectional data may include the impact of other events occurring at the same location. Stroombergen and Barsanti employ different causal inference techniques according to the data available for each outcome variable and their required explanatory variables. The three primary models applied are:

- **ordinary least squares (OLS) regression** - a standard regression model, used to estimate the impact on all outcomes. Explanatory variables used include a 2008 dummy variable to control for the effect of the economic downturn, a time trend and a dummy variable to identify the change in pre- and post-intervention outcomes (i.e. treatment effect).

- **generalised least squares regression** - this adjusts the model error term to account for correlation between residuals (that create bias in standard regression models). In this case, it was used to control for autocorrelation\(^{10}\) and spatial autocorrelation\(^{11}\).

- **difference-in-differences** - employed when a control group of roads/sites with similar characteristics was available.

Stroombergen and Barsanti (2012) also point out that different traffic levels at the different locations mean small variations at busy sites can overwhelm variation at quieter sites. The technical term for this phenomenon is heteroscedasticity, and it means the dispersion of residuals around the fitted regression line increases or decreases along a range of explanatory variable values. If the model does not take this uneven variability of residuals into account, statistical tests of significance can be invalidated. To overcome this potential issue, the evaluators employ models with heteroscedasticity-robust standard errors. However, robust standard errors cannot correct for bias due to mis-specified models or omitted variables.

The key findings from this EPIR are:

- a shift in annual average daily traffic (AADT) off the original route (SH17) and onto ALPURT B2 of around 12,000 to 18,000. There is no evidence of induced traffic from ALPURT B2

- a significant reduction in travel times, with average weekday savings of 14.5 minutes and holiday (weekend) savings averaging 22 minutes

- there is no strong evidence of ALPURT B2 generating additional employment or affecting local property prices

- the number of crashes is too low to allow robust inferences regarding the effects of ALPURT B2.

---

\(^{10}\) It is common for the errors (residuals) of models estimated using time series data to be correlated over time. This violates the regression assumption of errors being independently distributed.

\(^{11}\) Spatial autocorrelation occurs when observations made at different locations in the regression model are not independent. This also violates the regression assumption of errors being independently distributed.
3.2 International examples

Nicolaisen and Driscoll (2016) provide a comprehensive review of international practices and procedures of ex-post evaluations in the transport sector. Their analysis focuses on OECD countries, to ensure planning contexts within the studied jurisdictions are somewhat comparable, and to limit potential data or material availability issues. Parallels can be seen between the findings of Nicolaisen and Driscoll and the OECD (2015) relating to ex-post evaluation of regions. Nicolaisen and Driscoll’s main findings include:

- a significant difference in quality, coverage and comprehensiveness of the ex-post schemes
- data archiving and retrieval is a widespread problem for national transport planning agencies
- counterfactual approaches are varied, and often not addressed sufficiently.

The third finding implies that proper causal inference methods are rare in ex-post evaluation in transport. For instance, Nicolaisen and Driscoll reviewed eight systematic ex-post evaluation schemes in the UK, and note that only post opening project evaluation (POPE) of major schemes attempts to measure counterfactual causal mechanisms. The authors also note the ‘absence of consideration and communication of uncertainties for both endogenous and exogenous variables, such as economic growth rates, changes in travel behaviour, changes in policy, changes in fuel prices, [and] changes in transit fares’ (p22).

A summary of schemes assessed by Nicolaisen and Driscoll is presented in table 3.1. Detailed case study analysis is only applied to schemes within the UK and Norway, as these are perceived to be two of the better ex-post evaluation schemes currently in place. For example, only these two jurisdictions carry out a periodic meta-analysis of individual ex-post project evaluations (table 3.1).12 Nicolaisen and Driscoll perceive meta-analyses to be critical to the development of comprehensive datasets and to the identification of evaluation strengths and weaknesses.

3.2.1 England

Highways England undertakes POPE for major schemes (over £10m) and smaller local network management schemes (LNMS) under £10m. They involve an initial collection of pre-scheme baseline data, and ex-post analyses one and five years post implementation for major schemes (Finch 2015) and one year post implementation for LNMS. The evaluations compare ex-ante forecasts with ex-post results for outcomes such as journey times, congestion, journey time reliability, environmental impacts (air quality, noise) and safety.

Additional to these evaluations are independent ‘meta reports’. These are published biannually to identify common themes in evaluation data and examine the relationship between predicted and observed impacts and to compare the performance of different types of schemes across different regions (Highways England 2013).

3.2.2 France

In 1982, France introduced the Loi d’Orientation des Transports Intérieurs (LOTI), requiring all major transport projects to undergo ex-post analyses five to 10 years after the infrastructure came into operation (Worsley 2015). LOTI ex-post evaluations are compared with two sets of ex-ante numbers: estimates derived from an initial appraisal and re-evaluated estimates calculated after a mandatory public enquiry. Outcomes analysed relate to economic, social, environmental, safety and health outcomes (Chapulut et al 2005). Meunier (2010) found the average cost difference between the initial ex-ante

---

12 In this context, meta-analysis is analysis based on the results of multiple ex-post evaluation exercises.
analysis and the ex-post analyses is around 24%. However, this difference is only 10% for ex-ante estimates carried out after the public enquiry stage.

Table 3.1  Governmental ex-post transport project evaluation schemes

<table>
<thead>
<tr>
<th>Country</th>
<th>Transport sector</th>
<th>Title</th>
<th>Sampling</th>
<th>Time span</th>
<th>Evaluation years</th>
<th>Meta analysis</th>
<th>Projects evaluated</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Zealand</td>
<td>Road</td>
<td>Post implementation reviews (PIRs)</td>
<td>Sample of projects between NZ$0.5m–30m</td>
<td>2008 to present</td>
<td>1–3 years post opening</td>
<td>No</td>
<td>49</td>
</tr>
<tr>
<td>England</td>
<td>Road</td>
<td>Post-opening project evaluation (POPE) of major schemes</td>
<td>All projects over £10m</td>
<td>2002 to present</td>
<td>1 and 5 years post opening</td>
<td>Yes</td>
<td>75</td>
</tr>
<tr>
<td>England</td>
<td>Road</td>
<td>POPE of local network management schemes (LNMS)</td>
<td>All projects between £1–10m</td>
<td>2000 to present</td>
<td>1 year post opening</td>
<td>Yes</td>
<td>50</td>
</tr>
<tr>
<td>Scotland</td>
<td>Road</td>
<td>Trunk road infrastructure evaluation</td>
<td>All projects over £5m</td>
<td>2005 to present</td>
<td>1 and 3–5 years post opening</td>
<td>No</td>
<td>18</td>
</tr>
<tr>
<td>France</td>
<td>Road, rail, sea, and air</td>
<td>Loi d'Orientation des Transports Intérieurs (LOTI)</td>
<td>Projects over €82m</td>
<td>1984 to present</td>
<td>3–5 years post opening</td>
<td>No</td>
<td>50</td>
</tr>
<tr>
<td>Norway</td>
<td>Road</td>
<td>Post opening of monetised impacts of major projects</td>
<td>Projects over 200kr (3–5 chosen per year)</td>
<td>2006 to present</td>
<td>5 years post opening</td>
<td>Yes</td>
<td>22</td>
</tr>
<tr>
<td>United States</td>
<td>Rail</td>
<td>Before and after studies of new starts transit projects</td>
<td>All projects funded by the Federal Transit Administration New Starts grant</td>
<td>2006 to present</td>
<td>2 years post opening</td>
<td>No</td>
<td>13</td>
</tr>
<tr>
<td>Australia</td>
<td>Road</td>
<td>Ex-post economic evaluation of national highway projects</td>
<td>Case studies</td>
<td>2007 to present</td>
<td>Not clearly defined</td>
<td>No</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: Nicolaisen and Driscoll (2016)

3.2.3 Norway

The Norwegian Public Roads Administration (NPRA) has developed an evaluation methodology for road projects whereby five projects per year are chosen to undergo an ex-post evaluation. Only impacts that
can be monetised are considered at the ex-post stage, including travel time savings, vehicle operating costs, crash costs, induced traffic, inconvenience cost (ferry projects), noise nuisance, local air pollution, road maintenance costs, residual value of capital, cost of public funds and road investment cost (Kjerkreit 2008). Outcomes excluded from these evaluations include impacts to: quality of life, the natural environment, visual landscape, outdoor recreation and overall accessibility. Omitting these factors from the analyses reduces the value of the ex-post evaluation to some extent (Nicolaisen and Driscoll 2016).

Both the ex-ante and the post-opening impacts are calculated through specific computer software to keep the methodologies consistent. NPRA’s primary objective of the ex-post assessment is to measure the accuracy of the information given to the evaluators at the time of the initial planning stage, and to find out whether the benefits are acceptably larger, smaller or the same as what was forecasted (Kjerkreit 2008). Meta-analytic techniques are used to derive these trends, just as they are done by Highways England.

Kjerkreit’s (2008) study of the NPRA’s evaluations concluded that the economic performance of road projects, in net present value terms, was greater than forecasted for seven of the eight projects studied. This can be explained by underestimated traffic growth in the forecasting stage.

3.2.4 Australia

Australia’s Bureau of Transport and Regional Economics (BTRE) began systematically reviewing major road investments in response to the Australian Transport Council’s National guidelines for transport system management (NGTSM) (BTRE 2007). The NGTSM outline best practice for transport planning and evaluation in Australia and were last updated in 2006. The impacts considered by BTRE’s studies are purely economic. The central objectives of these case studies are to check the accuracy of the ex-ante estimates, identify sources of differences in predicted and actual results, and to learn from these differences for the benefit of future ex-ante analyses.

A key finding from BTRE’s ex-post projects is that ex-ante evaluations tended to underestimate road user benefits, an issue also experienced by Norway’s NPRA. Simple extrapolation of past travel data in the ex-ante analysis may explain this discrepancy. BTRE (2007) concluded future traffic ex-ante evaluations would benefit from data that allowed for greater sophistication in traffic modelling.

3.3 Summary

Three central objectives common to all the ex-post evaluations above are mirrored by the aims of New Zealand’s PIRs (NZ Transport Agency 2015). Table 3.1 shows the timing of local and international post-opening evaluations are also similar. However, New Zealand’s PIRs are generally applied to smaller projects than transport sector ex-post analyses in other countries. PIRs also appear to have a higher evaluation frequency, for instance, despite being a relatively recent scheme, a high number of projects have been evaluated (table 3.1).

Nicolaisen and Driscoll (2016) highlight the lack of proper cause-effect estimation within international transport evaluation schemes. This observation is also relevant to New Zealand’s PIRs. However, some evaluations do use statistical models to infer cause-effect relationships between an intervention and an outcome (eg NZ Transport Agency 2014) but this is not common practice. New Zealand’s PIRs could also benefit from the use of meta-analysis to critique and assess the effectiveness of impact evaluation schemes. Meta-analysis is perceived to be a distinguishing characteristic of ‘best practice’ ex-post evaluations.
4  Overview of key techniques and literature review of empirical applications

This section gives an overview of the main statistical techniques used for causal inference analysis and reviews applications of these techniques in the literature, with an emphasis on transport applications.

Each technique requires specific assumptions that need to hold for causal relationships to be established. Roughly speaking, these assumptions are easier to satisfy if the available data are similar to what would be produced by a randomised experiment (HM Treasury 2011). In other words, we can be more confident about the results of causal inference applied to data that is as similar as possible to that generated by a randomised experiment.

We have not attempted to provide a comprehensive survey of the theoretical literature on causal inference techniques (see Gelman and Hill 2007; Angrist and Pischke 2009; Imbens and Wooldridge 2009; Imbens and Rubin 2015; Abadie 2005; Athey and Imbens 2017, to review some recent developments).

4.1  Causal inference with simple regression models

If the data allows it, causal effects can be estimated with relatively simple regression models using cross-sectional or panel data. In practice, particularly in transport applications where the data has not come from a randomised experiment, such techniques will not produce reliable estimates of causal effects unless all factors that affect the outcome of interest can be observed and included as explanatory variables in the regression model.

However, even if these techniques cannot be applied due to data limitations, they are useful for understanding how causal effects can be estimated in simple situations, and the issues that may arise when doing so. For these reasons, we start with a discussion of simple regression models, before reviewing more advanced techniques.

4.1.1  Regression with cross-sectional data

Regression with cross-sectional data – where variables are measured at the same point in time or without regard to differences in time – is a widely used statistical prediction technique. Despite its relative ease of application to predictive inference, basic regression models are difficult to use for causal inference because strict assumptions are needed to support a causal interpretation of estimated regression coefficients (Gelman and Hill 2007).

The most basic cross-sectional regression model that could be used for ex-post evaluation, requires data on post-intervention outcomes of both treated and controlled units and their treatment status (coded as a dummy variable taking the value 1 if treated and 0 otherwise). The impact of the intervention can be modelled through the following regression equation 4.1, where the coefficient $\beta_1$ represents the average difference in treated and non-treated post-intervention outcomes:

$$ Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + e_i $$

(Equation 4.1)

Where:  $Y_i$ measures the outcome for unit $i$
$\beta_0$ is a constant
$\beta_1$ is the estimated average treatment effect
$T_i$ is the treatment variable, coded 1 for treated units and 0 otherwise
\( \beta_2 \) is a vector of estimated coefficients for the variables in \( X \)
\( X_i \) is a vector of other observed characteristics of unit \( i \) that are expected to affect outcomes
\( e_i \) is a random error term with zero mean and constant variance

In the simplest case, where there are no other characteristics aside from treatment that affect outcomes (i.e., \( X_i \) can be omitted), estimating this regression model is equivalent to simply estimating the mean outcomes of the treatment and control groups. A simple t-test can then be used to determine whether there is a significant difference in the two means – this is equivalent to testing whether the estimated coefficient \( \beta_1 \) is significantly different from zero.\(^{13}\)

It is also possible to include interaction effects between the variables in \( X \) and the treatment variable \( T \). This should be done if it is believed the effect of the treatment depends on one or more characteristics of the units captured by the variables in \( X \). However, in such cases the interpretation of the treatment effect is more complicated, because the size of the effect depends on the characteristics of the unit receiving treatment. One way to interpret the treatment effect when interaction terms are included in the model is to calculate the treatment effect for the 'average' value(s) of the relevant characteristic(s).

Intuitively, the estimated effect of the treatment, \( \beta_1 \), can only be justified as ‘causal’ if the difference in outcomes between the treatment and control groups is solely because of the presence of the intervention, given the variables that are controlled for by the data in \( X \). In other words, to give a causal interpretation to \( \beta_1 \), we must be able to assume that assignment of units to the treatment or control groups is effectively random given the variables in \( X \). If that is true, then the conditional independence assumption is satisfied (see section 2.5.2 above). In addition, there must be sufficient overlap in the characteristics of the treatment and control groups so the common support assumption is satisfied.

The primary reason for questioning the conditional independence assumption is that treated and control units may be systematically different in ways that are not measured or are not included in the model for any reason (i.e., are not represented by \( X \) in the model above). This is likely when treatment is not randomly assigned but depends on one or more unobserved characteristics. Such characteristics are referred to as confounding covariates if they affect both the probability of receiving treatment and the outcome. If confounding covariates are omitted from the estimated regression model (i.e., they are unobserved), the estimated value of \( \beta_1 \) will be biased. This can lead to finding \( \beta_1 \) is statistically significantly different from zero when there is no causal relationship between the treatment and the outcome, or failing to estimate a statistically significant value of \( \beta_1 \) when a causal effect actually exists. The possible existence of confounding covariates is therefore an important issue that needs to be considered in any analysis of causal effects. In addition, if there are other characteristics or factors not included in \( X \) and that affect the outcome but do not affect the probability of receiving treatment, the estimate of the treatment effect will be relatively imprecise, although it is not biased.

As discussed earlier, these assumptions are more plausible if the treatment is randomly assigned and carried out within a controlled experiment, but experiments are seldom feasible for transport interventions. Raballand et al (2011) provide a rare example where fare prices for a bus service in Malawi were randomly allocated to examine the relationship price and service provider profit.\(^{14}\) All other

\(^{13}\) It should be kept in mind that such tests assume the treatment and control groups have equal variance and the estimate coefficients are approximately normally distributed. If these assumptions are not valid then an unequal-variance t-test or non-parametric test should be used instead.

\(^{14}\) Random allocation of fares to a single bus service allowed researchers to observe the effect of price on profit without having concern for selection bias that may occur if multiple bus routes of varying prices and characteristics were sampled instead.
applications of causal inference techniques we have identified in the literature are based on non-experimental data.

Even when it is possible to estimate a basic regression model of treatment effects (such as the model above), we may not be confident the conditional independence and common support assumptions are satisfied. This is because it is difficult to be sure there are no unobserved confounding covariates that have not been included in the model. Furthermore, in practice, it is common to find a lack of sufficient overlap in the characteristics of units included in the treatment and control groups.

The causal inference techniques discussed in most of the rest of this report attempt to overcome some of the shortfalls of applying simple regression models such as the one above to non-experimental data by further exploiting relevant data so assumptions underpinning estimation of a causal effect are more credible. However, the potential bias created by unobserved confounding covariates is also an issue for the more sophisticated causal inference methods discussed in this report. While some methods (eg difference-in-differences (DiD) and fixed-effects models) are able to cope with certain types of unobserved factors, it is never possible to be sure the conditional independent assumption is satisfied. Thus, evaluators must carefully consider whether the assumptions of the model they are using are reasonable given their understanding of the potential causal relationships they are analysing.

4.1.2 Regression with panel data

Panel data, also known as longitudinal data, consists of outcomes and other characteristics observed over two or more time periods for the same set of observational units. For example, the annual count of fatal crashes, traffic volume and average speed may be observed for different road sections (observational units) in two or more years. This type of data provides information on how outcomes and characteristics for each unit evolved over time, as well as information about how these things vary across units at each point in time. Thus, panel data combines features of cross-sectional and time-series data.

If panel data is available, a ‘fixed effects’ regression model can be estimated to control for potential unobserved factors that are fixed across units or across time, and to eliminate them as possible sources of bias in estimated treatment effects. For example, an unobserved variable may vary across units but remain constant over time (eg a region-specific attitude toward speeding), or be the same across units but vary over time (eg nation-wide shocks that affect road traffic volumes). Unit-specific fixed effects can be controlled for by including dummy variables for units. Likewise, time-specific fixed effects can be accounted for through including dummy variables for time periods. However, fixed effects models cannot control for omitted variables that vary both across units and over time, since this would introduce more variables into the model than can be estimated.

With panel data, it is also possible to control for pre-intervention observations of outcomes, if these are believed to influence the effect of the intervention. For example, we might hypothesise that existing crash rates at intersections influence the effectiveness of safety interventions applied to intersections. To control for this, we can include the pre-intervention crash rate as an explanatory variable in a panel data regression model. Obviously, it does not make sense to control for post-intervention outcomes.

---

15 In each case, the dummy variable for one unit or one time period needs to be omitted. Including the full set of dummy variables in a regression model will make the explanatory variables perfectly multi-collinear, which is known as the ‘dummy variable trap’, and prevents estimation of the coefficients.
4.1.3 Model specification

For simplicity, the discussion above uses a linear model that can be estimated using OLS. This type of model implicitly assumes the dependent variable ($Y_i$) is continuous and can take both positive and negative values. In some situations, this may not be an appropriate assumption, for example if the dependent variable is the number of traffic crashes at particular locations, then this will be discrete numbers and cannot be less than zero.

In such cases an alternative model specification such as a Poisson or more general negative binomial model can be used. These models assume the natural logarithm of $Y_i$ is related to a linear combination of the dependent variables, and $Y_i$ takes a Poisson or negative binomial distribution. For example, the cross-sectional regression model above can be re-written as:

$$\ln(Y_i) = \beta_0 + \beta_1 T_i + \beta_2 X_i + e_i$$

(Equation 4.2)

This is a log-linear model, and when combined with the assumption of a Poisson or negative binomial distribution for $Y_i$, such models cannot be estimated using OLS. Instead a numerical technique such as maximum likelihood (ML) estimation must be used, which is readily available in most statistical software packages. Statistical significance tests can also be performed on the coefficients of such models, using an appropriate distribution for the test statistic, i.e. a conventional $t$-test is not appropriate and a different distribution must be used for the test.

More generally, generalised linear models (GLMs) can be used to estimate models under other distributional assumptions. Generalised linear mixed models (GLMMs) take this one step further to allow for unobserved random components.

Fixed-effects models are also often estimated using the ‘first difference’ of the dependent and independent variables, and excluding the dummy variables described above. The differencing has the effect of eliminating any effects on each cross-sectional unit that are constant over time.

Despite the differences in functional form and estimation technique, the basic issues regarding estimation of causal effects discussed above remain the same, with one important difference - the effects of any unobserved confounding covariates that are constant over time will be captured by the fixed effects in the model and this eliminates such factors as a potential source of bias.

Despite this advantage, it is important to remember fixed-effects models cannot control for unobserved confounding covariates that are not constant over time, and such factors may be important in some cases. Furthermore, estimating fixed-effects models can be problematic when some explanatory variables do not vary much over time, as it is difficult to separate the effects of such variables from the fixed-effects (Mare and Graham 2013). Such highly persistent variables can be important in transport applications, limiting the potential use of fixed-effects models.

4.1.4 Weighted regression models

The simple cross-sectional models described in section 4.1.1 above are also known as ‘regression adjustment’ (RA) models of treatment effects, i.e. the regression model is used to ‘adjust’ observed outcomes from the treatment and control groups to generate counterfactuals.

---

16 A Poisson model makes the restrictive assumption that the mean of the dependent variable is equal to its variance. The negative binomial model relaxes this assumption.
An alternative approach is to model the probability that a unit is selected for treatment, based on its characteristics. If treatment/control selection is a binary outcome, this can be done with a simple discrete choice model such as a logit or probit model. Such models estimate the probability that a unit is selected for treatment given its observed characteristics. The resulting probabilities can be used to estimate treatment effects either directly or in combination with a cross-sectional RA model.

Inverse probability weighted and augmented inverse probability weighted methods involve calculating weighted averages of outcomes across the treatment and control groups, where the weights are estimated from a treatment model as described above, i.e., the weights are based on propensity scores (described further in section 4.2 below). The weighting seeks to adjust for differences in characteristics that determined treatment status of the treatment and control groups.

It is also possible to use the treatment probabilities to calculate weights that are used in a weighted RA model (this method is known as inverse probability weighted regression adjustment). This allows factors that affect treatment status as well as outcomes to be controlled for in the estimation of treatment effects. This method is sometimes referred to as ‘doubly robust’ because it combines a regression model with a propensity score model in such a way the treatment effect is consistently estimated if just one of the two component models is correctly specified.

These methods are similar to the technique of propensity score matching (PSM) that we discuss in section 4.2 below. In our view, PSM has some additional advantages over simple probability-weighted methods, arising from the use of a matching algorithm to improve the balance of characteristics between the treatment and control groups. PSM can be combined with regression models also. For these reasons, we recommend the use of PSM over probability-weighted methods where no matching is done. However, it is possible to include a matching step prior to estimating a probability-weighted model as described above. The results of such analysis should be quite similar to those obtained from PSM.

We should also stress that all RA-based methods are only able to control for observed characteristics of units that may affect treatment status and/or outcomes. In reality, unobserved factors are also likely to play a role, and more sophisticated methods need to be used to overcome the potential bias created by unobserved factors. We return to this issue in later sections.

### 4.2 Propensity score matching

#### 4.2.1 Overview

PSM provides an alternative to the simple regression methods described above to control for observed confounding covariates. PSM can be used to help overcome a lack of ‘balance’ or lack of sufficient ‘overlap’ in the characteristics of treatment and control units and can help improve the accuracy of estimated causal effects relative to a simple regression model.

PSM typically involves two phases. First, each unit’s probability of receiving treatment is estimated (a propensity score is assigned to each unit) using a model of the relationship between receipt of treatment and variables known or suspected to influence treatment status (Boarnet 2007). Second, treated and non-treated units with similar propensity scores are matched in some way, and their post-intervention outcomes compared. We can attribute the average difference in outcomes to be the causal effect of the intervention, provided that the conditional independence assumption holds, given the factors that are encapsulated by the propensity score. Most statistical software can carry out these steps via a single PSM function.
4.2.2 Assumptions

4.2.2.1 The conditional independence assumption

Like conventional regression models discussed above, the conditional independence assumption must hold to infer a causal relationship from analysis using PSM. Instead of using a matrix of control variables, the conditional independence assumption can be based on a single propensity score calculated for each unit:

\[ T_i \perp (y_i^0, y_i^1) \mid P(X_i) \]  

(Equation 4.3)

Where:
- \( y_i^0 \) and \( y_i^1 \) are the potential outcomes for non-treated and treated units respectively:
- \( T_i \) is the treatment variable
- \( P(X_i) \) is the propensity score for unit \( i \) given its characteristics \( X_i \)
- \( \perp \) indicates statistical independence and \( \mid \) indicates conditionality

In other words, PSM methods assume that the propensity score captures all potential confounding covariates so treatment and potential outcomes are assumed to be independent for any unit, conditional on that unit’s propensity score. As with cross-sectional regression methods described above, if there are unobserved confounding covariates, PSM methods will also fail to produce unbiased estimates of causal effects.

4.2.2.2 The common support condition

The common support condition, also known as the overlap condition, states the probability of being selected into treatment is greater than zero but less than one, given a set of pre-treatment characteristics (see section 2.5.2 above for a formal definition). This ensures units with the same set of characteristic values have a positive probability of being treated or not (Caliendo and Kopeinig 2008). For values of characteristics where there is no overlap between the treatment and control groups, we cannot make causal inferences without imposing some additional assumptions.17

A larger overlap or region of common support is preferred as it allows for more matches between the control and treated units. In a PSM approach, the characteristics of observational units are summarised by their propensity score. Figure 4.1 shows two examples of propensity score distributions, one with a small area of common support (left diagram) and the other with a large area of common support (right diagram).

Figure 4.1 Propensity score distribution examples

---

17 For examples, see the discussion of regression discontinuity models and interrupted time series analysis (see section 4.2).
Although the right side of figure 4.1 illustrates a relatively large area of common support, lack of complete overlap between the control and treatment groups means units falling outside the area of common support cannot be used for the analysis without making additional assumptions. Weak overlap, such as in the left side of figure 4.1, may lead to inaccurate treatment effect estimates and can make estimation sensitive to the choice of model specification (Crump et al 2006).

Balance in the characteristics of the two groups is also important for estimation of treatment effects. Both illustrations in figure 4.1 show a clear imbalance of propensity scores with significantly higher averages belonging to the intervention groups. Although the implications of imbalance are similar to those of lack of overlap, Gelman and Hill (2007) believe the latter is a more serious problem, as lack of overlap places greater limitations on the causal conclusions that can be made and/or more reliance on assumptions.

PSM can help to correct for problems caused by lack of overlap and lack of balance. Lack of overlap can be addressed by using propensity scores to identify ranges of characteristics where there are no units in both the control and treatment groups and excluding such units from the analysis. Lack of balance can be addressed in the way that units in the treatment and control groups are matched with each other.

A key advantage of the PSM approach is that a large number of characteristics of treatment and control units can be used in the estimation of the propensity scores. Alternative approaches to matching that directly use the characteristics of the units (without calculating propensity scores) are only feasible if the number of characteristics to be used in matching is small.

### 4.2.3 Implementation

The aim of the propensity score estimation stage is to collapse multiple characteristics into a single variable. Logit and probit models are the most commonly used for the calculation of propensity scores where there is a binary treatment. To obtain propensity scores, the treatment indicator is used as the dependent variable, while factors suspected to influence both treatment assignment and outcomes are used as explanatory variables. Variables used to explain propensity scores should not be affected by treatment status (ie free of reverse causality), therefore Caliendo and Kopeinig (2008) suggest propensity score explanatory variables should either be fixed over time or measured before participation.

After propensity scores are calculated, a variety of matching algorithms can be used for the process of selecting treatment and control units for comparison. One popular method, nearest neighbour matching (NNM), pairs treated and control units of the closest propensity score. This technique can be altered to restrict the number of times that a single unit can be matched with another. For example, NNM ‘with replacement’ places no restriction on the number of matches for a given unit, whereas NNM ‘without replacement’ limits a unit to just one match (Caliendo and Kopeinig 2008).

Another common matching technique is to weight each unit by the inverse of its respective propensity score, known as inverse probability weighting. Once the propensity score is calculated, the treatment effect can be directly estimated by calculating a weighted average of outcomes for the treated and control groups (Kopeinig 2008). For example, one would use weights of \(1/PS_i\) and \(1/(1 - PS_i)\) for treated and control observations respectively (Gelman and Hill 2007). These weights can be used to calculate simple sample averages or used to estimate a weighted regression model.

---

18 NNM can also be used to match units directly, based on minimising some ‘distance’ function calculated from the characteristics of the units, ie without the intermediate step of calculating propensity scores. However, this is difficult to apply if the number of characteristics is large. In addition, modelling propensity scores makes use of the information about the relationship between characteristics and treatment status. For this reason, the discussion that follows relates to NNM applied to propensity scores only.
Matching methods should be selected to optimise both balance and overlap between the treatment and control groups. Thus, ‘the goal of PSM is not to ensure that each pair of matched observations is similar in terms of all their covariate values, but rather that the matched groups are similar on average across all their covariate values’ (Gelman and Hill 2007, p207). For large sample sizes, matching algorithms will give similar results, but the choice of matching technique can have a considerable impact on the results for small samples (Li et al 2013).

Once the matching has been done, treatment effects can be calculated by averaging the difference in outcomes between the treated units and the matched units in the control group. Statistical software such as Stata and R can be used to do this at the same time as performing the matching. Bootstrapping\(^\text{19}\) methods can be used to calculate the sample variance of the difference in means, for performing a significance test. As discussed in section 4.1.3 above, it may be sensible to assume that outcomes for the treated and control groups follow a Poisson or negative binomial distribution, eg if they are crash rates that cannot take on negative values. This involves using an appropriate distribution for any test of statistical significance of the difference in outcomes between the two groups.

4.2.4 Strengths and weaknesses of PSM

When applying PSM it is important to be aware of its strengths and weaknesses (Khandker et al 2010). A key potential weakness is the calculation of propensity scores is only based on observed characteristics of units, and unobserved factors that affect treatment status may create bias. Another disadvantage, as mentioned earlier, is PSM is a ‘data hungry’ technique and may result in a significant number of observations being discarded if they do not have good matches. However, PSM can be applied to cross-sectional data, possibly in combination with a regression model, and does not require observations at different points in time.

4.2.5 Examples

PSM methods appear to be relatively rare in the transport evaluation literature. We discuss below some of the few transport examples we have identified.

4.2.5.1 UK: speed limit enforcement

Li et al (2013) evaluated the impacts of speed limit enforcement cameras on road crashes in the UK. The researchers used PSM to select appropriate control groups to adjust for confounding factors. Data pertaining to 771 treatment sites and 4,787 control sites was available for the analysis. Distributions of propensity scores for treated and control sites showed a strong area of common support (figure 4.2). Only seven treated sites fell outside the region of support and were discarded for the matching stage.

Four types of matching algorithms were used to ensure the estimation did not depend on the chosen algorithm. For each method, the treatment effect was estimated through applying a t-test on the difference in outcomes between the matched treated and control groups, with the assumption that crashes follow a Poisson distribution. All the algorithms provided a similar treatment effect estimate, an expected result given the relatively large sample size. The findings suggest speed cameras led to a

---

\(^{19}\) Bootstrapping is the repeated re-sampling ‘with replacement’ (ie sampling whereby an observation can be drawn more than once) from a subset of the population data (eg matched data) in order to estimate parameters belonging to the wider population (eg mean value and variation).
significant reduction in crashes and were most effective in reducing crashes up to 200m from camera sites.\textsuperscript{20}

\textbf{Figure 4.2} Propensity score distribution

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure4.2.png}
\caption{Propensity score distribution}
\end{figure}

Source: Li et al (2013)

\subsection*{4.2.5.2 Canada: crash rates and safety investment}

Park and Saccomanno (2007) used PSM to control for the fact that roads with high crash rates are targeted for safety investments. They hypothesised that roads are selected after periods of ‘higher than usual’ crash rates, attributable to the random nature of traffic collisions. Therefore, these crash rates would return to a lower long-term expected value even without intervention (a phenomenon known as regression-to-the-mean). Consequently, ‘conventional observational models ascribe this reduction in collisions to the given treatment. This results in overestimation of treatment effect’ (Park and Saccomanno 2007, p112).

To test their theory, Park and Saccomanno applied PSM to Canadian highway–railway crossing data to estimate the causal effect of warning devices on the number of collisions (assuming collisions follow a negative binomial distribution). Like Li et al (2013), the sample size was fairly large, including 9,874 crossings, 1,274 of which were upgraded in some manner (ie treated).

Results were compared with past studies that had analysed the same (or similar) data yet had not controlled for confounding factors. PSM was shown to reduce selection bias, revealing a lower causal effect than was measured previously, which confirmed the researchers’ original hypothesis. They acknowledged that potentially unobserved factors causing selection bias were not controlled for by propensity score methods, meaning their estimates were still likely to have some level of bias, although the direction of this bias is unknown.

\subsection*{4.2.5.3 California: highway investment and urban growth}

Funderburg et al (2010) examined the association between new highway investments and urban growth in three California counties. Each county was analysed separately, with geographic area units located within a certain distance of new highways defining the treated area. Control geographic area units were selected from neighbouring areas that fell outside this distance threshold. The nearest neighbour method,

\textsuperscript{20} In such studies, it is important to note that further roll out of the intervention (speed cameras) may not be as effective as for the initial sites, if those sites were selected because they have high crash rates.
combined with a propensity score tolerance level of 0.1 (ie a matching unit’s score must be within ±0.1 of the corresponding treated unit), was used to ensure high-quality matches between factual and counterfactual units. Outcomes examined included changes in population and employment growth.

A difference-in-differences (DiD) model (see below) was then applied to the matched data to estimate the treatment effect. Combining PSM with DiD allowed the authors to control for both observed confounders (through PSM) and unobserved factors that had the same impact on all areas (through DiD). Their findings were twofold. First, ‘while improvements in surface transportation tend to have large impacts on growth patterns, the nature of the effects is materially dependent on the context of the highway investment…such as the type of highway improvement… and characteristics of the location’ (Funderburg et al 2010, p94).

Second, this study proved the usefulness and validity of quasi-experimental methods (ie techniques that attempt to mimic the key elements of a randomised experiment) in examining growth impacts of modern highways.

4.2.6 Limitations

The key limitations of PSM methods are the strict assumptions and richness of data needed for robust causal inference. All confounding covariates must be recorded in the data for the conditional independence assumption to be satisfied via the propensity score (Caliendo and Kopeinig 2008). Second, matching can only be implemented when there is adequate overlap in propensity score distributions of treatment and control groups. If no sufficiently close match can be found for some treated units, these treated units will be discarded, which may reduce estimation accuracy. Hence, a large number of observational units may be required to ensure adequate matches. In summary, PSM is a ‘data hungry’ causal inference method.

4.3 Difference-in-differences

4.3.1 Overview

The DiD technique provides a solution to bias caused by unobserved confounding covariates that are either constant over time or vary over time but affect outcomes in the treatment and control groups in the same way at any given point in time.

The logic behind DiD is that, if the treated and control units have parallel trends, one group can act as a counterfactual for the other. Thus, by subtracting the change in outcomes for the control group from the change in outcomes for the treatment group, we can estimate the treatment effect even when there are unobserved factors that might cause bias in a simple regression analysis:

\[
\text{Treatment effect} = \text{average change in treated outcomes} - \text{average change in control outcomes}
\]

To illustrate DiD further, suppose each of the four outcomes (before and after the intervention for treated and control units) can be represented by:

- \(a\) common constant for all observations
- \(\beta\) the effect of attributes specific to the treated units, aside from treatment
- \(T\) the treatment effect
- \(t\) the effect of time on all units in both the treatment and control groups

Table 4.1 illustrates how DiD works with this notation. The first difference (between pre- and post-intervention incomes for the treatment and control groups) eliminates the effect of any factors that have a constant effect on either group over time (ie \(a\) and \(\beta\) are eliminated). The second difference (between the
difference in outcomes between the two groups) eliminates any factors that have a constant effect on both groups at any point in time (i.e., \( t \) is eliminated). The only effect remaining after the double-differencing is the treatment effect (\( T \)).

**Table 4.1 Illustration of the DiD technique**

<table>
<thead>
<tr>
<th></th>
<th>Pre- intervention outcome</th>
<th>Post- intervention outcome</th>
<th>Difference in outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated units</td>
<td>( a + \beta )</td>
<td>( a + \beta + T + t )</td>
<td>( T + t )</td>
</tr>
<tr>
<td>Control units</td>
<td>( a )</td>
<td>( a + t )</td>
<td>( t )</td>
</tr>
<tr>
<td>Difference in differences</td>
<td></td>
<td></td>
<td>( T )</td>
</tr>
</tbody>
</table>

Source: Buckley and Shang (2003)

DiD estimation is illustrated graphically in figure 4.3. This shows the assumption that outcomes for treated and control units share the same trend. Given this, we implicitly assume the treated units would follow a parallel path to that of the control in absence of the intervention. Therefore, by effectively extrapolating the control trend onto the treated, we can isolate the change in outcomes caused by the intervention.

Figure 4.3 Illustration of the DiD model

This illustration also shows the errors that would result from naïve comparisons. If we simply calculated the difference in outcomes for the treatment group before and after the intervention, we could estimate the treatment effect to be \( t + T \), which is biased by the unobserved time-varying factors represented by \( t \). Similarly, if we calculated the difference in post-intervention outcomes between the treatment group and control group, we would estimate the treatment effect to be \( \beta + T \), which is biased by the unobserved systematic differences between the treatment and control groups represented by \( \beta \). Only the double-difference eliminates both the impacts of \( t \) and \( \beta \).

If the data is in panel form, i.e., if there is a pre- and post-intervention observation of the outcome for each unit, we can estimate a treatment effect via DiD using a simple regression model where the difference between pre- and post-outcomes for each unit \( (Y_{it} - Y_{0i}) \) is regressed on a constant and a dummy variable \( (D_i) \) indicating whether or not the unit was treated:
Overview of key techniques and literature review of empirical applications

\[ Y_{i1} - Y_{i0} = a + \beta D_i + e_i \]  \hspace{1cm} (Equation 4.4)

In practice, a DiD regression model with panel data above would normally include other explanatory variables to control for differences in outcomes between the control and treatment groups that can be observed. Such a regression model with panel data is a type of ‘fixed effects’ regression model, as discussed above.

DiD estimation can also be applied to repeated cross-sectional data, i.e., where pre- and post-intervention observations are not made on the same set of units. The intuition here is that, if units are randomly drawn from the same population, units from the pre-intervention cross-section can be used as ‘surrogates’ for units in the treated and control groups in the post-intervention cross-section (Stock and Watson 2003).

The DiD regression equation for repeated cross-sectional data is defined below, with the coefficient on the interaction between the treatment dummy and time period dummy (\( \beta_I \)) representing the treatment effect.

\[ Y_i = \beta_0 + \beta_D D_i + \beta_t t_i + \beta_I D_i t_i + e_i \]  \hspace{1cm} (Equation 4.5)

Where:

- \( Y_i \) is the outcome for unit \( i \)
- \( \beta_0 \) is a constant
- \( \beta_D \) is the estimated coefficient of being in the treatment group
- \( \beta_t \) is the estimated coefficient of being in the second time period
- \( \beta_I \) is estimated coefficient of being treated and in the post intervention group
- \( D_i \) is a treatment dummy (1 if unit \( i \) is in treatment group; 0 otherwise)
- \( t_i \) is a time dummy (1 if unit \( i \) is a post-intervention unit; 0 otherwise)
- \( e_i \) is a random error term.

The interaction between the treatment and time dummies (\( D_i t_i \)) is the product of each unit’s treatment indicator and time dummy covariates, coded 1 for treatment units belonging to the second time period and 0 otherwise. This enables the corresponding regression coefficient \( \beta_I \) to be interpreted as the average treatment effect on the treated, holding constant (i.e., controlling for) the effect of time (\( \beta_t \)) and the effect being in the treatment group (\( \beta_D \)).

In either case, it is possible to use a Poisson or negative binomial distribution assumption for estimating treatment effects in a DiD framework. This involves modifying the estimation of the regression models as described in section 4.1.3 above.

It is possible to combine DiD with PSM if sufficient data is available. This involves first estimating propensity scores and using a matching algorithm to obtain a matched set of treatment and control groups. The matched data is then used to perform DiD analysis as described above. This is a powerful technique that can improve the balance between treatment and control groups (via PSM) and can control for the effects of certain types of unobserved factors which may affect treatment and outcomes (via DiD). We give examples of the application of the PSM-DiD technique in chapters 6 and 9.

Fixed-effects regression models can also be applied to some of the same types of data that can be used for DiD analysis. Fixed-effects models were discussed in section 4.1.2 above. In the case where pre- and post-intervention data is available for the same sets of treated and control units, fixed-effects and DiD analysis will give the same result.

4.3.2 Assumptions

The key assumptions of DiD are that any potential source of bias due to unobserved confounding covariates is additive and constant over time (Khandker et al. 2010). First, the effects of unobservable
variables on outcomes must be additive for DiD to ‘difference out’ their influence. Second, the effects on outcomes of unobserved variables causing bias must be constant over time, or have the same effect on outcomes for all units at any point in time, for the parallel paths assumption to hold (Mora and Reggio 2012). Generally, the latter point is stressed more than the former.

4.3.3 Examples

4.3.3.1 London congestion charge

Li et al (2012) use DiD estimation to examine the causal impact of the London congestion charge (LCC) on road traffic casualties. Because past studies have found fuel taxes reduce road casualties, the authors hypothesise that congestion charging, effectively another form of road taxation, may also affect road crashes. Crashes within the central London ward\(^\text{21}\) form the treatment group, while crash data from central city wards outside of London is used for the control group. This ensures control units are independent of the treatment, as areas bordering the congestion zone are unreliable control units given their susceptibility to flow on effects from the scheme.

Li et al (2012) analyse pre-intervention crash data to assess whether the control group mimics the trend of the treatment group (ie testing the parallel paths assumption). Figure 4.4 shows the time trend of car casualties in the treatment group (London) and potential control groups (Birmingham, Leeds and Manchester).

The car casualty trend of Leeds best reflects that of London prior to the introduction of the LCC in 2003. Accordingly, car fatalities and injuries from Leeds are used as the control units; however, Manchester and Birmingham are found to be the most suitable control groups for bicycle and motorcycle deaths.

Observed variables suspected to influence outcomes are controlled by including them as additional explanatory variables. Such factors include the length of each type of road, the number of junctions and roundabouts, and traffic exposure in each ward. A ward’s population and employee counts are used as proxies for traffic exposure as this data was not available.\(^\text{22}\)

Figure 4.4 Count of car fatalities and injuries by city central area

Source: Li et al (2012)

\(^\text{21}\) A ward is the primary unit of British administrative and electoral geography with an average area of 14km\(^2\)

\(^\text{22}\) The authors acknowledged that DID estimation used on repeated data sets may suffer from the problem of serial correlation which in turn leads to biased regression estimates. However, a Durbin-Watson test showed no serial correlation in the residuals of the data.
Results from the DiD model (where a negative binomial distribution for crashes is assumed) show congestion charges significantly reduced the total number of car crashes, but increased motorbike and bicycle crashes. This outcome is consistent with previous research that found the number of inbound two-wheeled vehicles and the average traffic speed within the area had increased by 15% and 31% respectively since the introduction of the LCC.

4.3.3.2 Ontario legal alcohol limit

Carpenter (2006) evaluated the lowering of the legal alcohol limit to zero for 16–17-year-old drivers in Ontario, Canada. Data from the 1983–2001 Ontario student drug use surveys is used to examine whether the zero-tolerance policy reduced self-reported drinking and drunk-driving. The treatment effect is measured using DiD estimation. Observable factors such as gender, region, school grade (age) and cigarette tax23 are controlled by the model. Data from the same survey pertaining to slightly younger (14–15 year olds) or slightly older (17–20 year olds) individuals is used to create separate control groups.

Although both control groups should provide informative counterfactuals for the alcohol-related behaviours of 16 and 17 year olds, neither are ideal. The author explains that ‘younger youths, for example, have much lower alcohol-involved driving rates (see figure 4.5) than the older youths. One might also be concerned about older students if [high school] dropout behaviour is related to the same personal characteristics that determine alcohol involved driving’ (Carpenter 2006, p187).

Figure 4.5 Fraction of individuals reporting drunk driving in the Ontario student drug use survey

![Figure 4.5](image)

Source: Carpenter (2006)

An estimation using the 14–15 year olds as the control group returns a statistically significant treatment effect. However, the pre-existing trend in this group is not equivalent to that of 16–17 year olds, violating the parallel trend assumption. The 19–20 year olds’ trend does meet this requirement and, when employed, shows very little support for a significant effect of Ontario’s zero tolerance policy. Thus, Carpenter concludes that ‘Ontario’s age-targeted drunk driving law... was not responsible for reductions in Canadian youth road fatalities over the past two decades’ (p187).

4.3.3.3 New York cell phone use ban

Sampaio (2010) analysed the causal effect on road safety outcomes of New York’s ban on cell phone use while driving. The primary motive of this research was to critique the work of Nikolaev et al (2010) who evaluated the same road safety policy. Although Nikolaev et al acknowledged that unobservable variables

---

23 Cigarette taxes are included to account for the possibility that alcohol and cigarettes are related in youth consumption, either as substitutes or complements.
such as ‘road construction, safety education, introduction of new automobile safety features, and/or changes in alcohol and illegal substance control policies’ might also contribute to the negative effect of the cell phone ban, they failed to control for such factors. Hence Sampaio repeats the study but employs DiD in the hope of obtaining a more reliable estimation of policy effect.

Crash data from Pennsylvania, a state that did not implement the ban, is used as the control group. Treated and control data include two potential dependent variables: the number of fatal/injury car crashes per 100,000 licensed drivers each year. Independent variables include a time trend over the years 1997–2007, an intervention dummy coded 0 for years before the introduction of the ban (1997–2001) and 1 otherwise, and a dummy variable representing the fixed effect of a county (of which there are around 60 within each of the wider states). Sampaio’s data seems to uphold the parallel paths assumption, as seen by the pre-intervention trends of figure 4.6.

Figure 4.6  Fatal crashes in New York and Pennsylvania

![Figure 4.6](source: Sampaio (2010))

Sampaio (2010) estimates the ban to contribute to 1.9 fewer annual fatal crashes per 100,000 licensed drivers (with a strong level of statistical significance). Although this result is similar to that of the previous study, 1.4, Sampaio insists ‘their results being similar does not imply they [Nikolaev et al] have a good identification strategy, since [they did not account for the fact that] fatal accidents are naturally decreasing and they do not have a control group that allow good comparisons between treated and non-treated states’ (p769). Although this estimate improves on the previous analysis, Sampaio does not claim to have the final answer to the causal effect, as his analysis is restricted by the questionable assumption that other shocks influence both states in the same way (i.e. the common trends assumption).

4.3.4 Limitations

The primary limitation of DiD analysis is that any potential unobserved confounding covariates must be constant over time, or impact all units equally at any given time. We can have more confidence that this assumption is satisfied when past outcomes of the treated and control follow very similar trends (i.e. tend to move in parallel over time). If this assumption does not hold, the DiD estimate will not be able to provide an unbiased estimate of the casual impact of an intervention. However, this limitation can be reduced if variables influencing the outcome are observed, allowing differences in outcomes to be controlled in the regression model. Additionally, evaluators could first filter out poor control units via PSM, which in turn may improve the equivalence of treated and control pre-intervention trends.

Another limitation is that DiD estimates can be adversely affected by correlation in the errors of the fitted model, either within groups or across time (Bertrand et al 2004). Such correlation can lead to substantial
under-estimation of the variance of estimated causal effects, which increases the chance that a significant causal effect will be found when no such effect exists.

4.4 Instrumental variables

4.4.1 Overview

Instrumental variable (IV) methods are a general technique that offer a solution to what is known as the ‘endogeneity’ problem in regression models. Endogeneity occurs when one or more explanatory variables are correlated with the error term in a regression model, i.e., when there are factors that are not included in the regression model that affect outcomes and also affect the explanatory variables that are included. This can be caused by various issues, including the presence of unobserved confounding covariates when estimating treatment effects, as discussed above.

More generally, endogeneity problems can arise when an explanatory variable in a regression model is not truly independent, i.e., there is a two-way relationship between it and the dependent variable. It is easy to see this could be an issue in transport applications. For example, suppose we are interested in the effect of transport infrastructure on productivity. Investments in transport infrastructure may increase productivity, but it is also possible that such investments are targeted at areas where productivity gains (or potential gains) are highest. Thus, there is possibly a two-way relationship between productivity and investment in transport infrastructure in a given area. To estimate the true effect of investment on productivity, we must attempt to disentangle these effects.

If we are trying to estimate the effects caused by a discrete intervention or treatment, recall that bias in the estimation of treatment effects occurs when one or more factors not included in a regression model affect both the probability of treatment and the outcome. The effects of such factors show up in the error term of a regression model, and so the error term will be correlated with the treatment variable. This violates one of the key assumptions of OLS regression, and the estimated coefficient on the treatment variable will be biased. Essentially, in such a case, OLS regression cannot separate the effects that are caused by treatment alone, versus those caused by the omitted factors that affect both treatment and outcomes.

The previous section explained how DiD estimation can help overcome problems due to unobserved factors that have the same effect on all observational units at any point in time. IV methods offer a solution to problems due to unobserved factors that may have different effects on observational units at any point in time. This involves finding one or more variables (called ‘instruments’) that are strongly associated with the explanatory variable(s) of interest but do not directly affect outcomes (Khandker 2010). In other words, the only effect of the instrument on outcomes is via the explanatory variable(s) of interest, and there is no independent or direct effect of the instrument on outcomes.

To understand how IV regression works, imagine that the variation in an explanatory variable $X$ has two parts: one part is correlated with the error term, $e$, while the other part is not (Stock and Watson 2010). If the correlation between $X$ and the error term can be removed, the regression assumptions are maintained, and the estimated coefficient on $X$ is free of bias. IV regression does this through the help of one or more instruments that are used instead of using $X$ directly.

---

24 This is also called a ‘simultaneity’ problem.

25 See appendix A for a review of OLS regression assumptions.
A valid instrument is a variable that is both strongly correlated with the explanatory variable(s) of interest but not correlated with the error term. If a valid instrument can be found, a technique known as ‘two-stage least squares’ can be used to obtain an unbiased estimate of the effect of the explanatory variable(s) of interest on outcomes. This involves first regressing the endogenous explanatory variable(s), $X_e$, on the instrumental variable(s) and the other control variables of the initial model, to isolate the variation in $X_e$ that is uncorrelated with the error term. The corresponding first stage regression equation is as follows:

\[
X_{ei} = \beta_0 + \beta_Z Z_i + \beta_1 X_i + e_i
\]  

(Equation 4.6)

Where:
- $X_{ei}$ is the value of the endogenous regressor for unit $i$
- $\beta_0 + \beta_Z Z_i$ is the exogenous component of $X_{ei}$
- $Z_i$ is the value of the instrumental variable(s) for unit $i$
- $X_i$ are other control variables from the original model
- $e_i$ is the problematic (endogenous) component of $X_{ei}$ for unit $i$

The second stage involves replacing original values of the endogenous variable(s) with their corresponding predicted (fitted) values from the first stage regression, and then re-estimating the original model (Van de Walle 2009). Because the problematic error component, ($e_i$ in equation 4.6), is not captured by the variation in the fitted values for $X_{ei}$, estimated coefficients from the second stage regression should now be unbiased.

Application of IV is best done in statistical packages such as R and Stata, which have built-in methods for IV regressions such as two-stage least squares. The XLStat add-on for Excel also provides two-stage least squares estimation. It is important to note the standard errors of the coefficients estimated in the second stage regression need to be adjusted to account for the use of predicted values from the first stage in place of the endogenous variable(s). This means two-stage least squares is not quite as simple as estimating two regression models, and cannot be properly applied using the built-in regression functions available in Excel.

For the same reasons as discussed in section 4.1.3 above, a Poisson or negative binomial distribution for outcomes may be preferred. This is possible in an IV framework, and some statistical software includes methods for estimating IV models under alternative distributional assumptions. While the estimation technique differs, the conceptual issues are the same as for the linear regression models discussed above.

### 4.4.2 Assumptions

In practice, evaluators must choose instruments carefully. A valid instrument must be:

- relevant – sufficiently correlated with the explanatory variable of interest (eg the treatment indicator variable)
- exogenous – have no correlation with the error term

In practice, finding variables that meet these two requirements can be difficult.

Generally speaking, an instrument can be deemed relevant if the first stage F-statistic is greater than 10.$^{26}$ This rule of thumb does not apply when there is more than one endogenous variable (see Angrist and Pischke 2009 testing guidelines in such circumstances for a modification of this rule). Essentially, the

$^{26}$ An F-test of overall significance tests the hypothesis that the fit of the specified model is statistically different from (or better than) a model with no predictors, ie an intercept-only model.
greater an instrument’s relevance, the more variation of $X_e$ it can explain, and the more information is available for IV regression (Stock and Watson 2003). Weak instruments do not allow for reliable IV regression estimates.

It is impossible to test whether an instrument is exogenous or not when the endogenous variables are exactly identified, i.e., there is exactly the same number of instruments as endogenous variables. In such a case, evaluators must use expert judgement to discern whether the instrument is exogenous or not. If the endogenous variables are over-identified, i.e., the number of instruments exceeds the number of endogenous variables, several methods can be used to test the exogeneity assumption. A popular method is the Hausman test for exogeneity, whereby the error component from the first stage regression is included as an additional explanatory variable in the original model. If the coefficient corresponding on the first stage error term is statistically significant in this model, we can conclude the instrument is in fact endogenous and thus not appropriate for IV regression.

4.4.3 Examples

4.4.3.1 Spain: productivity impact of road transport infrastructure

Matas et al. (2015) examined the effect of road transport infrastructure on firm productivity in Spain. Data from the Spanish Structure of Earnings Survey over three different years (i.e., repeated cross-section samples) were used for the analysis. Wage data by geographical area unit was used as the dependent variable, as a proxy for productivity by area unit. The effect of transport infrastructure was included through an ‘effective density’ variable, which is essentially a measure of accessibility, as defined in Graham (2007). A number of firm and individual characteristics, such as a firm’s sector and an employee’s education and occupation, were controlled for given their expected influences on wages.

The potential two-way relationship between productivity (the dependent variable) and accessibility (the treatment) is a problem to overcome, as discussed above. As the productivity of firms in an area is likely to determine accessibility, the regression model cannot isolate the reverse effect (and the relationship of interest), i.e., the impact of accessibility on productivity. This endogeneity problem is dealt with through IV regression, where pre-improvement population (by area unit in 1900) forms the instrument. This variable is considered to be a valid instrument given its correlation with current accessibility yet independence from the random error term due to the long time lag.

A Hausman exogeneity test shows that IV estimates of the accessibility effect on wages is statistically different from that predicted in a simple regression model. Although this test gives evidence of the existence of endogeneity, the IV and standard regression estimates of the effect of accessibility on wages are very similar, and thus the endogeneity problem is found to be small. It is concluded that every 1% increase in accessibility tends to increase productivity by 0.06%.

4.4.3.2 Great Britain: impacts of road infrastructure on company performance

Gibbons et al. (2012) also estimated the causal impact of increased accessibility from improved transport infrastructure on firms’ performance. Great Britain data from 1998 to 2008 was used for the analysis. A combination of road network data and road improvement data was used to create a measure of accessibility to employment. Outcomes of interest included firm and area unit level employment and

---

27 Effective density is a measure of accessibility (distance) to employment by ward weighted by the inverse of the distance to all areas including the firm’s own ward.

28 Combining unit record data for individuals with higher-level ‘group’ data can cause bias as the group-level data may omit factors that determine individual outcomes. However, in practice this problem may be impossible to resolve as individual-level data is often not available for all relevant characteristics.
productivity. Electoral wards were used as geographical units, of which there are over 10,500 in Great Britain.

Three key sources of potential endogeneity were identified as:

- unobserved heterogeneity, e.g., the influence of area-specific yet unmeasurable productive advantages, which affect both employment accessibility and economic outcomes
- non-random assignment, e.g., transport improvements are intentionally placed in wards or close to firms with specific characteristics
- reverse causation, e.g., the potential for the outcome variables to drive changes to employment accessibility.

The problem of unobserved heterogeneity was addressed through the use of a fixed-effects model, whereby dummy variables for each ward capture underlying factors, e.g., productive advantages, specific to each ward. Fixed-effects models assume these ward-specific factors are constant over time and can be subtracted away from the causal effect by exploiting panel data. Non-random assignment is dealt with by restricting the analysis to firms that are close to improvements (within a 20km radius). This is possible given, within this distance band, the observed impact of improvements varies considerably, and improvements are unlikely to target specific firms or wards within such a small radius.

Simultaneity or reverse causation was overcome through IV regression. The authors believe that ‘changes in accessibility due to the relocation of employment across space may be directly affected by the outcome variable or correlated with unobserved shocks in the error term… which may lead to bias in the estimation’. To address this specific issue, their accessibility variable was predicted using the ‘pre-improvement [1997] spatial distribution of employment’. The predicted accessibility variable using this instrument should only vary with changes to the transport network, and not current levels of employment.

A second regression was estimated using these predicted values in place of actual accessibility. Like Matas et al (2015), the IV results were similar to that of the original regression, meaning the endogeneity due to reverse causation was minimal. Gibbons et al (2012) concluded that, in Great Britain, ‘a 10% improvement in transport-induced accessibility leads to about a 3% increase in the number of businesses and employment, up to 30km from the site of the improvement’.

4.4.3.3 US: effect of speed on fatalities and travel time

Ashenfelter and Greenstone (2002) analysed the effect of vehicle speed on road fatalities and travel time savings. To investigate these relationships, their study exploited data observed before and after a 1987 US road safety policy change, where the rural interstate speed limit increased from 55mph to 65mph. Although this policy was authorised by the federal government, only 21 states adopted the change at that time. By valuing time saved at the average hourly wage, the authors could calculate a monetary sum that an ‘adopting’ state is willing to accept in return for the increased risk of road deaths.

Annual data specific to both adopting (treated) and non-adopting (control) states was used for the analysis, including: vehicle miles travelled, fatal crashes and average vehicle speeds. First, a basic DiD model was used to estimate the average impact of the 65mph rule on road fatalities and average speed. DiD estimates showed a 31% increase in fatalities and a 4.5% increase in average speed for the ‘adopting’ states.

For exploratory purposes, the researchers regressed fatalities on speed to find no correlation between speed and road deaths. However, this is unsurprising if people choose to travel more slowly to reduce risk when driving on a more dangerous road, even if the speed limit is higher. Such compensatory driving behaviour would imply the speed variable was endogenous, causing the estimated effect of speed on fatalities to be biased. IV regression was employed to solve this problem, with the desired instrument being a variable that
was correlated with speed but did not affect fatalities, except through speed. Thus, a plausible instrument was whether the 65mph speed limit was introduced or not. Once speed was instrumented by the policy indicator variable, the second stage regression showed, on average, the adoption of the 65mph speed limit increased fatality rates by 35% and speeds by 3.5% (2mph) on rural interstate roads. Ashenfelter and Greenstone (2002, pii) concluded if time saved was valued at the average hourly wage, ‘adopting states were willing to accept risks that resulted in a savings of $1.54 million ($1997) per fatality’.

4.4.4 Limitations

The key limitation of IV regression is the difficulty in finding suitable instruments in practice. As seen in the examples above, a lagged treatment variable (or a time lag of a variable closely associated with the treatment variable) is a popular instrument of choice. This is because it is implausible that current outcomes have influenced past values of a treatment variable. Again, the applicability of such an instrument depends on data availability.

It is also important to note that using invalid estimates can produce worse results than using no instruments at all. In particular, when instruments are only weakly correlated with the endogenous variables, or if instruments are not truly exogenous (ie are correlated with the error term), IV estimation can produce biased and inconsistent estimates of causal effects. This is compounded by the fact that diagnostic testing cannot completely determine whether a given instrument is valid. The indirect method of estimation used in IV also means IV estimates tend to be less efficient (ie have higher variance) than non-IV methods.

4.5 Regression discontinuity

4.5.1 Overview

Regression discontinuity (RD) analysis can be used for causal inference when treatment is not randomly assigned but is clearly assigned according to a characteristic that can be observed and modelled (Angrist and Pischke 2009). Consequently, there is no overlap (or a discontinuity) between treated and control units with respect to the variable(s) that determine whether a unit receives the treatment.

For example, treatment could be assigned based on an observed unit characteristic, such as a specified age limit, where all individuals under this threshold are treated. In such a situation, there are no ages of individuals at which we observe both treatment and control observations,29 thus evaluators must ‘make strong modelling assumptions because we will be forced to extrapolate our model out of the range of our data’ (Gelman and Hill 2007, p213). It becomes increasingly difficult to be confident these assumptions are satisfied as we look further away from the cut-off point between the treatment and control groups. Hence it is common to limit RD analysis to values of characteristics around the assignment threshold, and when interpreting the results we should be careful not to apply these to units with characteristics that are very different from units around the cut-off point.

To illustrate how RD works, figure 4.7 shows a linear discontinuity, a non-linear discontinuity, and nonlinear relationship mistaken for a discontinuity. Essentially, RD analysis extrapolates a relationship from data on both sides of the cut-off (ie for both treated and controls), and estimates whether the difference in outcomes is attributable to the intervention or not, controlling for the other explanatory variables included in the model. Panel C of figure 4.7 illustrates how non-linearity can be mistaken for discontinuity if a model is not specified properly.

---

29 Sometimes discontinuity is not so distinct, this is called a ‘fuzzy’ overlap rather than a ‘sharp’ overlap.
RD can be implemented by estimating a regression model including variables measuring the treatment assignment and other variables expected to influence outcomes. The estimated coefficient on the variable used to represent treatment assignment is the estimated treatment effect, assuming there are no confounding covariates not included in the model. As discussed above, care should be taken to only interpret treatment effects for units having characteristics in the vicinity of the discontinuity. It is therefore not possible to make robust general statements about treatment effects estimated from RD analysis.

4.5.2 Assumptions

RD first requires clear knowledge of the mechanism of treatment assignment, eg value of the cut-off point. Second, the variable(s) capturing assignment to treatment must be measured prior to the treatment. Lastly, for both the treatment and control groups, the estimated relationship between the outcome and the explanatory variables in the model must hold outside the ranges of the explanatory variables that are observed for each group. In other words, average outcomes just below the cut-off are a valid counterfactual for units just above the cut-off, and vice versa. Lee (2008) notes these assumptions can be remarkably easy to satisfy compared with other causal inference techniques such as DiD or IV regression.

4.5.3 Examples

Very few transport intervention evaluations employ a true RD design. Although a number of transport papers claim to have applied RD to the data, more often than not, their method is in fact an interrupted time series (ITS) design (see section 4.6). Essentially, these articles have analysed a discontinuity with respect to a point in time (ie a break in a time series) rather than at a discontinuity with respect to a characteristic level (ie a break in a cross-section). Therefore, such articles are not summarised here but in the section on interrupted time series analysis (ITSA) below.

4.5.3.1 India: impacts of roads on school enrolment

Mukherjee (2012) is one of the few transport research papers that employs a true RD analysis. This study examined the impact of an Indian rural road scheme, a programme that constructed new ‘all-weather’ roads in villages that previously had limited road access, on the change in school enrolment in beneficiary villages. Because only those villages with populations of 500 and above were eligible for this intervention, Mukherjee used an RD design to compare student enrolments between villages close to this population cut-off. Unobserved factors correlated with both population size and school enrolment would cause Mukherjee’s estimates to suffer from omitted variable bias. Therefore, after carrying out the RD on the rural road scheme, the author followed recommendations from Lee and Lemieux (2009) and conducted (or simulated) RD analyses at the left and right-hand side medians of the horizontal axis (at village populations of 422 and 573). Neither of these simulated RDs show a statistically significant discontinuity, and in turn provide a robustness check of the treatment effect estimated in the original RD model. Final results show improved road access increases school enrolment by around 22%. The size of this effect is dependent on the general age cohort and the social background of the students.
4.5.3.2 Sierra Leone: road projects and rural market prices

Casaburi et al (2013) examined the impact of EU funded rural road projects in Sierra Leone to transport costs and rural market prices. To facilitate the allocation of funds, the EU prioritised roads according to an index variables relating to local economic data. The highest-ranked projects were chosen until the cumulative distance of these roads totalled 150km in each district. Because treatment was assigned according to a single variable, the priority index variable, a RD design was employed to analyse the causal effect of this scheme.

A wide range of data was collected for the RD analysis. A total of 47 roads were included in the study; however, the RD was applied to the closest 31 observations to the road rehabilitation cut-off. The researchers modelled a number of RDs for the different dependent variables (outcomes) including: transport costs, travel speed and local market prices for different types of crops. In general, these models included dummy variables to control for road characteristics and district-specific effects, and a third order polynomial to approximate the trend of the economic index variable.

They found improvement in road quality reduced transport costs by a statistically significant 59% RD results showed the effect of road improvements on crop price was heterogeneous and largely determined by local market characteristics such as the surrounding area’s level of productivity.

4.5.4 Limitations

Causal inference with RD is only feasible when treatment is assigned via a single observed characteristic. In general, this type of intervention allocation is rare; however, the common practice of prioritising road projects by their net present value (ie CBA) makes RD designs particularly relevant to road transport ex-post evaluations. This is because projects that are comparable with respect to their intervention type and net benefit forecast are likely to have similar characteristics and thus similar counterfactuals.

4.6 Interrupted time series analysis

4.6.1 Overview

An ITSA is essentially an RD where the ‘break’ occurs at a point in time. ITSA is more commonly used in impact analysis given its applicability to almost any type of intervention and the prevalence of time-series data.

In its simplest form, ITSA only requires the date of implementation and a time series of outcomes before and after an intervention. Although ITSA may be simpler than RD from an implementation perspective, Reis and Judd (2000) explain the ‘time’ element introduces statistical complications that are absent from RD designs, eg the need to control for long-term time trends, cyclical time trends and autocorrelation. If these issues are addressed appropriately, ITSA can provide a valuable impact evaluation study design, particularly for population-wide interventions (Bernal et al 2016). However it should be noted ITSA requires extrapolating trends beyond the time period observed in the data (ie a type of forecasting), which requires strong assumptions and is probably less robust than other causal inference methods.

A simple ITS regression equation can be defined as:

\[ Y_t = \beta_0 + \beta_1 X_t + \beta_2 T_t + \beta_3 X_t T_t + e_t \]  
(Equation 4.7)

Where:
- \( Y_t \) is the outcome at time \( t \) (dependent variable)
- \( X_t \) is time elapsed since the start of the intervention (same frequency as the outcome variable)
- \( T_t \) is a dummy treatment variable (coded 0 and 1 for pre- and post- intervention periods)
Ex-post evaluation of transport interventions using causal inference methods

\[ \beta_1 \] is the estimated effect of time on the outcome variable (i.e., the time trend)

\[ \beta_2 \] is the estimated change in the level of the outcome variable after the treatment

\[ \beta_3 \] is the estimated change in the time trend after the treatment

\[ e_t \] is a random error term with zero mean and constant variance

Only three variables are required in a basic ITSA: outcomes, time and the intervention dummy. However, information about attributes of units (i.e., covariates) can be included as other explanatory variables and thus controlled by the model. Once again, it is possible to use a Poisson or negative binomial distribution assumption for \( Y_t \) if this is more appropriate, as described in section 4.1.3 above.

If sufficient data exists to do so, the validity of ITS causal estimates can be tested by replacing the outcome variable with a ‘non-equivalent independent variable’, i.e., outcomes from a control group. Shadish et al. (2002) explain that such a variable ‘is predicted not to change because of the treatment but is expected to respond to some or all of the contextually important internal validity threats [i.e., threats to causal inference] in the same way as the target outcome’ (p. 509). If the treatment is found to have a statistically significant impact on the outcomes of the control group, then we can infer the response measured within the treated group was not caused by the intervention.

4.6.2 Assumptions

Similar to the key RD assumption, evaluators applying ITSA need to assume that effects of other explanatory variables on outcomes are similar before and after the treatment. In other words, data prior to the cut-off point in time provides an adequate counterfactual for outcomes after that point. This is a strong assumption if the difference in time is large.

4.6.3 Examples

4.6.3.1 London congestion charge

Percoco (2014) evaluated the causal effect of the London congestion charges (LCC) on pollution levels. From a theoretical perspective, the environmental implications of the LCC are ambiguous. For example, a reduced number of vehicles in the central area should lead to lower pollution levels; however, the authors also hypothesised that increased speeds due to less congestion might lead to more fuel consumption and vehicle emissions. An ITSA approach was adopted (although labelled an RD method) to analyse the break in the trend of pollutant concentration around the implementation of the scheme. Outcomes of 132 monitoring stations, located both inside and outside the LCC, were studied for their measurement of various pollutants from 2000 to 2013.

A spatially aggregated model was first estimated. In other words, the dependent variable was the sum of pollution from all sites and the treatment variable was denoted by years before and after the LCC’s introduction. A weather variable, a lagged dependent variable (to account for temporal dependence in concentration), and a time trend were controlled by the model. Consideration was also given to another vehicle emissions countermeasure, the low emission zone (LEZ), which was introduced not long after the LCC. The confounding effect of the LEZ was mitigated by either restricting the analysis period to before its implementation or including it as an additional explanatory (dummy) variable. A statistically significant reduction in the city’s pollution levels could not be attributed to the LCC when all of these factors were accounted for.

However, this aggregated model omitted a variable important to the outcome of interest: monitoring site distance from the charged area. To control for site heterogeneity, a spatial model was used whereby outcomes are specific to measuring sites (i.e., not aggregated) and an interaction variable between the treatment and monitoring station distance from the LCC boundary was introduced. This model showed a
significant decrease in the concentration of five pollutants in the treated area, and an increase in pollutant concentrations outside the charged area. Percoco’s finding is supported by evidence of traffic diversion from charged to uncharged routes as a consequence of the LCC, thus its relevance for city wide pollution is limited.

4.6.3.2 Italy: driver licence demerit points

De Paola et al (2010) estimated the causal effect of a demerit points system applied to driving licences, implemented in Italy, 2003. Under a demerit points system, certain driving violations incur penalty points according to the severity of the offence. Accumulation of points above a set threshold leads to driver licence suspension. Such a policy should encourage drivers to behave more carefully, promote safety and ultimately reduce road crashes and fatalities.

To test this hypothesis, the researchers used an ITSA to predict road casualty and driver offences applying a treatment variable (time of observation with respect to the policy introduction) while controlling for ‘month of the year, day of the week, an indicator for holidays, the monthly number of police patrols and speed cameras, average daily precipitations, average monthly gasoline price, [and] a dummy for highways’ (De Paola et al 2010, p8). Traffic crashes might also be related to time for a variety of reasons, including: technological progress in car safety, road maintenance and investments, and increasing driver awareness of traffic risks. Improvements to any one of these factors over the analysis period are likely to result in an overestimation of the treatment effect. To overcome this issue, the evaluators restricted the analysis to outcomes just before and just after the policy implementation. Consequently, factors that potentially change over longer periods were assumed to be constant during the relatively short analysis window, and thus ‘any jump in the dependent variable in proximity of the cut-off point can be interpreted as evidence of a treatment effect’ (De Paola et al 2010, p8).

ITSA was carried out on separate outcome variables (crash, injury and fatalities) using OLS and Poisson estimators. Both estimators showed a significant policy effect: a 10%, 21% and 35% reduction in crash, injury and fatalities respectively. Preference was given to the Poisson estimator as the dependent variables took on non-negative discrete values (ie are count data). Figure 4.8 shows a clear jump in the relationship between the outcomes and the time variable in proximity to the policy implementation date. De Paola et al concluded that the policy had a significant causal effect on road safety in Italy.

30 Conventional OLS regression assumes that errors are normally distributed around their true value. However, count data, especially in small samples (n<30), follows a distribution that is closer to that of a Poisson model, where the distribution mean is equal to its variance. Thus regression models that assume a Poisson distribution are generally used when analysing count data.
4.6.3.3 UK: crash rate impacts of motorway extensions

Olsen et al (2016) examined the effect of motorway extensions on crash rates for surrounding local roads. In the UK, motorways account for 21% of all traffic yet only 5.4% of fatalities and 2.7% of casualties (Department for Transport 2014). These statistics suggest motorways provide safer infrastructure for traffic, and lead to Olsen et al’s hypothesis that motorway extensions reduce crashes on nearby local roads.

Their analysis focused on the construction of several kilometres of new motorway running through Glasgow, Scotland. This road was built above existing roads (so did not replace any existing arterial routes) and opened in 2011. Local roads directly surrounding the motorway extension defined the treated area. Two separate control areas located within Glasgow were used: an area surrounding an existing motorway, and an area without a motorway. Study area boundaries were set so their respective social, economic and demographic characteristics were comparable.

ITS regression models using autoregressive integrated moving average (ARIMA) errors were fitted to monthly casualty (count) data.31 Time series of outcomes in each area show a falling trend in road crashes from 1997 to 2014. No significant difference in level or trend was found in the treated area since the motorway extension. Olsen et al’s results suggest policymakers cannot necessarily justify the construction of new urban motorway infrastructure under expectations it will reduce road crashes or casualties.

4.6.3.4 Sweden: bicycle helmet law and child head injuries

Bonander et al (2014) studied the effectiveness of a 2005 bicycle helmet law on head injury rates for children. Their analysis used Swedish hospital data pertaining to injured cyclists from 1988 to 2012. This data also contained information about a patient’s age (0–14 or 15+ years old), sex and injury diagnosis. The monthly proportion of head injuries among cyclists admitted to hospitals was used as the dependent variable. Outcomes were distinguished between 0–14 year olds and 15 years and older. A time trend and seasonal factors were also included. Differences in the intervention effect between children (treated units) and adults (a non-equivalent dependent variable) were tested using interaction terms to identify the ATET.

31 It is common for the errors (residuals) of models estimated using time series data to be correlated over time. This violates the regression assumption of errors being independently distributed. As a consequence, standard errors of the estimated coefficients will be wrong. Essentially, ARIMA models modify the error component of a time series model to correct for this problem.
A negative binomial distribution was used for the estimation. The bicycle helmet law was found to have a statistically significant effect (-7.8%) on head injuries for male children, yet no effect on female children.

4.6.4 Limitations

Reis and Judd (2000) stipulate three statistical problems of ITSA that are absent from other RD analysis using cross-sectional data:

- the presence of long-term time trends
- the cyclical nature of time series data
- the tendency of autocorrelation within time series data, i.e., adjacent observations are often more similar than observations further removed in time, which if present leads to a violation of a key regression assumption.

Incorrect specification of long-term trends and cycles can lead to biased estimates of the treatment effect. Failure to account for autocorrelation can lead to incorrect standard errors and significance tests.

4.7 Bayesian methods

4.7.1 Overview

Bayesian statistical methods are an alternative to the econometric methods discussed above and have been widely used in before-and-after road safety studies (Li 2013). The popularity of Bayesian methods in before-and-after road safety studies arises from the fact they can be implemented where no data for a control group is available, and they address two common problems in before-and-after studies:

1. **Imprecision** – for example, we could observe 30, 34, 27 and 36 crashes per annum over four consecutive years at a given intersection. Although the intersection’s fundamental level of safety might not have changed, the crash count varies from year to year due to unknown factors. Therefore, using a single year’s crash count, or even that averaged over a two to three year period, will give an imprecise value of the intersection’s actual level of safety.

2. **Regression to the mean (RTM) bias** – this is a type of selection bias that occurs when safety interventions are assigned to units with abnormally high crash rates (Datta et al 2003). Following the above example, if the same intersection is chosen for a road safety improvement after the year of 36 crashes, even without intervention, we would expect the annual crash rate to reduce and converge to its long run average value. Attributing the subsequent reductions in crash frequency to the intervention alone would lead to an overestimation of the treatment effect.

Both issues are attributable to the random nature of motor vehicle crashes at any intersection in any given year. To overcome these issues, Bayesian methods assume a location’s historical crash data is not the only source of information about its actual level of safety; more information can be found in the safety outcomes of similar locations (Hauer et al 2002). Bayesian statistics provide methods for combining these two sources of information to give a more precise estimate of a location’s level of safety than can be inferred from either source of information alone.

Therefore, the goal of Bayesian analysis of road safety interventions is to:

- Use all available information to more precisely estimate the starting (pre-intervention) level of safety of the treated units, defined as their (pre-intervention) expected or long run average crash rates.
Estimate the (post-intervention) counterfactual crash rate. This is done by adjusting the pre-intervention expected crash rate to account for the change in time and traffic volume (Amundsen and Elvik 2004). The pre-intervention expected crash rate is used instead of the actual pre-intervention crash rate to alleviate regression to the mean bias (Amundsen and Elvik 2004).

It should be noted that, while these methods can adjust the counterfactual crash count to accommodate the types of problems described above, the post-intervention crash count will still be volatile and this may make it difficult to estimate reliable causal effects.

Two general types of Bayesian analysis are possible: ‘empirical’ Bayes (EB) and ‘full’ Bayes (FB). Both methods are conceptually similar, and the main difference between these approaches is that EB methods work with point estimates whereas FB methods work with fully specified probability distributions. FB methods have some advantages of greater flexibility and potentially greater accuracy than EB methods, but are more complex. Accordingly, EB methods are more common in the literature and our discussion below covers these methods only.

The underlying motive of the EB method is that a unit’s observed crash record alone is not sufficient to understand its pre-intervention level of safety. Outcomes from data relating to similar units can be used through the safety performance function (SPF) to define the pre-treatment level of safety more precisely. The EB method also assumes this estimated pre-treatment level of safety can then be adjusted (to account for changes in traffic volume and time) to identify the counterfactual post-intervention level of safety.

### Implementation of EB methods for road safety evaluations

The following summary of the EB procedure is based on Hauer et al (2002), Persaud et al (2010) and Li (2013). These studies first define the treatment effect as the change in safety for a given location (unit):

$$\lambda - \pi$$  \hspace{1cm} (Equation 4.8)

where $\lambda$ is the counterfactual number of crashes for the treated locations and $\pi$ is the observed number of post-intervention crashes for the treated locations.

As mentioned above, the EB method does not use past crash count data ($y_0$) directly to arrive at the counterfactual estimate, $\lambda$, for any given location. First, the expected value of $y_0$, $E(y_0)$, is estimated. Typically, this is done with the aid of two sources of information (Hauer et al 2002):

1. Pre-intervention crash frequency of the treated locations ($y_0$)
2. The crash frequency expected of similar locations ($\mu$). In the language of Bayesian statistics this is known as a ‘prior’, ie it represents our best estimate of the crash frequency of a treated location prior to knowing its actual crash frequency.

Box 4.1 gives an example of a ‘prior’ commonly used in road safety studies.

Once we have calculated $\mu$ through the SPF, we use $\mu$ and $y_0$ to estimate the long run average or expectation of $y_0$: $E(y_0)$. These data points are typically combined by weighting each value and summing the products (Amundsen and Elvik 2004):

$$E(y_0) = w\mu + (1-w)y_0$$  \hspace{1cm} (Equation 4.9)
Box 4.1 The ‘prior’ and the safety performance function

In EB methods applied to road safety evaluations, the prior for any given location is derived from a structural equation that explains outcomes (e.g., annual crash count) as a function of site characteristics (e.g., lane width, traffic volume, and speed limit). In the literature, this equation is called the SPF and can be estimated via a regression on historical data relating to the treated units and other similar units (Hauer et al. 2002). A simple example of a SPF is:

$$\mu = 0.035 \times \text{AADT}^{0.6}$$

where $\mu$ is the expected annual crash count of a location, and AADT is the annual average daily traffic count. For example, if we are evaluating the safety effect of an intervention at a single intersection that has an AADT of 4000, absent any other information, we would expect the crash count to be 50.73 (0.35 x 4000$^{0.6}$).

It should be noted the SPF is itself a causal relationship, i.e., it attempts to predict the expected crash count at a location as a function of its characteristics. For all the same reasons as discussed in previous sections of this report, care must be taken when estimating such a function to ensure the coefficients are free from bias and as accurate as possible.

The weight, $w$, is calculated as a function of the mean and variance of SPF regression model (Datta et al. 2003). Essentially, the weight given to $\mu$ depends on the reliability of the SPF, i.e., the difference in safety of a specific site from the other similar sites (Hauer et al. 2002).

To estimate $\lambda$, a factor is applied to $E(y_0)$ to adjust for the pre- and post-intervention differences in time and traffic volumes (Persaud et al. 2010). The actual post-intervention crash count is subtracted from the counterfactual to estimate the change in safety to the treated, i.e., the treatment effect ($\lambda - \pi$).

4.7.3 Examples

4.7.3.1 Norway: safety effects of section control

Høye (2015) conducted a before and after study on the safety effects of section control at 14 sites in Norway. Section control is defined as a controlled stretch of road with speed cameras placed at either end. Outcomes analysed include the number of injury crashes and the number of people killed or seriously injured (KSI). The ‘before’ period was 36 months for all treated units, the ‘after’ period ranged between 13 and 36 months. Where possible, spill-over or crash migration impacts of speed cameras were also examined by analysing outcomes within 3km of the treated section.

As no comparison group was available, an SPF was estimated using historical Norwegian crash data and used to estimate crash rates of roads with similar characteristics. Predictor variables used in the SPF included traffic volume (AADT), speed limit, road category, and the number of lanes, junctions and bends in the section of road. Crash rates were assumed to follow a negative binomial distribution with a variable over-dispersion parameter. Once the SPF was estimated, the EB method was used to integrate the predictions of this equation with historical site-specific data to control for RTM bias. The relative

---

32 Over-dispersion does not exist in simple linear regression models, as the variance of the dependent variable is assumed to be independent of the mean. However, the distributions used to estimate count data (such as the Poisson or negative binomial) often specify a relationship between the variance and the mean (e.g., in the Poisson model, variance is equal to the mean). Over-dispersion is where the observed data variance is greater than the variance defined by the assumed distribution.
importance of the SPF and the site-specific data in the EB method was determined by the accuracy of the SPF model.

Høye (2015) found a 12% and 49% reduction in injury crashes and KSI respectively at treated sites. However, the decrease was only statistically significant for KSI outcomes. Near the treated sites (up to 3km in each direction), injury crashes were found to be significantly reduced by 46%; however, the number of KSI downstream were too few to draw any conclusions. Høye concluded section control was effective in reducing crashes, especially serious crashes, and crash reductions at or near to the control sections were more likely to occur than crash migration.

4.7.3.2 Norway: safety impacts of new urban arterial roads

Amundsen and Elvik (2004) analysed the road safety impacts of four new urban arterial roads in Oslo, Norway. The study looked at the rate of injury crashes and KSI, four years before and after the intervention at each location. Amundsen and Elvik controlled for general trends in the number of crashes by using the total number of crashes in the city of Oslo as a comparison group.

Even though the researchers subtracted crashes recorded at the treated sites from those for Oslo as a whole to avoid double counting, this type of comparison group can be problematic when the intervention affects surrounding roads. While Amundsen and Elvik acknowledged adjacent non-treated sites were likely to be affected by new arterial roads, it is not clear such roads were excluded from the area-wide comparison group, as best practice recommends. Regression to the mean was controlled for using the EB method.

The treatment effect is defined in terms of the percentage change in the expected number of crashes, as denoted by equation 4.10:

\[
\text{Change in expected number of crashes} = \frac{(A/B)}{(C/D)}
\]

(Equation 4.10)

Where:

- \(A\) - post-intervention crash count for the treated
- \(B\) - pre-intervention expected crash count for the treated
- \(C\) - post-intervention crash count for the control
- \(D\) - pre-intervention crash count for the control

Note that \(A, C\) and \(D\) are recorded data while \(B\) is an estimate of the expected number of crashes in the before period. \(B\) is derived using the EB method (as described in section 4.7.2) to control for RTM bias. Calculation of point estimate (eg treatment effect) variance and confidence interval is complex and more explanation is provided in Amundsen and Elvik (2004).

After controlling for the area-wide trend in crashes and RTM bias, a 9% reduction in the number of injury crashes was found for all four projects combined; however, it is not statistically significant. Amundsen and Elvik (2004) concluded that the safety impacts of new urban arterial roads are often counteracted by a simultaneous increase in traffic volume.

4.7.4 Limitations

The EB method relies heavily on a correctly specified prior, eg calculated using the SPF. Although this dependency is lessened by weighting the SPF according to its model accuracy, given that the weight assigned to the other key component (historical crash data) is directly proportional to this weight, a poorly specified SPF will place undue importance on past crash data, and consequently fail to overcome RTM bias.
effectively. In addition, as noted in the box above, the SPF itself must be carefully estimated to ensure its coefficients are unbiased. Sceptics of the EB method also believe the increased complexity and data needs (eg large sample of reference groups) do not warrant the effort required, as alternative and less sophisticated methods can produce equally reliable results (Persaud and Lyon 2007; Li 2013).

4.8 Multi-valued and continuous treatment methods

4.8.1 Overview

The causal inference methods reviewed above apply to a binary treatment, ie a unit is either treated or not and the treatment is the same for all units that receive it. However, in some cases treatments may not be binary and may take multiple or continuous values of the ‘level’ or ‘dose’ of the treatment. In such cases, average treatment effects are a function of the dose, rather than a single number, and causal inference seeks to estimate that function.

Throughout this report, we focus on causal inference methods for binary treatments, as this setup is used in the majority of the causal inference literature published to date. A smaller body of literature analyses treatments that are multivalued or continuous. This approach may be useful for certain types of transport interventions: table 4.2 provides some examples.

Table 4.2 Examples of transport interventions

<table>
<thead>
<tr>
<th>Binary treatments</th>
<th>Multi-valued treatments</th>
<th>Continuous treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td>tolled/ toll-free route</td>
<td>frequency of service</td>
<td>network capacity</td>
</tr>
<tr>
<td>presence of speed camera</td>
<td>multiple speed limits</td>
<td>length of segregated route</td>
</tr>
<tr>
<td>40km/h zone</td>
<td>cars per train</td>
<td>density of intersections</td>
</tr>
<tr>
<td>peak/ off-peak times</td>
<td>number of O-D routes</td>
<td>accessibility</td>
</tr>
<tr>
<td>pedestrianised/ unpedestrianised</td>
<td>number of network nodes</td>
<td>tax/subsidy rates</td>
</tr>
</tbody>
</table>

Source: Table 1 in Graham (2014)


4.8.2 Assumptions

Causal inference methods for multi-valued and continuous treatments seek to estimate APOs for doses. Unlike the ATE or ATET, these are not single numbers but vary with the level or dose of the treatment. For multi-valued treatments, ie treatments with a small number of different possible doses, the aim is to estimate an average treatment effect for each different dose. For continuous treatments, the aim is to estimate a function relating the APO to the level of the treatment. In such cases, applying a binary treatment effect approach would obscure estimation of the true treatment effect because it does not allow for variation in the dose of the treatment received by different units (Linkow et al 2015).

By allowing the treatment variable to be multi-valued or continuous, we assume units respond differently to various levels of exposure. Additional assumptions, such as units being ‘as if’ randomly assigned given the observed characteristics (ie conditional independence), will depend on the model used, but are conceptually similar to the assumptions required to estimate causal effects of binary treatments.
4.8.3 Examples

4.8.3.1 Georgia: impact of highway improvements

Linkow et al (2015) evaluated the impacts of a US$200m improvement to a large section of highway in the Republic of Georgia. Completed in 2008–2010, the project aimed to provide significant connectivity benefits, linking Tbilisi (the capital) to the bordering countries of Turkey and Armenia, and passing through one of the poorest and historically least accessible regions of Georgia.

A wide range of outcomes was analysed by Linkow et al, including changes to business activity (the number of industrial facilities, land use, food prices) transportation (traffic volume and speed, frequency of public transport services) and households (income, consumption, asset ownership). Information from household surveys, community-level surveys, and national road network data from 2003–2012 was amalgamated for the analysis.

The authors used DiD models to evaluate the treatment from both binary and continuous treatment perspectives. For the binary approach, the treatment group was defined as settlements within 30 minutes travel time of one of the improved roads. For the continuous treatment (or ‘dose-response’) approach, treatment was defined as the extent to which the intervention reduced travel time between the settlements and various relevant destinations. This approach was used because the investment impacted on different settlements in different ways in terms of the improved connectivity that was available.

Apart from the household data (which sampled different units each year), the data used by Linkow et al was a balanced panel (ie the same units observed in each time period). Therefore, ‘fixed effects’ were specified for community, transport, and business units and ‘random effects’ for household units.

The general structural equation used to estimate continuous treatment effects is expressed in box 4.2 (note the actual models estimated by Linkow et al also include fixed or random effects and other explanatory variables in some cases). The results from these models were compared with results from binary treatment DiD models (such as those reviewed in section 4.3 above).

---

Box 4.2 DiD equation and interpretation of coefficients under a continuous treatment approach

\[ Y_i = \beta_0 + \beta_D D_i + \beta_t t_i + \beta_I D_i t_i + e_i \]

Where:

- \( Y_i \) is the outcome for unit \( i \)
- \( \beta_0 \) is a constant
- \( \beta_D \) is the estimated coefficient for increasing the ‘dose’, not controlling for the effect of time
- \( \beta_t \) is the estimated coefficient of time
- \( \beta_I \) is the estimated coefficient of increasing the ‘dose’, controlling for the effect of time
- \( D_i \) is level of the continuous treatment variable or ‘dose’
- \( t_i \) is a time dummy (1 if unit \( i \) is a post-intervention unit; 0 otherwise)
- \( e_i \) is a random error term

As with the DiD equation for a binary treatment, the treatment effect is given by \( \beta_I \), the coefficient of the interaction between the dose and the post-intervention time period dummy. It can be interpreted as the effect of changing the dosage by one unit, holding constant control variables included in the model.
The main empirical results of Linkow et al (2015) are:

- Both the binary and dose-response results show a significant increase in the number of industrial facilities within a settlement.
- The dose-response model estimated significant reductions (p-value < 0.1) in household transport expenditure, but this was not supported by the binary models.
- Varying impacts were found with respect to food prices: both models showed statistically significant (p-value < 0.05) increases in beef and potato prices, and no change to wheat and milk prices.
- No significant effects were found on land use outcomes.
- Strong effects on daily vehicle counts and vehicle speeds, yet no significant impact was found regarding the frequency of public transport services to Tbilisi or regional centres.
- No meaningful treatment effects were found on household income, consumption and employment. However, the authors noted the unbalanced nature of the household data limited their ability to control for other household level characteristics. Estimates might also be limited by selection bias, as project roads connected to two international borders, yet control roads did not. This would be problematic to estimates where the outcome of interest was correlated with accessibility to Turkey or Armenia.

4.8.3.2 India: impacts of highway improvements on firms

Datta (2010) analysed the effect on nearby firms of India’s Golden Quadrilateral Program, a project that improved highways connecting its four largest cities from 2002 to 2006. In this case, the problem of selection bias (or endogeneity) is particularly clear: treated roads are likely to have greater economic returns than control roads given their unique connectivity to the country’s main cities. The author used an ingenious strategy to overcome this endogeneity issue: if the improved highways were not deviated to include or exclude in-between areas, then ‘highway construction can be treated as exogenous to the areas that the highway runs through’ (Datta 2010, p4).

Datta used both binary and continuous treatment approaches. The binary method defined treated firms as those in 19 cities that ‘coincidentally’ lay on the improved highway, and control firms as those in 18 cities located elsewhere (both groups excluded the four nodal centres). In the dose-response methods, driving distance from a firm’s closest city on the improved road was the continuous treatment variable. DiD models were used to estimate the impacts of these treatments on firms’ inventory levels, choices of supplier and perceived difficulty of transporting production. The relevant city population and industry dummy variables were used as controls. Unobserved fixed effects were mitigated through the DiD model.

The dose-response model was only used to analyse the inventory outcome, and showed that increasing a firm’s distance from the improved highway increased the days of inventory held (p-value < 0.01). Treatment effects from the binary models supported this finding, and suggested that firms on upgraded highways were more likely to have changed suppliers (presumably in favour of ones that were more cost-effective) and were less likely to report transportation as a major issue for production. These findings align with theory; however, the data did not allow the parallel trends assumption to be tested, and consequently the results cannot be fully validated.

4.8.3.3 Japan: urbanisation effects on travel behaviour

Parady et al (2014) looked at the causal link between the built environment and travel behaviour in the city of Hiroshima, Japan. The authors highlight two limitations hindering similar estimates in past studies.

---

33 This was estimated from a binary treatment perspective only
First, the use of a binary treatment variable (e.g., urban vs. suburban areas) does not adequately represent the varying levels of urbanisation seen in the real world. Second, the problem of ‘self’ selection bias, i.e., residents choose to live in neighbourhoods to suit their transport preferences. Failure to control for underlying differences in people’s travel preferences will lead to biased treatment effect estimates. Parady et al. (2014) used a continuous urbanisation variable and PSM to overcome these two modelling issues.

As the standard PSM model can only be applied to treatments in a binary setting, Parady et al. (2014) used the generalised propensity score (GPS) technique, proposed by Imai and van Dyk (2004) and Hirano and Imbens (2004). This adaptation of traditional PSM is explained in box 4.3.

**Box 4.3 Generalised propensity scores for evaluation of continuous treatments**

GPS models are designed specifically for multi-valued or continuous treatments and are similar to classic PSM in many ways. An obvious difference is that GPS allows propensity scores to be continuous, whereas PSM fit propensity scores based on the binary assignment to treatment and control groups. Therefore, an OLS regression can be applied to estimate propensity scores, rather than a logit or probit model. Once scores are estimated, stratification or subclassification methods can be used to estimate treatment effects, where treatment and control units are compared within propensity score intervals (Hirano and Imbens 2004; Imai and van Dyk 2004).

Parady et al.’s data was primarily sourced from a 2013 survey of 600 individuals, which provided information about each respondent’s attitudes, household, transport use and job location. Outcome and control variables were taken directly from this data, while the treatment variable was estimated (i.e., a latent variable) from land use data pertaining to each respondent’s geographic area, including: population density, average area of housing per person, ratio of households living in multifamily residences, ratio of renting households and density of commercial facilities. Figure 4.9 shows the estimated level of urbanisation in the case study area.

**Figure 4.9 Map of urbanisation level in Hiroshima city**

For the first stage of estimation (estimation of propensity scores), Parady et al. (2014) predicted the generalised propensity score using the latent variable of urbanisation (dependent variable) and individual
characteristics (explanatory variables). Individuals were then stratified based on propensity scores and treatment effects were estimated within each interval. The average treatment effect was the weighted average of within-strata estimates. Results show increasing urbanisation had a negative effect on non-work car trip frequency and distance, and a positive effect on non-work active mode travel frequency and distance. Propensity score results were compared against those from a simple OLS regression approach. As hypothesised, the simple approach overestimates the treatment effect (by 6%-36%) due to its failure to control for selection bias. However, the propensity score estimates used by Parady et al may themselves suffer from bias owing to potential unobserved confounding variables.

4.8.4 Limitations

Aside from being applicable only in multi-levelled or continuous treatment settings, dose-response methods have the same limitations as their binary treatment counterparts. Additional restrictions will also be specific to the model used. For example, stratification in generalised propensity score estimation may be limited when strata do not contain units across the full spectrum of ‘doses’, making it impossible to analyse the full range of the dose-response relationship. Furthermore, because this approach is relatively new to causal inference, the theory and its applications are still developing.

4.9 Summary

Each technique described above has a different way of overcoming potential sources of bias in the estimation of treatment effects, given the evaluators’ presumptions regarding the source(s) of bias and the data available. In reality, potential confounding variables are both observed and unobserved. Thus, a combination of causal inference techniques can be used.

For example, Funderburg et al (2010) used a combination of PSM and DiD to estimate the impact of new highway investments to urban growth, while Ashenfelter and Greenstone (2002) used IV in support of DiD to analyse the impact of vehicle speed on fatalities and travel time savings. The simplicity of the DiD technique makes it a popular method of choice for causal inference estimation, either as the sole estimator or in conjunction with other techniques. Van de Walle (2009) also points out the benefits of combining causal inference techniques, recommending that, even if selection bias is known or suspected to come from observed factors (which would suggest the use of PSM), PSM can be used in conjunction with either DiD or IV regression to address the potential for unobserved selection bias. Ideally both could be used; however, the difficulty of finding a valid instrument may prohibit the use of IV.

Generally, reliable causal inference methods require data relating outcomes and relevant characteristics of both treatment and control units, recorded before and after the intervention. Some evaluations, however, will not meet these requirements. Often this is because an intervention’s scope is nationwide (prohibiting the use of a comparison group), or the available data does not include relevant characteristics of the units of observation (prohibiting the ability to directly control for selection bias). Under such restrictions, evaluators will have to employ time series or cross sectional methods. These methods generally cannot control for all of the factors that influence outcomes, and only provide causal estimates under a restrictive set of assumptions. This limits their ability to quantify and validate a causal relationship; however, if applied carefully this type of analysis can still be useful.
5 Evaluation and selection of techniques

This chapter evaluates methods for causal inference, provides for how to select a method, and summarises the key methods used in specific project/policy circumstances. We also give recommendations for setting up data collection prior to intervention, so a robust ex-post evaluation can be done.

5.1 Characteristics of methods

Causal inference techniques are distinguished by their data requirements, the method used to determine the counterfactual and ability to control for unobserved factors (table 5.1).

A rich dataset with pre- and post- intervention data for a set of treated and control units that have a range of similar characteristics, plus data on other factors expected to affect outcomes, allows for the control of observed and unobserved factors that affect outcomes, and ultimately best practice causal inference.

Table 5.1 Summary of causal inference methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Minimum data requirements</th>
<th>Explicit control group required</th>
<th>Unobserved factors controlled for</th>
<th>Key assumptions needed for causal inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross- sectional regression</td>
<td>Post intervention</td>
<td>• Outcomes • Treatment status • Control variables</td>
<td>✓</td>
<td>Conditional independence, ie assignment of units to the treatment or control groups is effectively random given the control variables included in the model. Common support- ie all units have a positive probability of being treated or not (note in practice this assumption is often not checked in cross- sectional models).</td>
</tr>
<tr>
<td>Panel regression</td>
<td>Pre and post intervention</td>
<td>• Outcomes • Treatment status • Control variables</td>
<td>✓</td>
<td>Confounding variables are captured in the data, or if unobserved, are constant over time or across groups. Common support. ie all units have a positive probability of being treated or not (note in practice this assumption is often not checked in panel regression models).</td>
</tr>
<tr>
<td>Propensity score matching</td>
<td>Post intervention</td>
<td>• Outcomes • Treatment status • Control variables</td>
<td>✓</td>
<td>Conditional independence, ie assignment of units to the treatment or control groups is effectively random given the control variables included in the propensity score. Common support, ie all units have a positive probability of being treated or not.</td>
</tr>
<tr>
<td>Difference-in-differences</td>
<td>Pre and post intervention</td>
<td>• Outcomes • Treatment status • Control variables</td>
<td>✓</td>
<td>Parallel paths, ie potential unobserved confounding factors are additive and have an equal influence on the treated and control groups at any point in time.</td>
</tr>
</tbody>
</table>
### Table: Evaluation and selection of techniques

<table>
<thead>
<tr>
<th>Method</th>
<th>Minimum data requirements</th>
<th>Explicit control group required</th>
<th>Unobserved factors controlled for</th>
<th>Key assumptions needed for causal inference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instrumental variables regression</strong></td>
<td>Post intervention</td>
<td>• Outcomes</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Treatment status</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Instrument (often pre-intervention)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Regression discontinuity</strong></td>
<td>Post intervention</td>
<td>• Outcomes</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Predictors of treatment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interrupted time series</strong></td>
<td>Pre and post intervention</td>
<td>• Outcomes</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Treatment status</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Empirical Bayes</strong></td>
<td>Pre and post intervention</td>
<td>Before and after outcomes for the treated location</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Pre-intervention outcomes and predictors for similar locations to calculate the prior, e.g., SPF.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multi-valued or continuous treatments</strong></td>
<td>Post intervention</td>
<td>• Outcomes</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Treatment status</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Control variables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5.2 Choosing a method based on available data

Figure 5.1 provides guidance for selecting a causal inference method after an intervention has occurred, based on the available data. Best practice methods are those that exploit the multidimensional nature of panel data if it is available (e.g., panel models and DiD) to separate the treatment effect from the effects of other observed and some types of unobserved factors.
Cross-sectional methods can deliver reliable causal estimates if, and only if, all factors affecting outcomes are measured and recorded in the data. Because this assumption is often difficult to satisfy, methods such as DiD, panel models, and IV are generally preferred over cross-sectional methods, but these have greater data requirements.

In all cases where treatment and control groups are compared, it is important for the characteristics of the two groups to have sufficient overlap. PSM can be used to help improve the ‘balance’ of the characteristics of the two groups, but it cannot make up for a lack of data in the first place, and in some cases, PSM involves dropping observations from the dataset. PSM can be combined with cross-sectional, panel and DiD models as an initial step prior to estimating the regression models to improve the balance of the treatment and control groups. Alternatively, the propensity scores generated by PSM can be used as weights in a regression model estimated by weighted least squares.

If a cross-sectional control group cannot be established, evaluators must use time series data to infer a causal effect. These methods (ITSA and empirical Bayes) require stronger assumptions to estimate a causal impact. For example, empirical Bayes methods rely on a large sample of reference units to correctly specify the prior and estimate the counterfactual. ITSA faces issues arising from the difficulty of modelling time series data, such as controlling for trends and seasonal factors that affect outcomes over time independent of the intervention being evaluated.

Analysis of post-intervention surveys is the least reliable method, yet may be an evaluator’s only option if appropriate data is not available. Such surveys seek to estimate the impacts of an intervention after the fact by asking people about its impacts. It is difficult to establish a reliable counterfactual as survey respondents may fail to properly recall the state of the world prior to the intervention, or may fail to understand exactly what counterfactual they are being asked to consider. Methods applying post-intervention survey data are not regarded as causal inference techniques, and hence not discussed in this report.

**Figure 5.1 Causal inference method selection flow chart**

*Regression Discontinuity (RD) is a special type of cross section regression used when there is no overlap between treatment and control groups.*
5.3 Data collection to enable causal inference analysis

When an intervention is being considered, it is important that evaluators specify the data to be collected for ex-post evaluation. If possible, data on outcomes and characteristics of treatment and control units should be collected ex-ante as well as ex-post to enable one of the more robust methods in figure 5.1 to be implemented. If a randomised experiment is not possible, ideally the data collected for evaluation will:

- be a panel dataset containing observations of pre- and post-intervention outcomes of the treatment and a control group
- include data on other factors that are expected to influence outcomes of either group, so these can be used as explanatory variables
- if random assignment to treatment and control groups is not possible, be assigned based on observed characteristics
- have a broad overlap in the characteristics of units in the treatment and control groups.

Prior to intervention, once the outcome(s) of interest are identified, evaluators need to carefully consider the following:

- What is the counterfactual scenario? Is it a continuation of the status quo, or will an intervention occur anyway?
- What units are a sufficient control group? Ideally, these units will have similar characteristics to the treatment units but will not be directly or indirectly affected by the intervention. In transport applications, this may require the treatment and control groups to be physically distant from each other.
- How and why the outcomes of the treated units may be different from the potential control units, in absence of an intervention? If the two groups are systematically different in ways unrelated to the intervention, these differences need to be measured so they can be controlled for when estimating the effect of intervention. In addition, if the results of the evaluation need to be generalised beyond the treatment group, there must be sufficient overlap in the characteristics of the treatment and control units.

5.4 Methods for evaluating different types of intervention

Transport interventions differ in terms of scale (local, regional or national) and objectives or target outcomes (e.g., road safety, accessibility, agglomeration benefits, emissions). As explained above, the choice of evaluation method is constrained by data availability, so we cannot provide definitive guidance on the choice of method for a given type of intervention.

However, certain types of intervention and certain types of outcomes lend themselves to particular techniques for ex-post evaluation, or make others infeasible. We list the key observations below:

- One-off local projects may be difficult to evaluate via PSM or other purely cross-sectional methods, as these require multiple control and treated units, but a suitable control group may be difficult to find for a one-off project and it may not be possible to define multiple treated units for the project (depending on the nature of the project);
- Agglomeration or accessibility benefits are often evaluated using IV regression due to its ability to overcome selection bias, i.e., areas chosen for new transport links/developments are more likely to have
more economic and employment potential than control areas. In such cases, long-term historical population data is often the instrumental variable of choice, as this variable (at least in theory) should be:

- relevant – accessibility interventions tend to target more populous areas
- exogenous – cannot be influenced by current levels of accessibility.

- Projects or policies applied in specific geographic areas are commonly evaluated using DiD or panel data methods as other (non-treated) areas can provide acceptable counterfactuals if data on pre- and post-intervention outcomes is available.

- Homogenous local or regional projects selected by the same CBA (eg multiple bridge widening projects that have undergone the same CBA) favour a RD design as those ‘nearly treated’ units may become appropriate counterfactuals. This is because homogenous projects with comparable net benefit prospects are likely to have similar characteristics and thus similar counterfactuals.

- National policies applied to the whole population or all geographic areas are often evaluated via ITSA or empirical Bayes methods, whereby pre-intervention outcomes become the ‘control’ units in absence of a proper control group.

5.5 Conclusions on methods

The methods have different data requirements so the data available limits the choice of method, as shown in figure 5.1. If considering methods for ex-post analysis prior to intervention, this enables plans to be put in place for the collection of the appropriate data.

Given full information, or the opportunity to collect appropriate data, the ‘best’ causal inference method is usually conditional on the characteristics of the intervention. Ideally, evaluators would

- randomly assign the treatment
- have a large number treated and control units
- measure as many important unit attributes as possible
- observe all of the relevant data before and after the treatment.

Although the random allocation condition is often not realistic in a policy setting, evaluators may be able to overcome the problems of non-random assignment (this is primarily the problem of omitted variable bias) if the other criteria are met. These conditions enable evaluators to:

- make statistical inferences from a sub-population (condition 2)
- control for important variables influencing the outcome (condition 3)
- exploit panel models to control for some unobserved variables influencing the outcome (condition 4)

If these conditions are fulfilled, the PSM-DiD model is perhaps the simplest but most effective technique for mitigating against different sources of omitted variable bias. This model uses PSM to select the best possible counterfactual units given the observed variables, and then applies a DiD regression to control for certain types of unobserved variables.
6 Method application guidance

This section provides practical guidance on how to do ex-post analysis of transport interventions using causal inference methods. Techniques covered include:

- simple methods using cross-sectional regression models
- difference-in-differences methods
- fixed effects regression with panel data
- PSM, on its own or in combination with other methods
- IV regression
- ITSA.

All these methods except PSM are illustrated in this section using the implementation of safer speed areas in Hamilton city as a simple case study (see section 6.1). The small size of this dataset meant PSM could not be implemented successfully; however, an additional case study of Auckland’s Northern Busway in chapter 7 illustrates PSM and shows how PSM can be combined with other techniques such as DiD. In addition to this report we have provided an Excel spreadsheet, and R and STATA code that implements most of the methods described above, using the Hamilton case study.

The analysis of the safer speed areas intervention in Hamilton is intended to be for illustrative purposes only. A full analysis of this intervention would need to consider a wider range of factors and impacts on safety. In addition, the analysis is limited by a small treatment group and a short time-frame. We therefore do not provide overall conclusions on the effectiveness of this intervention in this report.

6.1 Hamilton city safer speed areas case study

Hamilton City Council’s (HCC) implementation of safer speed areas reduced speed limits from 50km/h to 40km/h on 10 streets in eight different residential areas around Hamilton city in 2011. Speed limits of 40km/h were chosen according to research that suggests a large reduction in the probability of death when impact speeds are decreased from 50 to 40km/h (ACC and LTSA 2000).

The affected streets were (Turner et al 2014):
1 Barrington Drive located in the Huntington area
2 Heath Street located in the Saint Andrews area
3 Forsyth Street located in the Saint Andrews area
4 Casey Avenue located in the Fairfield area
5 Willoughby Street located in the Whitiora area
6 Hammond Street located in the Hamilton Lake area
7 Blackburn Street located in the Dinsdale area
8 Lethborg Street located in the Dinsdale area
9 Howell Avenue located in the Riverlea area
10 Hudson Street located in the Riverlea area.

34 www.saferspeedarea.org.nz/
Ex-post evaluation of transport interventions using causal inference methods

We were unable to include all 10 streets in the analysis. Some streets either had too few crashes for valid statistical inferences (1 or 0 over the analysis period), or had no available traffic count data (an important explanatory variable). In some cases, the first issue was overcome by combing connected streets into a single unit (e.g., Lethborg St and Blackburn St). Our treatment group is:

1. Barrington Drive
2. Heath Street
3. Willoughby Street
4. Dinsdale Road
5. Howell Ave

Although Hamilton’s safer speed areas were initially part of a national demonstration project, they were made permanent by the HCC in September 2011. The intervention was carried out in two distinct phases (Turner et al. 2014). First, engineering devices were installed (where necessary) to slow traffic to within 5 km/h of the 40 km/h speed limit. The second phase introduced speed limit signs and road markings. Given the short timeframe between these two phases, our analysis looks at the effect of the safer speed areas intervention as a whole, rather than the individual impacts of the two phases of implementation.

The HCC considered the 2011 safer speed areas a success and introduced new areas in 2013 and 2014; however, our analysis relates only to the original intervention areas and roads. We used data from 2009 and 2010 to measure the number of crashes prior to intervention, and data from 2012 and 2013 to measure post-intervention outcomes. The analysed data includes 29 different roads, comprising six treated roads and 23 control roads in Hamilton city.

Control roads were selected primarily on whether relevant traffic count data could be retrieved. Traffic volume is an important factor that should be controlled for in the evaluation of the impact of the safer speed areas on crash count. Such data is available for many Hamilton city roads on the HCC website. Control sites located in residential areas, i.e., with a 50 km/h speed limit, were also prioritised. We specifically excluded streets on the border of the safer speed areas from the control group as these streets might have also been affected by the intervention as motorists possibly did not immediately realise they had driven out of the reduced speed limit area.

Data and code files for this case study are provided in appendix C. The key variables in the dataset are:

- **crash count:** total for a given street or combination of streets during the pre- and post-intervention periods (2009–10 and 2012–13 respectively)
- **treated:** a dummy variable indicating whether a street was treated or not (1 = treated)
- **vehicle kilometres travelled (VKT):** the pre-treatment average daily traffic on a given street multiplied by the length in km, used as a measure of traffic ‘exposure’ on each street
- **cross intersections:** number of cross intersections on a given street
- **T intersections:** number of T intersections on a given street.

---

6.2 Simple cross-sectional regression models

In this section, we describe the estimation of treatment effects using simple cross-sectional regression models of post-intervention outcomes. As discussed in section 4.1, such methods do not usually produce reliable estimates of causal effects due to concerns about bias created by omitted variables. However, we have included them here for comparison with other techniques and because more sophisticated analysis is not always possible due to data constraints.

Cross-sectional analysis may also be used when some data is available over time but for a limited number of periods or if observations for different units are only available at different points in time. In such cases, it may be necessary to treat all data as if it was at the same point in time, and combine it into a ‘pooled’ cross-section. While such pooled cross-sectional analysis is generally inferior to the use of panel data models or DiDs, in some cases it may be all that is possible due to data limitations.

The discussion below focuses on cross-sectional regressions of post-intervention outcomes for a treatment group and a control group. It is also possible to estimate simple regression models that compare pre- and post-intervention outcomes for a treatment group alone. In such a case, the pre-intervention outcomes are used to establish the counterfactual. However, it is difficult to generalise the results of such analysis beyond the specific treatment group because the absence of an explicit control group means there is some chance the observed effects of the intervention are due to the characteristics of the units in the treatment group.

6.2.1 Data and software requirements

Cross-sectional regression requires data on post-intervention outcomes from treated and control units (e.g., sections of road, intersections, area units, firms or people). For reliable analysis, the number of units in both groups (sample size) should be reasonably large. As a rule of thumb, around 25 units are needed for valid statistical inferences, and preferably about half of these observations would be in the treatment group. Ideally, all variables (aside from treatment) influencing the outcome of interest should also be included in the analysis dataset. These variables must be measured prior to the intervention.

Cross-sectional regressions that assume the outcome of interest is normally distributed can be done in Excel or other spreadsheet software. However, there are two cases where this may not be appropriate:

**Count data**: if the outcome of interest is a count that cannot be negative values, e.g., the number of crashes in a year, it is better to use regression models that assume a Poisson or negative-binomial distribution. This type of model is not available in Excel’s built-in regression analysis.

**Heteroscedasticity**: this is a phenomenon where the variance of residuals (around the fitted regression line) increase or decrease with an explanatory variable (see figure 6.1). In such cases, regression models with ‘robust’ standard errors should be employed. Heteroscedasticity can only be identified after a regression has been fitted; methods to detect this problem are briefly explained in the following section.

Count data and heteroscedasticity can be modelled correctly through special functions in Excel add-in packages or other analytical software such as R or Stata.


6.2.2 Cross-sectional regression in Excel

6.2.2.1 Data layout

Data should be organised so each row contains a specific unit and its outcome level, treatment status, and explanatory variable levels. Table 6.1 demonstrates this layout using part of the HCC safer speed areas data (see appendices C and D). Explanatory/control variables recorded in the data should be observed pre-treatment. Post-treatment observations of controls should only be used if they are fixed over time, and thus equal to their pre-treatment level (Gelman and Hill 2007).

Table 6.1 Example of cross-sectional data format

<table>
<thead>
<tr>
<th>Unit name</th>
<th>Post intervention crash count</th>
<th>Treated</th>
<th>Pre-intervention crash count</th>
<th>Pre-intervention VKT</th>
<th>No. of cross intersections</th>
<th>No. of T intersections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrington Dr</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>501,240</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Heath-Mahana St</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2,498,505</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Willoughby St</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>2,757,341</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Dinsdale Rd</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3,533,200</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Crosby Rd</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2,730,200</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Hillcrest Rd</td>
<td>7</td>
<td>0</td>
<td>6</td>
<td>4,865,450</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Maeroa Rd</td>
<td>6</td>
<td>0</td>
<td>8</td>
<td>5,323,890</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

6.2.2.2 Exploratory analysis

Table 6.2 shows the mean values of each variable in the HCC safer speed areas dataset. In this example, treated and control streets have different characteristics, eg treated streets have a significantly lower VKT. Such imbalance of treatment and control group characteristics is a sign of non-random assignment, which must be considered in any estimates of causal effects. It seems likely that lower VKT accounts for some of the observed difference in crash count between the treated and control groups.

Table 6.2 also shows a simple comparison of the mean post-intervention crash count for the treated and control groups. The difference is highly statistically significant on the basis of a simple t-test, with roads in the treatment group having, on average, about six fewer crashes in the post-intervention period compared with roads in the control group. This naive comparison suggests the intervention was highly successful at reducing crashes in the treatment group of roads relative to the control group.

However, such a simple comparison does not account for any systematic differences between the treatment and control group that may have affected the number of crashes. As well as the difference in...
VKT already noted, it is clear from table 6.2 that the mean crash count in the pre-intervention period for the treated group is lower than the control group. As we will see below, more sophisticated analysis leads to substantially lower estimates of the treatment effect once the observed differences between the treatment and control groups are considered.

Table 6.2  Example of average values by treatment status

<table>
<thead>
<tr>
<th>Treatment status</th>
<th>Mean post-intervention crash count</th>
<th>Mean pre-intervention crash count</th>
<th>Mean pre-intervention VKT</th>
<th>Mean no. of cross intersections</th>
<th>Mean no. of T intersections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated group</td>
<td>2.17</td>
<td>4.50</td>
<td>1,814,090</td>
<td>1.17</td>
<td>4.17</td>
</tr>
<tr>
<td>Control group</td>
<td>8.13</td>
<td>9.70</td>
<td>6,489,160</td>
<td>0.83</td>
<td>4.65</td>
</tr>
<tr>
<td>Difference (treated – control)</td>
<td>-5.96</td>
<td>-5.20</td>
<td>-4,675,070</td>
<td>0.34</td>
<td>-0.49</td>
</tr>
<tr>
<td>P-value (treated vs control)</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.64</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Correlation values indicate the extent to which the variation of one variable is like that of another. In Excel, correlations between two different arrays can be calculated through the ‘CORREL’ function. Table 6.3 shows a correlation matrix of part of the HCC speed zone data. Strong correlations, ie those greater than 0.50 or less than -0.50, are highlighted.

The correlation matrix indicates the variables likely to be important in a regression model, ie those strongly associated with the outcome variable. It also highlights explanatory variables that are highly correlated with each other. In a regression model this is known as multicollinearity, and regression coefficients of highly correlated variables can become unreliable as they can change dramatically in response to small changes in model specification. Therefore, multicollinearity poses a threat in causal inference if it involves the treatment variable.

The correlation matrix of table 6.3 tells us there is a strong relationship between post-intervention crash count and VKT (0.70), and a moderately strong relationship between VKT and number of T intersections (0.55). The pre-intervention crash count is also moderately correlated with the number of cross- and T-intersections, and the post-intervention crash count is highly correlated with the pre-intervention crash count. Even though treated streets supposedly had relatively high crash counts, the treatment indicator has a weak negative correlation with the pre-intervention crash count (-0.25). Fortunately, there is no sign of multicollinearity involving the treatment indicator variable.

Table 6.3  Correlation matrix

<table>
<thead>
<tr>
<th>Treatment status</th>
<th>Post intervention crash count</th>
<th>Treatment indicator</th>
<th>Pre intervention VKT</th>
<th>Pre-intervention crash count</th>
<th>No. of cross intersections</th>
<th>No. of T intersections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crash count</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treated</td>
<td>-0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VKT</td>
<td>0.70</td>
<td>-0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-intervention crash count</td>
<td>0.82</td>
<td>-0.25</td>
<td>0.85</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of cross intersections</td>
<td>0.44</td>
<td>0.11</td>
<td>0.38</td>
<td>0.54</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>No. of T intersections</td>
<td>0.49</td>
<td>-0.06</td>
<td>0.55</td>
<td>0.56</td>
<td>0.05</td>
<td>1.00</td>
</tr>
</tbody>
</table>
6.2.2.3 Implementation in Excel

Regression in Excel can be done with the ‘LINEST’ function or the regression analysis wizard. We explain the steps to fitting regression through the LINEST function in box 6.1.

**Box 6.1 Excel’s LINEST function**

The LINEST function uses the ‘least squares’ method to estimate a linear regression equation that best fits the selected data, and then returns an array that describes the fitted model. LINEST requires:

- known ‘y’ s – the array of outcomes (dependent variable) for the regression
- known ‘x’ s – the array of explanatory variables for the regression
- const – select TRUE to include a constant (intercept)
- stats – select TRUE to include regression statistics in the output values.

Because LINEST is an array function, an array of blank cells must be selected before entering the formula. The selected array must contain five rows and \(k+1\) columns, where \(k\) is the number of explanatory variables used in the regression.

The table below describes the LINEST output when three explanatory variables have been used. The LINEST output presents estimated regression coefficients in reverse order in the first row and their corresponding standard errors in the second row. Regression statistics are contained in three rows in the first two columns below these values.

<table>
<thead>
<tr>
<th>Coefficient (3rd or last explanatory variable)</th>
<th>Coefficient (2nd explanatory variable)</th>
<th>Coefficient (1st explanatory variable)</th>
<th>Coefficient (constant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error (3rd or last explanatory variable)</td>
<td>Standard error (2nd explanatory variable)</td>
<td>Standard error (1st explanatory variable)</td>
<td>Standard error (constant)</td>
</tr>
<tr>
<td>(R^2)</td>
<td>Regression standard error</td>
<td>#N/A</td>
<td>#N/A</td>
</tr>
<tr>
<td>F statistic</td>
<td>Residual degrees of freedom</td>
<td>#N/A</td>
<td>#N/A</td>
</tr>
<tr>
<td>Sum of square errors</td>
<td>Sum of squared residuals</td>
<td>#N/A</td>
<td>#N/A</td>
</tr>
</tbody>
</table>

Note: Excel puts \#N/A in the bottom-right cells when multiple explanatory variables are used.

The simplest possible cross-sectional regression model involves the outcome variable regressed on the treatment variable and a constant:

\[
Crash\ count_i = \beta_0 + \beta_1 \cdot Treated_i + e_i
\]  

(Equation 6.1)

Where:

- \(Crash\ count_i\) is the outcome for each unit, \(i\)
- \(\beta_0\) is a constant to be estimated
- \(\beta_1\) is the estimated treatment effect
- \(Treated_i\) is the treatment indicator for unit \(i\)
- \(e_i\) is an error term representing the unexplained variation in the crash count, assumed to have zero mean and constant variance.
Estimating this simple model corresponds to performing a t-test for the difference in mean crash count between the treatment and control groups. The t-test performed on a regression model such as this assumes equal variance of crash counts in the two groups. Alternatively, an unequal variance t-test can be performed using Excel's built-in data analysis functions. The results are summarised in table 6.4.

Table 6.4 Cross-sectional regression model output

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>8.13***</td>
<td>1.31</td>
</tr>
<tr>
<td>Treated</td>
<td>-5.96**</td>
<td>2.90</td>
</tr>
<tr>
<td>R²</td>
<td>0.14</td>
<td></td>
</tr>
</tbody>
</table>

A better approach to estimating treatment effects with a simple regression model is to include other explanatory variables, aside from the treatment indicator, which may have affected the outcome of interest. For example, equation 6.2 specifies a regression model for the HCC data that controls for other characteristics of the roads in the treatment and control groups.

\[
\text{Crash count}_i = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \text{VKT}_i + \text{Pre}_i \cdot \text{crash count}_i + \beta_4 \cdot X \cdot \text{intersections}_i + \beta_5 \cdot T \cdot \text{intersections}_i + e_i
\]  

(Equation 6.2)

The results are presented in table 6.5. This model explains a much greater percentage of the variation in crash counts (R² = 0.71) than the simple model above (R² = 0.14 in table 6.4). The treatment effect has reduced to -3.80 and is now only statistically significant at the 10% level.

Table 6.5 Cross-sectional regression model output including VKT and intersections variables

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.05</td>
<td>1.41</td>
</tr>
<tr>
<td>Treated</td>
<td>-3.80*</td>
<td>2.13</td>
</tr>
<tr>
<td>Pre-intervention VKT</td>
<td>-1.71E-07</td>
<td>2.96E-07</td>
</tr>
<tr>
<td>Pre-intervention crash count</td>
<td>0.58***</td>
<td>0.20</td>
</tr>
<tr>
<td>Cross intersections</td>
<td>0.54</td>
<td>0.81</td>
</tr>
<tr>
<td>T intersections</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>R²</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>

If all predictors in the model are statistically significant, the model is likely to be specified 'as best as possible' given the available data. However, models with coefficients which are not significant at the 10% level (p > 0.10) may be better specified by removing factors that poorly explain the outcome variable. Models with a reduced number of predictors are known as 'restricted' or 'reduced' models, and of course the treated variable will remain in the reduced model regardless of its statistical significance. If the
treatment effects estimated in the full and restricted regressions differ substantially, an F-test can be applied to decide on the preferred model specification and treatment effect.36

In our example, neither of the treatment effects estimated in the full and or a restricted model that excludes the intersections variables are statistically significant at the 5% level. Therefore, our cross-sectional regression analysis suggests the safer speed areas intervention cannot be shown to have reduced the number of vehicle crashes in the treated areas, once we incorporate observed differences between the treatment and control groups in our analysis. Instead, most of the difference in post-intervention crash counts between the treatment and control groups appears to be attributable to differences in the pre-intervention crash counts.

6.2.3 Cross-sectional regression with Poisson or negative binomial models

Stata and R (and other packages) can estimate the cross-sectional models described above under more appropriate distributions for crash count data, such as Poisson and negative binomial distributions as discussed in section 4.1.3.

In their standard form, Poisson and negative binomial models assume there is a log-linear relationship between the crash count and the explanatory variables (see section 4.1.3). This means that to interpret the coefficient on the treatment dummy variable in a cross-sectional regression we must first exponentiate it. The resulting figure is then the multiplicative effect of the intervention on the dependent variable. To calculate a treatment effect that is comparable with the linear models presented in section 6.2.2, we can multiply the estimated effect of the intervention by the mean of the dependent variable for the control group.

For example, in the safer streets case study, a simple negative binomial model with no other explanatory variables estimates a coefficient on the treatment effect of -1.32. This implies the intervention multiplies the crash count by \( \exp(-1.32) = 0.27 \). The mean post-intervention crash count of the control group is 8.13, therefore the estimated treatment effect is \( 0.27 \times 8.13 = 5.96 \), ie the intervention is estimated to reduce the mean crash count per street by about six crashes during the two-year treatment period.

When controls (pre-intervention VKT, pre-intervention crash count, cross-intersections, and T-intersections) are added to these models, the estimated coefficient on the treatment indicator \(-0.89\) for a negative binomial model, which implies treatment multiplies the crash count by \( \exp(-0.89) = 0.41 \). Since the effect is multiplicative and there are other explanatory variables in the model, there is no longer a single estimate of the reduction in crash count estimated to be caused by the intervention. Instead, an estimate can be calculated by using the fitted model to predict crash counts for all units with the treatment dummy variable set to zero and set to one. The difference in these predictions for each unit is the estimated effect of treatment given the characteristics of each unit. These differences can then be averaged across all units to arrive at an average treatment effect that is comparable to the linear models in section 6.2.2. For the safer streets example we estimate an average reduction of 4.55 crashes per street in the treatment period.

Table 6.6 summarises the treatment effects across four cross-sectional regression models. Of these, the best estimate is the ‘simple negative binomial plus controls’ model, given that it specifies an appropriate distribution and controls for different characteristics of the treatment and control groups.

---

36 This test estimates whether the residual sum of squares in the reduced model is significantly greater than that of the full model, in other words, whether the variables removed from the model are jointly valuable to the regression specification. Independent analysis of each variable’s statistical significance will not suffice as this ignores the joint significance (or importance) of variables to the model. Applying an F-test on our data shows that the number of X and T intersections are not jointly significant.
These results suggest that assuming a more appropriate distribution for crash counts restores the statistical significance of the safer streets intervention, in comparison with the regression results assuming a normal distribution reported in tables 6.4 and 6.5. However, the estimated impact of the intervention is smaller if characteristics of the streets are controlled for. These results suggest the intervention reduced the number of crashes per street during the two-year treatment period by about 4.5 per street.

<table>
<thead>
<tr>
<th>Outcome variable = crash count</th>
<th>Normally distributed crash count</th>
<th>Normally distributed crash count + controls for characteristics</th>
<th>Negative binomial distribution for crash count</th>
<th>Negative binomial distribution for crash count + controls for characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment effect</td>
<td>- 5.96**</td>
<td>- 3.80*</td>
<td>- 5.96***</td>
<td>- 4.55**</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

6.2.4 Summary

Steps to carrying out causal inference using cross-sectional regression are summarised below:

1. **Organise the data**: rows should contain a unique unit (observation) and its outcome level, treatment status and other co- variate levels.

2. **Explore the data**: check for balance in observed characteristics between the treated and control groups. Identify predictors that are highly correlated with the treatment variable and, if possible, transform problematic variables to overcome multicollinearity.

3. **Fit a ‘full’ or ‘unrestricted’ regression** (include all explanatory variables). Use an appropriate distribution assumption, eg Poisson or negative binomial for crash counts.

4. **Fit a ‘reduced’ or ‘restricted’ regression** (remove insignificant explanatory variables). Use an appropriate distribution assumption, eg Poisson or negative binomial for crash counts.

5. **Compare full and reduced models** via the F-test. Re-introduce omitted variables (or test them independently) if they are found to be jointly significant. Identify preferred regression specification.

6. **Identify treatment effect** using the preferred model specification.

6.3 Difference-in-differences methods

DiD methods can relax some of the restrictive assumptions required to estimate valid treatment effects in the cross-sectional regression models described above. DiD models control for the effects of unobserved variables that are fixed across units yet vary over time (ie time- specific effects), or fixed over time but vary by unit (ie unit- specific effects).

There are two main ways a DiD estimate of a treatment effect can be obtained:

1. **By doing a simple comparison of the means of the treatment and control groups, before and after the intervention. A t-test can be used to check whether the treatment effect is statistically significant, assuming outcomes are normally distributed. If the normality assumption is not appropriate, a non-parametric test can be used.**
2 By estimating an appropriately specified regression model. This enables factors that do not have a constant effect across units or time to be controlled for by including them as explanatory variables in the regression model.

Both methods are described below. The first method has the advantage of being very simple to implement; however, regression-based methods may give more accurate estimates of the treatment effect if other factors need to be controlled for as described above.

It is also possible to combine either form of DiD estimation with PSM techniques for greater accuracy. We discuss these more advanced techniques in section 6.5 below.

6.3.1 Simple DiD analysis via comparisons of means

Essentially, the DiD estimate of a treatment effect is given by:

\[ \text{Treatment effect} = \text{Difference in outcomes for the treatment group} - \text{Difference in outcomes for the control group} \]

where in each case the ‘difference in outcomes’ is a comparison of post- and pre-intervention outcomes for the relevant group. This calculation is performed using the mean outcomes for each group in each time period:

\[ \text{Treatment effect} = (\text{Mean post-intervention outcome for treatment group} - \text{Mean pre-intervention outcome for treatment group}) - (\text{Mean post-intervention outcome for control group} - \text{Mean pre-intervention outcome for control group}) \]

Thus, a DiD estimate can be obtained by calculating the four relevant means, and using them to calculate the treatment effect as described above. This is straightforward to do in a spreadsheet.

Table 6.7 illustrates this method using the Hamilton safer streets case study. For the treatment group, the mean crash count was 4.50 crashes per street in the pre-intervention period and 2.17 in the post-intervention period. For the control group the respective means are 9.70 and 8.13. The DiD calculation involves first calculating the post- vs pre-intervention difference in means for each group: -2.33 for the treatment group and -1.57 for the control group. The DiD estimate of the treatment effect is then the difference between these two differences, ie -0.77. This implies the safer streets intervention reduced the average number of crashes per street in the two-year treatment period by 0.77.

<table>
<thead>
<tr>
<th></th>
<th>Treatment group</th>
<th>Control group</th>
<th>Difference in difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-intervention</td>
<td>4.50</td>
<td>9.70</td>
<td></td>
</tr>
<tr>
<td>Post-intervention</td>
<td>2.17</td>
<td>8.13</td>
<td>0.77</td>
</tr>
</tbody>
</table>

If the data is a balanced panel (ie the units in the control and treatment groups do not change over time) then the DiD estimate of the treatment effect can also be calculated by calculating the post- vs pre-intervention change in outcome for each unit, averaging those changes over the treatment and control groups, and taking the difference in those means. For example, in the safer streets case study the mean change in crash count for the treatment group is -2.33 and for the control group is -1.57. The difference between these two means is also -0.77.

With balanced panel data it is also possible to perform a basic test of statistical significance of the DiD estimate of the treatment effect in Excel. This can be done by applying the t-test function to the post- vs pre-intervention differences in crash count for the treatment and control groups. The options for a two-
sided, unequal variance t-test should be specified. In the safer streets example this gives a p-value of 0.60, ie there is no evidence the treatment effect is significantly different from zero.

### 6.3.2 DiD regression models

DiD can also be estimated using regression models, in Excel and other statistical software. Estimating a DiD regression requires outcomes measured for treated and control units before and after an intervention. Unlike other panel data methods, the pre- and post-intervention outcomes can be measured on entirely different units (ie repeated cross-sections where the cross-section changes over time). For example, a treated unit included in the pre-intervention group may (or may not) be recorded post-intervention.

If the data is a ‘fully balanced’ panel (ie outcomes are measured for the same set of units over time) with exactly two time periods, the treatment effects estimated by DiD will be the same as a panel fixed effects model (see section 6.4), and hence a fixed effects model is another way of obtaining a DiD estimate, if the data is a balanced panel with two time periods. However, in other cases (where the panel is not balanced and/or if there are more than two time periods), DiD and fixed effects models will give different results.

In addition to outcomes data, a time series of historical outcomes for treated and control groups is useful for checking the ‘parallel paths’ assumption of DiD (explained below). It is also possible to include other explanatory variables, such as VKT in the Hamilton case study that may not have constant effects on outcomes across time or across units.

#### 6.3.2.1 Data layout

Table 6.8 uses some of the Hamilton city case study observations to demonstrate the layout required for DiD regression in Excel. Essentially, the data can be categorised into four groups:

1. Treated outcomes (and covariates) before the intervention
2. Control outcomes (and covariates) before the intervention
3. Treated outcomes (and covariates) after the intervention
4. Control outcomes (and covariates) after the intervention.

The Hamilton case study data is an example of a ‘fully balanced’ panel where pre- and post-intervention outcomes are measured for the same cross-section of roads. However, as explained above, DiD is still feasible when the cross-section changes over time.

Three dummy variables are required in DiD regression:

- treatment dummy – coded 1 for treated roads and 0 for controls, regardless of the time period
- period dummy – coded 1 for rows relating to the post-intervention period, 0 for the pre-intervention period
- interaction dummy – the product of the treatment and time period dummies.

Explanatory variables that change over time and between units, such as VKT, should also be included in a DiD model.

---

37 The t-test also assumes the differences in crash counts for the two groups are normally distributed. If this is not true, an alternative test may need to be used. One possibility is a non-parametric test such as the Wilcoxon rank-sum test. However, Excel does not have built-in functions to perform such tests.
### 6.3.2.2 Checking the ‘parallel paths’ assumption

A critical assumption of DiD analysis is the trends in the outcome variable for the treatment and control groups would be ‘parallel’ if the treatment had not occurred. This assumption can be validated to some extent through analysing past trends. For example, figure 6.2 shows historical crash counts for treated and control groups from the HCC safer speed areas dataset. Their trends have converged over the three years prior to the 2011 intervention, meaning control sites appear to be a reasonable counterfactual for the treated group.

#### Figure 6.2 Historical paths of treated and control sites

![Image showing historical crash counts for treated and control sites](image-url)
6.3.2.3 DID regression analysis

DiD regression can be done using the LINEST function or the regression analysis wizard. A basic DiD analysis involves estimating a regression model where the three dummy variables (treated, time period and interaction) are the only explanatory variables. Applying this to the Hamilton case study yields the following equation and output:

\[
\text{Crash count}_{it} = \beta_0 + \beta_1 \cdot \text{Treated}_{it} + \beta_2 \cdot \text{Period}_{it} + \beta_3 \cdot \text{Interaction}_{it} + e_i
\]  

(Equation 6.2)

<table>
<thead>
<tr>
<th>Table 6.9 Summary: simple DiD model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome variable = change in crash count</strong></td>
</tr>
<tr>
<td>Explanatory variables</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Treated</td>
</tr>
<tr>
<td>Period</td>
</tr>
<tr>
<td>Interaction</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

In this model, the estimated coefficient on the interaction variable is the estimated treatment effect. This basic DiD model estimates a treatment effect of -0.77 which is not statistically significant at the 10% level. This model also estimates that, on average, treated streets have 5.2 fewer crashes than control streets, and streets in the post-intervention period have 1.57 fewer crashes. However, none of these effects are statistically significant at the 10% level.

The basic DiD model can be improved if data relating to explanatory variables that vary over time and by unit are available. In the HCC analysis, VKT is an example of such a variable. Other control variables in the case study data, such as the numbers of T and cross intersections, have unit-specific effects that do not vary over time. Such effects are controlled automatically in the DiD model and should not be included as explanatory variables.

An equation and regression summary (table 6.10) of the DiD model ‘plus controls’ is specified below. The estimated treatment effect is again the coefficient on the interaction term. Including VKT as an explanatory variable has further reduced the estimated treatment effect to -0.32 and it remains statistically insignificant. The VKT variable is highly statistically significant.

\[
\text{Crash count}_{i} = \beta_0 + \beta_1 \cdot \text{Treated}_{i} + \beta_2 \cdot \text{Period}_{i} + \beta_3 \cdot \text{Interaction}_{i} + \beta_4 \cdot \text{VKT}_i + e_i
\]  

(Equation 6.3)

<table>
<thead>
<tr>
<th>Table 6.10 Summary: DiD with controls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome variable = crash count</strong></td>
</tr>
<tr>
<td>Explanatory variables</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Treated</td>
</tr>
<tr>
<td>Period</td>
</tr>
<tr>
<td>Interaction</td>
</tr>
<tr>
<td>VKT</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01
6.3.3 Summary

The steps for carrying out causal inference using DiD regression are summarised below:

1. **Organise the data:** into the format shown in figure 5.1.

2. **Explore the data:** if using additional explanatory variables check for balance in characteristics between the treated and control groups. Identify predictors that are highly correlated with the treatment variable and, if possible, transform problematic variables to overcome multicollinearity.

3. **Fit a DiD regression:** include treatment, time period, and interaction dummies, and other explanatory variables if appropriate. F-tests can be used to exclude insignificant explanatory variables if necessary.

4. **Identify treatment effect** as the coefficient on the interaction term.

6.4 Panel fixed-effects models

Panel fixed-effects models exploit the nature of panel data to control for unobserved factors that have fixed effects over time on outcomes for each unit. Essentially, fixed-effects models are a simplified version of the DiD model when data is in panel form. An advantage of panel data methods is that they can use information from more than two time periods whereas DiD methods are limited to comparing two periods. On the other hand, estimation of panel models requires that at least some of the units in the cross-section remain the same over time, while DiD can be applied even if the cross-section is a completely different set of units at each point in time.

6.4.1 Data and software requirements

Fixed-effects regression requires panel (or ‘longitudinal’) data on treated and control units over two or more time periods (e.g. pre- and post-intervention). An important criterion for estimating treatment effects via a fixed-effects model is that explanatory variables that vary over time and units (i.e., without a ‘fixed’ component) must be recorded in the panel data. Fixed-effects models can be estimated in Excel and other statistical software.

In addition to this data, a time series of historical outcomes for treated and control groups is also useful for checking the ‘parallel paths’ assumption. A control group is deemed an appropriate counterfactual if its past trend is parallel (or at least very similar) to that of the treated group.

6.4.2 Fixed-effects regression in Excel

6.4.2.1 Data layout

There are various methods for estimating fixed-effects models. The simplest method to implement in a spreadsheet is a regression of the change in the dependent (outcome) on the change in explanatory variables and a treatment dummy variable (e.g. table 6.11).
Table 6.11  Panel fixed-effects regression data example

<table>
<thead>
<tr>
<th>Unit name</th>
<th>Change in crash count (post-intervention vs pre-intervention)</th>
<th>Treated</th>
<th>Change in VKT (post-intervention vs pre-intervention)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrington Dr</td>
<td>0</td>
<td>1</td>
<td>-54,728</td>
</tr>
<tr>
<td>Heath-Mahana St</td>
<td>-4</td>
<td>1</td>
<td>-696,946</td>
</tr>
<tr>
<td>Willoughby St</td>
<td>-5</td>
<td>1</td>
<td>-1,578,377</td>
</tr>
<tr>
<td>Dinsdale Rd</td>
<td>-2</td>
<td>1</td>
<td>-220,825</td>
</tr>
<tr>
<td>Crosby Rd</td>
<td>4</td>
<td>0</td>
<td>-80,300</td>
</tr>
<tr>
<td>Hillcrest Rd</td>
<td>1</td>
<td>0</td>
<td>282,875</td>
</tr>
<tr>
<td>Maeroa Rd</td>
<td>-2</td>
<td>0</td>
<td>-313,170</td>
</tr>
</tbody>
</table>

6.4.2.2  Exploratory analysis
Like DiD regression, the parallel trends assumption should be tested prior to the estimation of a fixed-effects model. See section 6.3.2.2 for an example of this analysis.

6.4.2.3  Panel fixed-effects regression analysis
Fixed-effects regression can be carried out through the LINEST function or the regression analysis wizard. In its simplest form, fixed-effects estimation involves regressing the change in outcomes on a treatment dummy variable. Applying this model to the Hamilton case study data yields the following equation and regression output (table 6.12):

\[ \Delta \text{Crash count}_i = \beta_0 + \beta_1 \cdot \text{Treated}_i + e_i \]  
(Equation 6.4)

Table 6.12  Summary: simple panel fixed-effects model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.57</td>
<td>1.04</td>
</tr>
<tr>
<td>Treated</td>
<td>-0.77</td>
<td>2.29</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

This model estimates a treatment effect of -0.77, which is not statistically significant at the 10% level or below. Since this data is the special case where outcomes are measured in exactly two time periods on the same set of units in each period, the estimated treatment effect from the fixed-effects model is the same as the DiD estimate in table 6.9 above.

This estimate will be biased if we have omitted an explanatory variable that changes over time and differs across units. VKT is an example of such a variable. Therefore, we repeat this model and introduce ‘change in VKT’ as an additional explanatory variable (output in table 6.13):

\[ \Delta \text{Crash count}_i = \beta_0 + \beta_1 \cdot \text{Treated}_i + \beta_2 \cdot \Delta \text{VKT} + e_i \]  
(Equation 6.5)
Table 6.13 Summary: panel fixed-effects model (controlling for VKT)

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.42</td>
<td>1.02</td>
</tr>
<tr>
<td>Treated</td>
<td>-0.33</td>
<td>2.26</td>
</tr>
<tr>
<td>VKT</td>
<td>1.23E-06</td>
<td>8.54E-07</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

Table 6.13 shows when the change in VKT is controlled for, the treatment effect reduces from 0.77 to 0.33 fewer crashes in the treated area, and it is still not statistically significant at the 10% level. Again, this is the same as the DiD estimate in section 6.3.3 in this special case.

### 6.4.3 Panel fixed-effects regression in R or Stata

If we estimate a fixed-effects model using the change in outcomes as the dependent variable, the dependent variable can take on negative values, and a linear regression is generally appropriate. This can be done in R or Stata by estimating the same models described above.

In Stata, panel regression models can also be estimated using the ‘xtreg’ command. In R, the ‘plm’ package offers several panel regression techniques. Both R and Stata can estimate panel data models when the panel is ‘unbalanced’, i.e., the units in the cross-section are not the same over time; however, such advanced techniques are beyond the scope of this report.

### 6.4.4 Summary

Steps to carrying out causal inference using panel fixed-effects regression are summarised below:

1. **Organise the data**: rows should contain a unique unit and its: change in the outcome level, treatment status and change in co-variates that are not fixed across units.
2. **Explore the data**: check the validity of the parallel paths assumption.
3. **Fit a fixed-effects regression**: regress the change in outcome vs treatment indicator variable and change in any other explanatory variables if appropriate.
4. **Identify treatment effect** as the estimated coefficient on the treatment indicator variable.

### 6.5 Propensity score matching

As discussed below, the Hamilton safer streets case study turned out to be unsuitable for PSM analysis. Applications of PSM to evaluating Auckland’s Northern Busway are described in chapter 7.

PSM is an alternative method of using information about observed differences between a treatment group and a control group. Using measured attributes, PSM refines the control and treatment groups so the two groups have relatively balanced characteristics and sufficient overlap. This is particularly important if the estimated treatment effect needs to be generalised beyond the specific treatment group for which it was estimated.

---

Imbalance refers to a fundamental difference between treated and control groups with respect to one or more pre-treatment variables. Lack of complete overlap occurs when the range of characteristics encompassed by one group are not supported by that of the other treatment type.
Unlike the other techniques described in this section, PSM does not provide an estimate of the treatment effect directly, however estimates of the treatment effect can be obtained through looking at average differences between the matched treatment and control groups, or by regression analysis using the matched data.

PSM can be an effective method of refining the treated and control groups so they resemble a randomly selected sample. Used with cross-sectional regression, only observed factors can be controlled for; however, PSM can be an even more powerful tool when combined with DiD, if suitable data is available.

6.5.1 Data and software requirements

PSM requires a relatively large number of units in the treatment and control groups. Data should include the outcome variable, a treatment indicator and a wide set of explanatory variables that describe unit attributes numerically or categorically. Essentially, any variable that may have influenced a unit’s likelihood of being treated should be recorded in the data.

PSM revolves around estimated propensity scores, ie estimates of the probability that a unit was selected for treatment. Since treatment is a binary variable, PSM requires logit or probit regression analysis, neither of which can be carried out in standard Excel. Although some Excel add-ins have logit or probit functions, more advanced statistical software can estimate these models and the matching stage in one function. These functions often provide different ways of matching the data and a summary of the change in balance, and overlap from the matching process. Therefore, we highly recommend PSM is carried out in software with these capabilities, rather than in Excel.

6.5.2 Analysis in R or Stata

6.5.2.1 Data layout

Data for estimating the propensity scores model should be in cross-sectional form (eg table 6.1). It is important all explanatory variables are measured before the intervention, or are variables that do not change over time and so cannot be affected by the intervention.

6.5.2.2 PSM analysis

PSM involves two phases: (1) estimate a model of propensity scores (ie probability of being treated for each unit), and (2) match treated and control units based on these propensity scores and use the matched units to calculate a treatment effect. The treatment effect calculation can be done by calculating and comparing weighted means of the treatment and control groups, or by estimating a cross-sectional regression or DiD model using the matched data and optionally using the propensity scores as weights in the regression. The following guidance assumes evaluators carry out PSM via a single function in statistical software, which typically requires the following to be specified:

Specify explanatory variables for the propensity scores model: the dependent variable is the treatment dummy (eg 1 for streets in safer speed areas, 0 otherwise) and explanatory variables are those which (potentially) influence treatment probability (eg VKT, and number of T intersections and cross intersections). If available, the pre-intervention value of the outcome of interest (eg crash count) can be included as an explanatory variable. This is particularly important if evaluators suspect units were targeted based on their pre-intervention outcome level, eg if a safety intervention was targeted at areas with high crash rates. The estimated propensity model is used to calculate a propensity score for each unit and is a value between 0 and 1 that can be thought of as the probability a unit was selected for treatment given its characteristics.
**Matching method:** several algorithms can be used to match treated and control units based on propensity scores. Methods available in most PSM packages include:

- Nearest neighbour: pairs treated and control units of the closest propensity score.
- Calliper or radius matching: sets the maximum distance (in propensity score) between two units for a valid match; units with no match inside this distance will be discarded.
- Sub-classification or interval matching: this technique divides the area of common support into intervals that contain units with attributes of similar levels.

The *control-to-treated ratio*: specifies the number of control units matched to each treatment unit. This is particularly useful when there can be multiple control units with the same or similar propensity score to that of a single treated unit. Setting the control-to-treated ratio is like deciding whether to match with or without replacement. For example, ‘with replacement’ places no restriction on the number of matches for a given treatment unit, whereas ‘without replacement’ limits a unit to just one match.

It is also possible to specify whether to discard units that fall outside the area of common support (ie outside the range of characteristics of units that are in both the treatment and control groups), and thus exclude them from the estimation of a treatment effect after matching.

Given that the Hamilton case study data contains only six treated sites and 25 control sites, it would normally be considered too small for PSM to be applied. However, for demonstrating this technique, we tried two PSM methods (PSM 1 and PSM 2). In both cases, propensity scores were estimated using pre-intervention VKT, pre-intervention crash count, the number of T intersections and the number of cross intersections as explanatory variables.

The first method, PSM 1, uses the nearest neighbour matching method, with a control-to-treated ratio of 1:1, and discards the treated and control units outside the area of common support. The second method, PSM 2, also uses the nearest neighbour method; however, the control/treated ratio is set to 1:2, and the only control units outside of the region of common support are discarded. Matching results are illustrated in figure 6.3.

Figure 6.3 shows large regions of the propensity score with no common support, ie no units in both the treatment and control groups. This implies, in general, control units in this dataset are poor counterfactuals for the treated group. Hence when units are matched under rather strict requirements, eg ‘without replacement’ (or treated/control ratio =1:1) in conjunction with discarding units outside the region of common support (PSM 1), only two units from each group are matched. When we relax the matching conditions by including ‘unsupported’ treated units and increasing the control/treated ratio, our matched sample increases (PSM 2). Gelman and Hill (2007) emphasise that matching methods should be selected to optimise both balance and overlap between the treatment and control groups. In this example, neither can be achieved without making the sample size too small to estimate a treatment effect.

If the matching algorithm produces control and treatment groups of sufficient size, the treatment effect can be calculated by one of:

- Comparing the means of the treatment and control groups. If there is more than one matched control unit per treatment unit, the outcomes for the control units need to be weighted, with weights based on propensity scores. This can be done automatically by statistical software.
- Estimating a cross-sectional or DiD regression model on the matched data, optionally using the propensity scores as weights. If weights are applied, each treated unit receives a weight of 1 and each control unit receives a weight of $p/(1 - p)$ where $p$ is its propensity score.
6.5.3 Summary

Steps to carrying out causal inference using PSM are summarised below:

1. **Organise the data for the propensity scores model**: into a cross-section of pre-intervention explanatory variables and a treatment indicator.

2. **Specify inputs and arguments of the matching function**: these include the explanatory variables for the propensity score model, the matching method, control-to-treated ratio, and whether to discard or keep unsupported units.

3. **Test different PSM specifications**: find model that optimises balance and overlap between the treatment and control groups.

4. **Identify treatment effect**: use regression (a cross-sectional model, or DiD if possible) to estimate the treatment effect, using the matched data. Propensity scores can be used as weights in a weighted least squares model. Alternatively, the means of the matched control and treatment groups can be compared using a t-test or other suitable hypothesis test.

6.6 Interrupted time series

ITSA is commonly applied in ex-post evaluation given its simple data requirements and interpretation. It is useful when no control group can be established, e.g., to estimate the effects of national policy changes. However, to estimate a treatment effect it requires extrapolating beyond the range of observed data, and it is difficult to be sure all other factors affecting the outcome of interest over time have been controlled for. This means ITSA is not usually a preferred technique unless there is no suitable data to use another technique involving an explicit control group.

6.6.1 Data and software requirements

ITSA requires a time series of outcomes for the treated unit(s), observed before and after an intervention. A time trend variable (e.g., first time period is 1, second is 2 etc) and a treatment dummy (encoded 0 for pre-intervention periods, 1 for post-intervention periods) are the only necessary explanatory variables; however, if possible it is good practice to include other explanatory variables that may have affected...
trends in the outcome of interest over time. A control group is not used in this method, and the
counterfactual is based on the pre-intervention outcomes for treated units. It may also be necessary to
handle seasonal fluctuations in the outcome of interest, eg by including seasonal dummy variables or by
seasonally adjusting the data prior to analysis.

ITSA can be carried in Excel or more advanced software. R and Stata offer methods that can correct for
‘autocorrelation’ problems that can arise when analysing time-series data, and may produce more reliable
estimates of treatment effects than simple regression models estimated in Excel. This is discussed further
in section 6.6.3.

6.6.2 Analysis in Excel

6.6.2.1 Data layout

In time series analysis, data is listed for each period in chronological order, where the columns contain the
corresponding outcome, time trend and indicator variables. Table 6.14 provides a subset of time series
data for the first eight quarters of the Hamilton case study. Outcomes are the sum of crash count for the
entire treatment group (six streets) in each quarter.

Although not always necessary, increasing the time frequency (eg from annual to quarterly) increases the
size of the dataset. When working with data higher than annual frequency (eg quarterly or monthly), it is
important to be aware of any seasonal effects that may be present in the outcome of interest. We have
included quarterly dummy variables to try to capture these effects. The quarterly crash count is also quite
volatile, so table 6.14 also includes a four-quarter moving average of the crash counts. Moving average
transformations of the outcome variable can be an effective way of smoothing the data to remove excess
volatility and/or seasonality and make it easier to estimate a treatment effect.

Table 6.14 Example of data layout for ITSA

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Crash count</th>
<th>Crash count MA(a)</th>
<th>Time trend</th>
<th>Treated</th>
<th>Treated *</th>
<th>Q2 dummy</th>
<th>Q3 dummy</th>
<th>Q4 dummy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>1</td>
<td>2</td>
<td>2.00</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>2</td>
<td>4</td>
<td>2.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>1</td>
<td>2</td>
<td>2.25</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>2</td>
<td>4</td>
<td>3.50</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>3</td>
<td>2</td>
<td>4.25</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>4</td>
<td>6</td>
<td>4.50</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) moving average

6.6.2.2 Exploratory analysis

Figure 6.4 presents total crash counts for the treated group of roads in our analysis of Hamilton city’s
safer speed areas.

The crash count time series varies significantly over time; this is expected given the randomness of vehicle
crashes. The four-quarter moving average time series removes much of the quarterly variation and
appears to show a decreasing time trend that starts before the intervention period. On average, the post-
intervention crash count seems lower than its pre-intervention level. However, this observation is limited
by the small number of post-intervention periods. From the graph, it is also not clear if the lower number

86
of crashes in the intervention period is due to the safer streets intervention, or simply a continuation of an earlier declining trend (or a combination of both effects).

Figure 6.4  Time series of quarterly crash counts treated group

6.6.2.3  ITSA in Excel

This section explores two ITSA methods. The first uses the quarterly crash count as the dependent variable and seasonal dummy variables to control for any seasonal effects. The second uses the four-quarter moving average of the crash count as the dependent variable, which eliminates seasonal effects (so seasonal dummy variables are not required) and smooths out other sources of excess volatility.

6.6.2.4  ITSA with seasonal dummies

ITSA with seasonal dummies requires the following explanatory variables: the treatment dummy, a time trend, an interaction between the time trend and treatment dummy, and seasonal dummies if the data has seasonal effects. The regression equation (applied to the Hamilton case study) is set out below and a regression summary is in table 6.15. This type of model allows the intervention to affect both the level and trend of the outcome variable. These effects are represented by the estimated coefficients on the treated and interaction variables, respectively.

\[
\text{Crash count}_t = \beta_0 + \beta_1 \text{Treated}_t + \beta_2 \text{Time trend}_t + \beta_3 \text{Treated} \times \text{Time trend}_t + \beta_4 Q2_t + \beta_5 Q3_t + \beta_6 Q4_t + e_t \quad \text{(Equation 6.6)}
\]

Table 6.15 shows this model estimates an average reduction of 3.09 crashes per quarter in the intervention period, but that the trend in crashes is increasing at a rate of 0.03 per quarter faster than the trend in the pre-intervention period. Neither of these coefficients is statistically significant at the 10% level or below, and an F-test reveals the two coefficients are also not jointly significant. This analysis therefore suggests there is no significant difference in the level and trend of crashes per quarter in the intervention period compared with the pre-intervention period. The lower rate of crashes in the post-intervention period therefore appears to be a continuation of a trend that started prior to the intervention.
Table 6.15  ITSA with controls for time and seasonality

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.01***</td>
<td>0.65</td>
</tr>
<tr>
<td>Treated</td>
<td>-3.09</td>
<td>11.84</td>
</tr>
<tr>
<td>Time trend</td>
<td>-3.43E-03</td>
<td>0.02</td>
</tr>
<tr>
<td>Treated x time trend</td>
<td>0.03</td>
<td>0.23</td>
</tr>
<tr>
<td>Q2 dummy</td>
<td>0.57</td>
<td>0.66</td>
</tr>
<tr>
<td>Q3 dummy</td>
<td>0.14</td>
<td>0.67</td>
</tr>
<tr>
<td>Q4 dummy</td>
<td>0.61</td>
<td>0.66</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

6.6.2.5 ITSA with smoothed dependent variable

Another method to control seasonality and other sources of volatility requires first smoothing of the dependent variable. There are several techniques for smoothing, the simplest of which is to calculate a moving average of the outcome of interest. The smoothed ITSA regression (equation 6.7) and summary (table 6.16) for the case study data are presented below:

\[ \text{Crash count} \; MA_t = \beta_0 + \beta_1 \cdot \text{Treated}_t + \beta_2 \cdot \text{Time trend}_t + \beta_3 \cdot \text{Treated} \times \text{time trend}_t + e_t \]  

(Equation 6.7)

Table 6.16  Smoothed ITSA

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.44***</td>
<td>0.24</td>
</tr>
<tr>
<td>Treated</td>
<td>-6.97</td>
<td>5.31</td>
</tr>
<tr>
<td>Time trend</td>
<td>-0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Treated x Time trend</td>
<td>0.11</td>
<td>0.10</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

Table 6.16 shows the smoothed model estimates a greater change in level (-6.97) and trend (+0.11) compared with the previous model. Again, neither effect is statistically significant at the 10% level. However, an F-test shows the two treatment effects are jointly significant at the 1% level, suggesting the safer speed areas reduced the average level of crashes in the intervention period, but slowed down the rate of reduction in crashes over time.

If the intervention affects both the level and trend of crashes, the treatment effect is not constant during the treatment period. One way to generate an overall estimate of the treatment effect is to calculate it for each quarter in the treatment period and sum those results. It is also important to remember that in this example the dependent variable of the ITSA analysis is the total crash count across all six streets where the intervention applied. To compare these estimates with the other methods where the analysis was conducted using street-level data, it is necessary to divide the average treatment effect by six to obtain the estimated average change in the number of crashes per street during the treatment period.

Applying this approach yields an estimated treatment effect of -2.13 for the model with seasonal dummy variables and -1.74 for the smoothed model.
6.6.3 Analysis in R or Stata

Statistical software like R or Stata includes additional functions for working with time series data. These functions allow for better estimation of treatment effects with count data and estimating time series models with auto-correlated errors. We applied these functions to the Hamilton case study data (see the Stata and R scripts for detailed code in appendix C).

6.6.3.1 Poisson ITSA with seasonal dummies

As with other regression techniques, an appropriate distribution for crash count data should be assumed. We were not able to estimate a negative binomial model in cases where the explanatory variables were purely dummy variables and linear time trends; however, a Poisson model could be estimated.

Re-estimation of the ‘ITSA with seasonal dummies’ model but under the Poisson distribution yields a combined average treatment effect over the treatment period of -1.42, but this effect is not statistically significant (p-value = 0.17). As the Poisson model is multiplicative it is not possible to separate this into level and trend effects.

6.6.3.2 ITSA with autoregressive errors

Regression models estimated with time-series data often suffer from the problem of serial correlation of the residuals, i.e., the residual in one period is correlated with the residual in the previous period. If not corrected, serial correlation will lead to incorrect standard errors and invalid significance test results.

There are various ways of dealing with serial correlation. A simple approach is to explicitly model the time-dependence of the residuals, i.e., to estimate a model where it is assumed the residual in one period partly depends on the residual in the previous period. Such models can be estimated in R or Stata using an ARIMA model instead of a simple linear regression.

We applied this approach to the linear model estimated above. The estimated effect of the treatment on the crash count level and time trend is -4.45 and -0.003, respectively. Although these effects are not significant in of themselves, are jointly significant at the 6% level (p-value = 0.056).

6.6.4 Summary

The steps for applying ITSA to ex-post evaluations are:

1. **Organise the data**: into time series format and create the time trend and necessary dummy variables. Create a smoothed transformation of the dependent variable (e.g., a moving average) if necessary, or handle any seasonality via seasonal adjustment or seasonal dummy variables.

2. **Graph the data**: plot a historical time series of the outcome variable and a smoothed transformation of the outcome variable (if necessary), to visualise the time trends.

3. **Estimate an ITSA model**: including a time trend, treatment period indicator, and interaction variable, as well as seasonal dummy variables if necessary. If data is available, other explanatory variables that may have affected the outcome of interest over time can also be included in the model. Test the significance of the post-intervention change in level and trend. Use an F-test to test joint significance of the treatment indicator and interaction effect.

4. **Identify treatment effect**: if the coefficients on the treatment indicator and interaction effect are individually and/or jointly statistically significant, calculate the treatment effect during the treatment period by combining these two effects as described above.

Evaluators must remember that causal inference using ITSA is limited by the assumption that effects of factors aside from the intervention are similar before and after the intervention. In other words, data prior
to the intervention period provides an adequate counterfactual for outcomes after that point in time, given the explanatory variables included in the model. This is a strong assumption if the difference in time is large, since it is possible other factors not included in the model have affected the outcome of interest over time.

6.7 Instrumental variables regression

IV methods provide a way of estimating the impact of an intervention when treated units are not randomly selected and selection was not carried out based on characteristics that can be observed. In regression analysis, non-random assignment based on unobserved characteristics can mean the treatment indicator variable suffers from selection bias or omitted variable bias (OVB or SB), ie there may be an unobserved (and thus uncontrolled) variable that impacts both the treatment and the outcome. In such a situation, the treatment indicator variable may be said to be ‘endogenous’, ie treatment and outcomes are (partially) jointly determined by some outside factor(s) that cannot be observed.

The idea of IV regression is to find one or more ‘instruments’ (variables) that help predict the treatment variable yet do not impact the outcome variable, and essentially to use these instruments in place of the treatment indicator, to estimate the treatment effect. If a suitable instrument can be found, an IV regression can be used to estimate a treatment effect that is free from bias. However, it is not always possible to find a suitable instrument and this may limit the applicability of IV methods.

6.7.1 Data and software requirements

IV regression to estimate a treatment effect requires the same data as simple regression analysis of a treatment effect, plus one or more IV for the treatment indicator.39 Evaluators must choose instruments carefully. A valid instrument must be:

- relevant – sufficiently correlated with the explanatory variable of interest (ie the treatment indicator variable)
- exogenous – have no impacts on the outcome of interest aside from any direct effects.

An understanding of how units were selected for treatment can help evaluators identify appropriate instruments. For example, the HCC targeted streets near to schools and/or with historically high crash rates.40 Given this information, the following instruments were chosen for the Hamilton case study:

- a dummy variable indicating whether a school is located on or adjacent to a given street
- the number of people aged under 15 within a given street’s census area unit (CAU) in 2006
- the number of crashes for a given street in 2000.

These are believed to be valid instruments because they reflect factors that affected selection into the treatment group but are not expected to be related to the other explanatory variables in the model.

IV regression is easily carried out in statistical programs such as R or Stata. Although IV regression is feasible in Excel, it is tedious and may be inaccurate when the treatment indicator is binary (eg units are either treated or controlled). This is because IV regression involves estimating a model of the treatment indicator, and models with binary dependent variables should be estimated with a suitable non-linear regression model.

---

39 Two or more instruments, ie over-identification of the treatment variable, are preferred as this allows instruments to be tested for exogeneity.

40 HCC website: www.hamilton.govt.nz/our-services/transport/movingaround/Pages/Safer-Speed-Areas.aspx
such as a logit or probit model. Such models cannot be estimated in Excel without a statistical add-in
package. Therefore, we recommend evaluators use statistical software to carry out IV regression.

6.7.2 IV regression in R or Stata

6.7.2.1 Data layout

Both cross-sectional and panel data can be used for IV regression. For illustration, we applied IV
regression to the cross-sectional analysis in the Hamilton case study (ie where the dependent variable is
the post-intervention crash count for treated and control streets). The cross-sectional regression equation is:

\[
\text{Crash count}_i = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \text{Pre.VKT}_i + \beta_3 \text{Pre.crash.count}_i + \beta_4 \text{Xintersections}_i + \beta_5 \text{Tintersections}_i + \epsilon_i
\]  

(Equation 6.8)

In this model, although we have controlled for the pre-intervention crash count, we may suspect the
treatment indicator is not fully exogenous, ie there are some other unobserved factors that affect the
probability of treatment and crash count. We applied IV regression to attempt to overcome that problem.

Table 6.17 shows a subset of the data: the outcome interest (crash count), explanatory variables and three
IV, to be used for the IV analysis.

<table>
<thead>
<tr>
<th>Unit name</th>
<th>Crash count</th>
<th>Treated</th>
<th>Pre-VKT</th>
<th>Pre-crash count</th>
<th>No. of X intersections</th>
<th>No. of T intersections</th>
<th>Under-15 population</th>
<th>Near school</th>
<th>2000 crashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrington Dr</td>
<td>1</td>
<td>1</td>
<td>501240</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>984</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heath-Mahana St</td>
<td>1</td>
<td>1</td>
<td>2498505</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>1200</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Willoughby St</td>
<td>1</td>
<td>1</td>
<td>2757341</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>294</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Dinsdale Rd</td>
<td>1</td>
<td>1</td>
<td>3533200</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>816</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Crosby Rd</td>
<td>1</td>
<td>0</td>
<td>2730200</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>816</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hillcrest Rd</td>
<td>7</td>
<td>0</td>
<td>4865450</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>585</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maeroa Rd</td>
<td>6</td>
<td>0</td>
<td>5323890</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>573</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

6.7.2.2 Exploratory analysis

IV regression is only required if the treatment variable is endogenous, ie suffers from selection bias due to
some important yet unobservable variable. If this is not the case then using IV regression may give a less
accurate estimate of the treatment effect than a simple regression, as the instruments are used in place of
the treatment indicator in the estimation of the treatment effect.

The Hausman test for endogeneity is commonly employed to test this assumption. The Hausman test has
two steps:

1. Regress the endogenous variable (treatment indicator) on all explanatory variables from the original
regression, and the instrument(s), eg:

\[
\text{Treated}_i = \beta_0 + \beta_1 \text{Pre.VKT}_i + \beta_2 \text{Pre.crash.count}_i + \beta_3 \text{Xintersections}_i + \beta_4 \text{Tintersections}_i + \beta_5 \text{U15.pop}_i + \beta_6 \text{Near.school} + \beta_7 \text{2000.crashes} + \epsilon_i
\]  

(Equation 6.8)

2. Regress the original outcome variable on all explanatory variables plus the residuals from the first
regression, eg:
Ex-post evaluation of transport interventions using causal inference methods

\[ \text{Crash count}_i = \beta_0 + \beta_1 \cdot \text{Treated}_i \]
\[ + \beta_2 \cdot \text{Pre.VKT}_i + \beta_3 \cdot \text{Pre.crash.count}_i + \beta_4 \cdot \text{Xintersections}_i \]
\[ + \beta_5 \cdot \text{Tintersections}_i + \beta_6 \cdot \text{Stage.1.Residuals}_i + e_i \]  
(Equation 6.9)

Essentially, the Hausman test evaluates whether error terms of these two models are correlated. If the residuals (errors) from the first step are statistically significant in the second regression model, we can conclude the treatment variable in the original model is endogenous.

If, as in this case, the treatment indicator is a binary variable, an appropriate model such as logit or probit should be used in the first stage.

Using the Hamilton safer streets dataset, we found the stage 1 residuals are statistically significant at the 10% level in the second stage regression (p-value = 0.058). Therefore, assuming our instruments are exogenous, the treatment variable is likely to be endogenous and the estimated treatment effect will suffer from selection bias.

6.7.2.3 IV regression analysis (two stage least squares)

Like the Hausman test, IV regression requires two stages. These regressions can be estimated simultaneously via a single IV function in statistical software, or calculated manually. We applied the two-stage least squares ('2SLS') method to the data described above.

2SLS - stage 1

Stage 1 of 2SLS is a repetition of the first regression in the Hausman test. Because the dependent variable is binary, we estimated this using a probit model. Unlike the Hausman test, it is the predicted values of this model, and not the residuals, that are used in the following stage. If the instruments used in stage 1 fulfil the criteria above, predicted values of the treatment should now be exogenous and can be used in place of the treatment indicator in the second stage regression.

In R, it is necessary to estimate the first stage using the glm command to estimate a probit model for the treatment indicator, and then calculate the fitted values from that model. In Stata, the treatreg command offers IV estimation of a treatment effect using a probit model for the first stage and linear regression for the second stage, with a single command (see the attached code for an example).

2SLS - stage 2

Stage 2 of 2SLS takes the original regression model (ie with the outcome of interest as the dependent variable) and replaces the original treatment indicator with its predicted values from stage 1. Re-estimation of this model should yield a treatment effect that is free of selection bias. We provide the second stage regression equation and results (table 6.18) for a linear regression.

Table 6.18 shows the predicted treatment variable is highly significant and the IV method estimates the safer speed intervention had a significant impact (8.16 fewer crashes) on the treated streets.

However, the treatment effect could still be biased if the instruments were irrelevant or endogenous. Relevant instruments should be independently or jointly significant in the stage 1 regression. Instruments can only be tested for exogeneity if they are over identified, ie there is more than one instrument for the treatment variable. This test requires estimation of different 2SLS estimates where only one instrument is used in the first stage regression. If 2SLS estimates of the treatment effect are sufficiently close according to an F-test, the instruments are deemed to be exogenous.
6.7.3 Summary

The steps for applying IV regression via 2SLS to ex-post evaluations are:

1 Organise the data: define the regression model with the possibly endogenous treatment variable, and collect data on instruments.

2 Hausman test: carry out the test of whether the treatment indicator is endogenous.

3 IV regression/2SLS: carry out both stages of IV regression simultaneously via an IV function or manually. Ideally, a logit or probit model should be used for the binary treatment indicator in the first stage model.

4 Test instruments: analyse stage 1 coefficients to see if instrument(s) are relevant (use F-test if multiple instruments are employed). If possible, test that instruments are exogenous.

5 Identify treatment effect: the estimate is valid if instruments are relevant and endogenous. The estimated treatment effect is the coefficient on the predicted values of the treatment indicator in the second stage model.

6.8 Comparison of results

Table 6.19 summarises and compares the estimated treatment effects of the Hamilton safer speed areas intervention under the various methods described above. The estimated treatment effect in each case is the reduction in mean crash count for treated streets during the treatment period.

These results imply the following:

- A simple comparison of mean post-intervention crash counts between the treatment and control groups suggests the intervention prevented about six crashes on average per street in the treatment group, during the treatment period. A simple t-test also indicates this effect is highly statistically significant.

- Controlling for the different characteristics of the treatment and control groups via a simple cross-sectional regression reduces the estimated treatment effect to a reduction of 3.8 crashes per street during the treatment period, and this effect is only weakly statistically significant. If an appropriate distribution for crashes is assumed (i.e., a negative binomial model is used), the estimated treatment effect increases to -4.6 crashes, and the statistical significance improves.

- The more sophisticated fixed-effects and difference-in-differences approaches produce the same results in this case and can control for some types of unobserved factors that may have affected crash counts. These more robust methods suggest that the safer speed areas intervention did not have any
statistically significant impact on crash counts for the treated streets in comparison with the control group. From the data it is clear both the control and treatment groups experienced a reduction in crashes after the intervention. The analysis suggests the reduction in crashes in the treatment group is not significantly different from that observed in the control group, i.e., there is no evidence of a reduction in crashes above the pre-existing trend.

- ITS and IV estimates were provided for illustration; however, given the data that is available (i.e., pre- and post-intervention observations for both a treatment and control group), difference-in-differences or fixed-effects models are better methods for conducting this evaluation.

- In the ITSA models, if serial correlation is corrected for then the intervention is estimated to have caused a reduction in about 2.3 crashes on average per street during the treatment period and this effect is weakly statistically significant (p-value 0.06).

Overall, this example highlights the importance of controlling for differences in observed characteristics that may have affected the outcome of interest (crash counts in this case), and, where possible, the influence of unobserved factors via a difference-in-differences or fixed effects evaluation. When such factors are sufficiently accounted for, an apparently strong treatment effect can disappear. However, it is also important to note that methods applied to panel data (e.g., fixed-effects and DiD models) can underestimate a treatment effect when explanatory variables in the model are highly persistent over time (i.e., do not vary much over time), due to the correlation between the fixed-effects components of these models and persistent explanatory variables.

As noted in the introduction to this section, additional analysis is necessary to reach definitive conclusions about the Hamilton safer speed areas intervention. The results presented here are only intended for the purpose of illustrating the techniques that we have described.

Table 6.19 Comparison of estimated treatment effects for the Hamilton safer streets intervention

<table>
<thead>
<tr>
<th>Technique</th>
<th>Variation</th>
<th>Estimated treatment effect</th>
<th>Treatment effect p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive comparison (t test)</td>
<td>Normal distribution</td>
<td>-5.96</td>
<td>0.00</td>
</tr>
<tr>
<td>Cross-sectional regression, with control variables</td>
<td>Normal distribution</td>
<td>-3.80</td>
<td>0.09</td>
</tr>
<tr>
<td>Fixed-effects model, controlling for VKT</td>
<td>Normal distribution</td>
<td>-4.55</td>
<td>0.02</td>
</tr>
<tr>
<td>Difference-in-differences (DiD)</td>
<td>Negative binomial distribution</td>
<td>-0.33</td>
<td>0.89</td>
</tr>
<tr>
<td>Interrupted time series</td>
<td>No other controls</td>
<td>-0.77</td>
<td>0.87</td>
</tr>
<tr>
<td>Interrupted time series</td>
<td>Controlling for VKT</td>
<td>-0.32</td>
<td>0.91</td>
</tr>
<tr>
<td>Instrumental variables</td>
<td>Normal</td>
<td>-2.13</td>
<td>0.24</td>
</tr>
<tr>
<td>Interrupted time series (moving average)</td>
<td>Normal + auto regression</td>
<td>-2.28</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-1.42</td>
<td>0.17</td>
</tr>
<tr>
<td>Instrumental variables</td>
<td>Normal</td>
<td>-1.74</td>
<td>0.00</td>
</tr>
<tr>
<td>Instrumental variables</td>
<td>Normal</td>
<td>-8.16</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Case study: Auckland’s Northern Busway

7.1 Introduction

In this section, we apply some of the empirical techniques discussed in this report to evaluate aspects of the Northern Busway in Auckland. This case study is complementary to the Hamilton safer speed areas study in chapter 6 and provides further illustration of the use of causal inference techniques, including the combination of PSM with other techniques.

The key objectives of the busway are to (NZ Transport Agency 2012):

- increase accessibility to public transport
- reduce travel times of bus users along SH1
- increase person-carrying capacity of the Harbour Bridge
- minimise adverse environmental effects of private motor vehicle use
- enhance activity in city centres by improving accessibility and capacity.

We were constrained by data availability and were not able to evaluate all objectives or potential effects of the busway. In terms of the objectives above, we have analysed:

- impacts on accessibility to public transport (measured as changes in use of bus services)
- business and employment activity in areas close to busway stations on the North Shore
- vehicle volumes on the Harbour Bridge (but we were not able to measure people volumes directly).

We also analysed other outcomes that are not related to the busway’s objectives but may still be of interest, including:

- impacts on customer satisfaction of public transport users
- travel time reliability and congestion experienced by vehicles on the Northern Motorway (we were not able to measure travel times of bus users).

We have not attempted to provide an overall ex-post assessment of the benefits and costs of the busway. Instead, we have analysed outcomes for which we could obtain sufficient data to apply robust causal inference methods. This case study is intended primarily to illustrate the application of these techniques and, on its own, is not sufficient to evaluate the overall impacts of the busway. Due to data constraints, a full analysis of the impacts of the busway may need to rely on some qualitative analysis as well as quantitative analysis. Such analysis is outside the scope of this report.

7.2 The Northern Busway

The busway connects parts of Auckland’s North Shore with the central business district. Initially, five stations were built on the North Shore (Albany, Constellation, Sunnynook, Smales Farm and Akoranga), with services stopping at all stations on the North Shore, and stopping at Victoria Park in the CBD before terminating at Britomart. The busway infrastructure between Albany and the CBD was completed and officially opened in February 2008. Services from the Albany and Constellation stations started in December 2005 using the general traffic lanes on the motorway, but for the purposes of our analysis we assume the busway was in operation from the beginning of 2008 onwards.
Between Constellation and Akoranga stations, the busway consists of two-way segregated, dedicated lanes alongside the Northern Motorway. This enables buses using the busway to avoid much of the congestion that occurs on the Northern Motorway. From Akoranga station to the Harbour Bridge there is also a dedicated bus lane in the south-bound direction only. Elsewhere, busway services travel in general traffic motorway lanes (eg between Constellation and Albany stations in both directions) or on local roads; busway services also use some bus-only lanes within the CBD.

Albany and Constellation stations provide extensive park-and-ride facilities and all stations are served by local feeder buses. Bicycle parking facilities are also provided at every station, but it is not permitted to take bicycles on buses.

An additional park-and-ride station at Hibiscus Coast was opened in mid-2015, with buses travelling on the normal motorway lanes between Hibiscus Coast and Albany stations. As this is relatively recent and much of the data we have used relates to earlier periods, we have limited our analysis to the parts of the busway and busway services from Albany southward. Figure 7.1 provides an overview of the busway stations and route we have analysed.

The Northern Express (NEX) bus service provides frequent service between all busway stations and the CBD using a dedicated fleet of modern, air-conditioned buses (many of which are now double-deckers). As the NEX service was introduced as the same time as the busway, it is not possible to separate the effects of the busway from the effects of the NEX service. Although we refer to our analysis as estimating the effects of the busway, it really estimates the effects of the busway plus the NEX service.

Patronage on the Northern Busway has grown rapidly (figure 7.2). Auckland Transport (AT) only reported busway patronage relating to the NEX service prior to July 2015, even though the busway is also used by some other non-NEX services between the North Shore and the CBD, and by some local North Shore services for portions of their routes. From July 2015, this was changed to include all busway services, generating a one-time increase in the busway patronage figure. It meant patronage growth rates before
and after this time would not be not directly comparable. Our estimates of the impact of the busway also included the effects of these other services using the busway infrastructure.

**Figure 7.2  Patronage on the Northern Busway.**

![Graph showing busway patronage over time]

Source: Auckland Transport

### 7.3 Effects of the busway

The following analysis looks at the Northern Busway’s impact on a range of outcomes, including the:

- AADT over the Harbour Bridge
- travel time variability and congestion experienced by vehicles on the Northern Motorway
- number of people working and businesses established close to the busway on the North Shore
- proportion of commuters living near the busway travelling to any work location in Auckland by bus
- proportion of commuters living near the busway travelling to work in the CBD by bus
- proportion of AT HOP cardusers living near to the busway travelling to the CBD by bus in 2016
- reported levels of customer satisfaction of passengers using busway services
- total bus patronage in Auckland.

A summary of our analysis and findings for each of these outcomes are presented below.

---

41 A reusable prepay smart card for travel on trains, ferries and buses around Auckland
7.3.1 Harbour Bridge AADT

The busway was expected to reduce traffic levels on the parallel Northern Motorway due to mode transfer to public transport (NZ Transport Agency 2012). We used AADT data from the Transport Agency to evaluate this outcome.

Data from three telemetry sites on the Auckland Harbour Bridge was combined to form an annual time series over years 2001 to 2015 of average daily traffic volumes on the bridge. Figure 7.3 shows a large reduction in the Harbour Bridge AADT from 2007 to 2008, and subdued growth thereafter (until 2015).

Figure 7.3 Auckland Harbour Bridge annual average daily traffic 2001–2015

Given that the Harbour Bridge is a single, unique site, it is not possible to estimate the effects of the busway on AADT by comparing a treatment and control group. Instead, we apply an ITSA to evaluate the joint statistical significance of these two effects. As noted elsewhere in our report, ITSA is less robust than other causal inference methods, and the results of this analysis should be interpreted with caution.

In this analysis, the treatment occurs from 2008 onwards, and observations prior to 2008 are used to help establish the counterfactual. Explanatory variables include a treatment dummy, a time trend variable, an interaction between the time trend and the treatment dummy, and real national GDP. We have also included dummy variables to capture the construction period for the Victoria Park Tunnel (most of 2010 and 2011). This was a major construction project just south of the Harbour Bridge that may have disrupted traffic volumes on the bridge during its construction.

Given the relatively small number of observations available, it was not possible to test many explanatory variables in this model. The model’s function form is:

---

42 Auckland Harbour Bridge: centre span, left clip-on and right clip-on
43 Real GDP retrieved from Statistics NZ’s Infoshare: Economic Indicators, National Accounts, GDP(E), Chain volume, Actual, Total (Qrtly- Mar/ Jun/ Sep/ Dec)
\[
AADT_t = \beta_0 + \beta_1 \cdot \text{Treated}_t + \beta_2 \cdot \text{Time trend}_t + \beta_3 \cdot \text{Real GDP}_t + \beta_4 \cdot \text{VPT construction} \\
+ \beta_5 \cdot \text{Treated} \ast \text{time}_t + \epsilon_t 
\]  
(Equation 7.1)

Regression results from table 7.3 show the busway is estimated to reduce AADT by 8,378 vehicles per day in 2008 but to increase traffic volumes by 460 per day with each post-intervention year. Although these two effects are not statistically significant by themselves \((p > 0.05)\), an F-test infers their joint impact is significant at the 10% level \((p\text{-value} = 0.085)\).

We applied the Breusch–Godfrey test\(^{45}\) to assess the presence of auto-correlation (also known as serial correlation) in the time series model which, if present, can lead to incorrect estimates. The test shows no evidence of non-zero correlations among lagged residuals \((p\text{-value} = 0.094)\).

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>10,777</td>
<td>81,300</td>
</tr>
<tr>
<td>Treated</td>
<td>-12,056</td>
<td>7,013</td>
</tr>
<tr>
<td>Time trend</td>
<td>-3,317</td>
<td>3,532</td>
</tr>
<tr>
<td>Real GDP</td>
<td>0.95</td>
<td>0.55</td>
</tr>
<tr>
<td>Treated x time trend</td>
<td>460</td>
<td>1,328</td>
</tr>
<tr>
<td>VPT construction dummy</td>
<td>2,361</td>
<td>2,677</td>
</tr>
</tbody>
</table>

* \(p\text{-value} < 0.10\), ** \(p\text{-value} < 0.05\), *** \(p\text{-value} < 0.01\)

Given that the dependent variable (AADT) is a count, it may be more appropriate to estimate the treatment effect with a Poisson model, using the same variables as described above. As discussed in section 6.2.3, the log-linear relationship of the Poisson model means its regression coefficients do not have a simple additive interpretation. Therefore, we calculate the combined treatment effect by:

1. Retrieving the fitted values of AADT from the Poisson ITSA model from 2008 onwards, ie estimated AADT with the busway
2. Using the model coefficients to predict the counterfactual AADT from 2008 inwards, ie set the treatment dummy to 0 in that period and use the model to estimate AADT in each year
3. Taking the average difference between (1) and (2) over the treatment period to estimate an overall average treatment effect.

Figure 7.4 illustrates results from the Poisson model predicted AADT with the busway (values from calculation 1), predicted AADT without the busway (values from calculation 2), and actual Harbour Bridge AADT. The difference between the two predicted lines is the estimated impact of the busway on AADT in each year.

\(^{44}\) The initial reduction of 8,378 vehicles per day is calculated as the coefficient on the treated dummy \((-12,056)\) plus the value of the treated * time trend interaction effect in 2008 \((460 \ast 8)\).

\(^{45}\) This is one of a handful of tests used to test for auto-correlation in a time series model. Essentially, all of these models test how well a given data point can be predicted by previous data points.
The difference between these two predicted values suggests, on average, the busway reduced Harbour Bridge AADT by 6,500 vehicles or by about 3.9% of the counterfactual AADT in each year from 2008 to 2015. The treatment effect is statistically significant at the 1% level.

Figure 7.4 Harbour Bridge AADT: actual and Poisson ITSA predicted values

7.3.2 Travel time variability and congestion

The original economic evaluation estimated the single largest source of benefits from the busway would be derived from traffic decongestion along the parallel SH1 (NZ Transport Agency 2012). Such benefits were estimated to be around 40% of the project’s total economic benefit.

We investigated the busway’s impact on travel time variability (reliability) and congestion through data provided by AT’s traffic system performance monitoring reports. These reports contain travel times for the majority of Auckland’s motorways and other key routes. The road network is disaggregated into short segments (approximately 0.5–2km long) and travel times are recorded three times a day (morning, inter and evening peak), over a five- day period (Monday to Friday), twice a year (November and March).

We define the treatment group as 10 segments on the motorway alongside the busway (from the Sunset Rd overbridge to the Fanshawe St off ramp). The control group is made up of 25 segments from other parts of the Auckland city state highway network. All North Shore segments were excluded from the control group in case they were indirectly affected by the busway, ie to satisfy the stable unit treatment value assumption (SUTVA).

For consistency, we only use measurements of the morning peak traffic travelling in the city-bound direction. Pre-intervention data is taken from the March 2005 survey, while the March 2009 and 2012 surveys are used as two separate post-intervention periods. We use two post-intervention periods to allow for the fact that the busway’s impact on congestion and reliability may differ in the short and long term. In
particular, any immediate reduction in congestion may eventually be offset by induced traffic on the Northern Motorway. As data is in a panel form, we used DiD for the analysis.46

The two outcomes analysed are travel time variability (TTV)47 and congestion index (CGI).48 Unfortunately, no appropriate control variables were available for this analysis, thus the DiD regression (equation 7.2) for these models is:

\[ Y_{it} = \beta_0 + \beta_1.Treated_{it} + \beta_2.Time\ trend_{it} + \beta_3.Treated*Time_{it} + e_t \]  

(Equation 7.2)

where \( Y_{it} \) is either TTV or congestion for road segment \( i \) in time period \( t \).

Results from the TTV models are presented in table 7.2. The 2009 vs 2005 model does not show a statistically significant difference in TTV at the 5% level (p-value > 0.05); however, the model using 2012 as the post-intervention year shows a weakly significant (p-value < 0.10) increase (worsening) in TTV. The estimated impact of a 0.44 unit increase in TTV is more than 100% of the average TTV for segments in the treatment group in 2005 (0.38). However, the large standard error indicates this estimate is relatively imprecise.

Table 7.2 Regression summary – travel time reliability

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Outcome = TTV 2009 vs 2005</th>
<th>Outcome = TTV 2012 vs 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.66***</td>
<td>0.66***</td>
</tr>
<tr>
<td>Treated</td>
<td>-0.28</td>
<td>-0.28*</td>
</tr>
<tr>
<td>Period</td>
<td>0.46**</td>
<td>-0.31**</td>
</tr>
<tr>
<td>Treated x period</td>
<td>-0.45</td>
<td>0.44*</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

Table 7.3 summarises results from the DiD models where the CGI is the dependent variable. In this case, although the treatment effects are negative for both models, neither shows a statistically significant difference in congestion (p-value > 0.10).

Table 7.3 Regression summary – congestion index

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Outcome = CGI 05- 09</th>
<th>Outcome = CGI 05- 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.92***</td>
<td>0.92***</td>
</tr>
<tr>
<td>Treated</td>
<td>1.41***</td>
<td>1.41**</td>
</tr>
<tr>
<td>Period</td>
<td>0.05</td>
<td>-0.33</td>
</tr>
<tr>
<td>Treated x Period</td>
<td>-0.05</td>
<td>-0.30</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

46 The earliest data pertains to 2005, meaning a time series analysis would not be appropriate given the lack of pre-intervention time periods.

47 Calculated as the maximum travel time of the daily averages, minus the minimum travel time of the daily averages, divided by the average of the five daily travel time averages, in each year.

48 Minutes of delay per kilometre, where the expected travel time is derived from a segment’s speed limit.
In summary, there is little evidence the busway reduced TTV or congestion along the Northern Motorway parallel to the busway. In fact, there is weak evidence of a worsening in TTV when comparing TTV in 2012 with 2005.

It is possible the absence of evidence for reduced TTV or congestion caused by the busway could be due to induced demand on the Northern Motorway. In other words, after the combined capacity of the motorway and busway increased, there may have been an increased incentive to travel by private vehicle on the motorway. Such induced traffic may have offset any improvements in TTV or congestion caused by the busway, making such improvements difficult to observe.

These findings should be taken with caution as we could not control for time-varying factors, such as any additional motorway lanes or other upgrades, or changes in the population of the connecting areas.

7.3.3 North Shore businesses and employment

We have analysed the busway’s impact on the number of employees and business ‘geographic units’ in locations on the North Shore that are relatively close to the busway.49 The busway’s impact on the same outcomes in the CBD was not evaluated due to several anticipated difficulties:

- lack of sample size – there are only three CBD area units (insufficient for statistical analysis)
- unique characteristics of CBD area units – it would be impossible to find enough appropriate control area units. Thus, the evaluation would be limited to time series regression, which in our view could not adequately isolate the busway effect from all other factors affecting CBD employment.

Data from Statistics NZ’s business demographics and 2006 and 2013 census datasets was used for the analysis. Information from each of these datasets can be disaggregated at the CAU level. A CAU is roughly equivalent to a suburb and thus the collection of Auckland CAUs forms a cross section that can be used for analysis. Data from CAUs over time provides a panel dataset that can be used for fixed-effects or DiD models.

GIS software was used to calculate the road distance from the centre point of each CAU distance to the CBD (Britomart). We also calculated whether any part of each CAU was within a 1 or 2km radius of a train station or busway station.

Our base dataset included 406 CAUs in the Auckland region.50 The treated group consists of 15 North Shore CAUs in close proximity to busway stations. The remaining 391 CAUs are all potential control units. Robust causal inference requires no interference between control and treated units (the stable unit treatment value assumption, see section 2.5.2). To satisfy this requirement, we filtered the potential control units to exclude 43 CAUs that had some area within 2km of a busway station.

We used PSM to select the control group based on CAU characteristics measured in 2006 (ie prior to the busway’s completion). This method helps identify control CAUs with similar business-related characteristics to the treatment group. Characteristics used to match units include road distance from the centre point of each CAU to the CBD, and a number workplace characteristics from the 2006 census (ie prior to the busway’s completion), such as the proportion of employee types working in the area unit (eg labourers, professionals, technicians), and employee travel to work modes. Table B.1 in appendix B provides the full list of attributes used in the PSM algorithm.

---

49 A ‘geographic unit’ is a single physical location that a business operates at. Large businesses may have more than one geographic unit.

50 Seven gulf harbour CAUs in the Auckland region were excluded from the base data.
The nearest neighbour method was used in conjunction with a 1:2 treated:control ratio. Figure 7.5 shows that treated units’ propensity scores are evenly distributed between 0 and 1, whereas the majority of matched control units lie between 0 and 0.4. Ideally, only units with common support, ie matched units with propensity scores between 0 - 0.4, would be analysed. However, there is an insufficient number of treated units with matched control units in this propensity score range, therefore we use all matched CAUs in the following analysis. Figure 7.6 provides a geographic representation of the matched treatment and control CAUs resulting from this analysis.

Figure 7.5 Distribution of propensity scores

Data on outcomes and characteristics for each CAU is available for 2006 and 2013, so we can apply DiD estimation to the matched data, ie a PSM- DiD model. A key assumption of the DiD model is that the treated unit’s counterfactual trend (ie trend without busway) and control unit’s actual trend are parallel (ie
the parallel paths assumption). Although this is impossible to test in reality, we can partially verify this
assumption using historical data. We construct a pre-treatment time series of the average number of
employees and geographic units per CAU for each group. The employee time series (figure 7.7) shows the
two treatment types have a similar, but not parallel, trend. However, the geographic unit time series
(figure 7.8) suggests the parallel trends assumption is certainly plausible for this outcome.

Figure 7.7  Pre-treatment time series of average employee count per CAU, 2000–2008

![Pre-treatment time series of average employee count per CAU, 2000–2008](image)

Figure 7.8  Pre-treatment time series of average business count per CAU, 2000–2008

![Pre-treatment time series of average business count per CAU, 2000–2008](image)

Our first PSM-DiD model includes all potentially relevant explanatory variables in the dataset, and the full
DiD model has the following function form:
\[ Y_{it} = \beta_0 + \beta_1 \text{labourers proportion}_{it} + \beta_2 \text{machinery operators and drivers proportion}_{it} + \beta_3 \text{managers proportion}_{it} + \beta_4 \text{sales workers proportion}_{it} + \beta_5 \text{technicians and trade workers proportion}_{it} + \beta_6 \text{bicycle to work}_{it} + \beta_7 \text{drove to work}_{it} + \beta_8 \text{motorcycle to work}_{it} + \beta_9 \text{other mode to work}_{it} + \beta_{10} \text{car passenger to work}_{it} + \beta_{11} \text{bus to work}_{it} + \beta_{12} \text{train to work}_{it} + \beta_{13} \text{treated}_{it} + \beta_{14} \text{period}_{it} + \beta_{15} \text{treated \times period}_{it} + \epsilon_t \]  

(Equation 7.3)

where \( Y_{it} \) is the employment or geographic unit count for CAU \( i \) in time period \( t \).

Table 7.4 presents the regression summaries for both business-related outcomes. The models estimate that, on average, the busway increased the number of employees and business geographic units in each CAU by averages of 1,082 and 126 respectively. However, neither of these effects are significant at the 10% level or better.

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Outcome = employee count</th>
<th>Outcome = geographic unit count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard error</td>
</tr>
<tr>
<td>Constant</td>
<td>-13,341</td>
<td>10,452</td>
</tr>
<tr>
<td>Workplace labourers over age 15 proportion</td>
<td>6,280</td>
<td>18,481</td>
</tr>
<tr>
<td>Workplace machinery operators and drivers over age 15 proportion</td>
<td>20,702</td>
<td>20,470</td>
</tr>
<tr>
<td>Workplace managers over age 15 proportion</td>
<td>20,516</td>
<td>16,418</td>
</tr>
<tr>
<td>Workplace professionals over age 15 proportion</td>
<td>-4,643</td>
<td>14,481</td>
</tr>
<tr>
<td>Workplace sales workers over age 15 proportion</td>
<td>-13,152</td>
<td>15,539</td>
</tr>
<tr>
<td>Workplace technicians and trades workers over age 15 proportion</td>
<td>2,892</td>
<td>20,548</td>
</tr>
<tr>
<td>Workplace travel to work bicycle proportion</td>
<td>-36,595</td>
<td>96,491</td>
</tr>
<tr>
<td>Workplace travel to work drove a private vehicle proportion</td>
<td>23,244***</td>
<td>5,818</td>
</tr>
<tr>
<td>Workplace travel to work motor cycle or power cycle proportion</td>
<td>40,399</td>
<td>105,740</td>
</tr>
<tr>
<td>Workplace travel to work other proportion</td>
<td>35,883</td>
<td>61,676</td>
</tr>
<tr>
<td>Workplace travel to work passenger in a private vehicle proportion</td>
<td>8,358</td>
<td>61,445</td>
</tr>
<tr>
<td>Workplace travel to work public bus proportion</td>
<td>11,139</td>
<td>36,364</td>
</tr>
<tr>
<td>Workplace travel to work train proportion</td>
<td>343,452**</td>
<td>164,602</td>
</tr>
<tr>
<td>Treated</td>
<td>376</td>
<td>1,170</td>
</tr>
<tr>
<td>Period</td>
<td>-909</td>
<td>1,119</td>
</tr>
<tr>
<td>Treated \times period</td>
<td>1,082</td>
<td>1,645</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

Regression results from a reduced model (where insignificant control variables from the full model are omitted) are presented in Table 7.5. Although the impacts estimated in the reduced model are greater, neither are statistically significant at the 10% level or better.
We conclude there is no evidence of an impact of the busway on the number of employees or business units on the North Shore. Instead, growth in businesses and employment in these areas appears to have been caused by other factors aside from the busway which also caused the number of businesses and employment to increase in control CAUs.

### 7.3.4 Commuters travelling by bus to work

This analysis looks at the Northern Busway’s impact on the proportion of commuters living near busway stations travelling to any work location by bus. Information from the 2006 and 2013 censuses was used to form a panel of Auckland CAUs that includes the outcome of interest (proportion of commuters travelling to work by bus) and a wide range of CAU characteristics, all measured at CAU level.

The treated units include 25 North Shore CAUs located near a busway station. This expands on the 15 treated units used in the analysis of North Shore business and employment outcomes to include additional CAUs along the North Shore’s eastern coastline. We used a larger treatment area when analysing commuter behaviour because of the access to the busway stations provided by local bus services and park-and-ride facilities at some stations.

Control CAUs located within 2km of a busway station were excluded to help satisfy the SUTVA assumption. PSM was used to select an appropriate control group from Auckland CAUs and improve the balance and overlap of the treatment and control groups. Table B.1 in appendix B lists the CAU-level attributes used in the PSM algorithm. In this case, the variables used in the PSM analysis focused on household and commuter-related characteristics of people living in each CAU. We used the nearest neighbour matching technique, and matched two controls for every treated unit. As propensity scores of both treatment groups are distributed sufficiently, we discarded treated units outside the region of common support without greatly compromising the number of matched units, as shown in figure 7.9. A geographic representation of the matched CAUs is presented in figure 7.10.
After matching, we applied a DiD model to estimate the busway’s impact on the proportion of residents of each CAU travelling to work by bus. The corresponding DiD regression equation is described in equation 7.4.
Proportion of bus commuters\(_{it}\) =  
\[ \beta_0 + \beta_1 \cdot \text{couple with children proportion}_{it} + \beta_2 \cdot \text{full time employed proportion}_{it} + \beta_3 \cdot \text{bicycle to work proportion}_{it} + \beta_4 \cdot \text{drove to work proportion}_{it} + \beta_5 \cdot \text{train to work proportion}_{it} + \beta_6 \cdot \text{walked to work proportion}_{it} + \beta_7 \cdot \text{median hh member count}_{it} + \beta_8 \cdot \text{median age}_{it} + \beta_9 \cdot \text{median hh income}_{it} + \beta_{10} \cdot \text{no vehicle proportion}_{it} + \beta_{11} \cdot \text{2 vehicles proportion}_{it} + \beta_{12} \cdot \text{3 or more vehicles proportion}_{it} + \beta_{13} \cdot \text{occupied private dwelling proportion}_{it} + \beta_{14} \cdot \text{rented dwelling proportion}_{it} + \beta_{15} \cdot \text{treated}_{it} + \beta_{16} \cdot \text{period}_{it} + \beta_{17} \cdot \text{treated x period}_{it} + e_{it} \]  
(Equation 7.4)

As above, a key assumption of the DiD model is that the treated unit’s counterfactual trend (ie trend without busway) and control unit’s actual trend are parallel. Historical time series data for the outcome of interest is not available, meaning the parallel paths assumption cannot be tested. Regression results in table 7.6 show that several explanatory variables are strongly significant, including the treatment effect coefficient (the treated * period interaction variable). This model estimates that the busway increased the proportion of bus commuters in each CAU by 1.6 percentage points on average (p-value = 0.02). For perspective, the average proportion of commuters travelling by bus in 2013 in the control group is 8.7% thus a 1.6 percentage point increase from the busway is rather substantial.

We also re-estimated this model using only statistically significant control variables. Results from the reduced model (table 7.7) show a small increase in the treatment effect, from 1.6% to 1.9% An F-test shows that the insignificant control variables of the full model do not have a joint statistical significance at the 5% level, in other words, taken together they do not add any meaningful variation to the full model. Therefore, our preferred model specification is the reduced model, which estimates the busway increased the proportion of bus commuters in each CAU by 1.9 percentage points on average (p-value < 0.01).

Table 7.6  PSM- DiD regression results – full model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient(^{(a)})</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.276**</td>
<td>0.108</td>
</tr>
<tr>
<td>Treated</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Period</td>
<td>0.014**</td>
<td>0.006</td>
</tr>
<tr>
<td>Couple with children proportion</td>
<td>0.127**</td>
<td>0.062</td>
</tr>
<tr>
<td>Full time employed over age 15 proportion</td>
<td>0.062</td>
<td>0.060</td>
</tr>
<tr>
<td>Home travel to work bicycle proportion</td>
<td>-0.463</td>
<td>0.521</td>
</tr>
<tr>
<td>Home travel to work drove a private vehicle proportion</td>
<td>-0.340***</td>
<td>0.062</td>
</tr>
<tr>
<td>Home travel to work train proportion</td>
<td>-0.354</td>
<td>0.748</td>
</tr>
<tr>
<td>Home travel to work walked or jogged proportion</td>
<td>-0.249**</td>
<td>0.116</td>
</tr>
<tr>
<td>Mean number of household members</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td>Median age</td>
<td>-0.003**</td>
<td>0.001</td>
</tr>
<tr>
<td>Median household income</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>No motor vehicle proportion</td>
<td>0.408**</td>
<td>0.128</td>
</tr>
<tr>
<td>Occupied private dwellings separate house proportion</td>
<td>-0.012</td>
<td>0.026</td>
</tr>
<tr>
<td>Rented dwellings proportion</td>
<td>-0.026</td>
<td>0.048</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>-0.241***</td>
<td>0.062</td>
</tr>
<tr>
<td>Two motor vehicles proportion</td>
<td>-0.099</td>
<td>0.067</td>
</tr>
<tr>
<td>Treated x period</td>
<td>0.016**</td>
<td>0.007</td>
</tr>
</tbody>
</table>

\(^{(a)}\) Proportion of commuters travelling by bus

\(\star\) p-value < 0.10, \(\star\star\) p-value < 0.05, \(\star\star\star\) p-value < 0.01
Case study: Auckland’s Northern Busway

### Table 7.7  PSM- DID regression results – reduced model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.043</td>
<td>0.046</td>
</tr>
<tr>
<td>Treated</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Period</td>
<td>0.012***</td>
<td>0.004</td>
</tr>
<tr>
<td>Couple with children proportion</td>
<td>0.116**</td>
<td>0.048</td>
</tr>
<tr>
<td>Full time employed over age 15 proportion</td>
<td>0.189***</td>
<td>0.032</td>
</tr>
<tr>
<td>Home travel to work drove a private vehicle proportion</td>
<td>-0.359***</td>
<td>0.032</td>
</tr>
<tr>
<td>Home travel to work walked or jogged proportion</td>
<td>-0.202**</td>
<td>0.092</td>
</tr>
<tr>
<td>Mean number of household members</td>
<td>0.053***</td>
<td>0.010</td>
</tr>
<tr>
<td>No motor vehicle proportion</td>
<td>0.373***</td>
<td>0.121</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>-0.275***</td>
<td>0.048</td>
</tr>
<tr>
<td>Two motor vehicles proportion</td>
<td>-0.106*</td>
<td>0.057</td>
</tr>
<tr>
<td>Treated x period</td>
<td>0.019***</td>
<td>0.007</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

### 7.3.5 Commuters travelling by bus to the CBD

The analysis in the previous section looked at the propensity of people living near a busway station to commute to any work location by bus. Given that the busway primarily serves commuters travelling between the North Shore and the CBD, we performed additional analysis of the busway’s impact on travel to work in the CBD.

Statistics NZ’s 2006 and 2013 travel to work data, which provides commuter travel information at a CAU-to-CAU level, was used to construct the outcome variable: the proportion of residents of each CAU travelling to work by bus to the three main CBD area units.51 Explanatory variables include CAU-level demographic and travel information from the 2006 and 2013 censuses, and geographic proximity variables that define or categorise a CAU’s distance from the CBD, train stations and busway stations.

The treated units include 25 North Shore CAUs located near busway stations, and the treatment group is the same as in section 7.3.4. Control CAUs located within 2km of the busway were excluded to help satisfy the SUTVA assumption. PSM was used to improve the balance of control and treated units. Table B.1 in appendix B lists the CAU-level attributes used in the PSM algorithm; as in the previous section these are related to personal and commuting characteristics of residents of the treatment group.

We used the nearest neighbour matching technique, and matched two controls for every treated unit. Figure 7.11 shows that the propensity scores of control units are not well distributed. Consequently, discarding treated units outside the area of common support will compromise the size of the matched dataset. All 75 matched units were therefore used in the subsequent analysis. A geographic representation of the matched CAUs is presented in figure 7.12.

---

51 Auckland Central East, Auckland Central West, and Auckland Habourside
Once again, we applied DiD regression to the matched data to estimate the busway’s impact on the outcome of interest.
Proportion of bus commuters to CBD\textsubscript{it} = \(\beta_0 + \beta_1 \cdot \text{couple with children proportion}_{it} + \beta_2 \cdot \text{full time employed proportion}_{it} + \beta_3 \cdot \text{train to CBD proportion}_{it} + \beta_4 \cdot \text{drove to CBD proportion}_{it} + \beta_5 \cdot \text{median hh member count}_{it} + \beta_6 \cdot \text{median age}_{it} + \beta_7 \cdot \text{median hh income}_{it} + \beta_8 \cdot \text{no vehicle proportion}_{it} + \beta_9 \cdot \text{2 vehicles proportion}_{it} + \beta_{10} \cdot \text{3 or more vehicles proportion}_{it} + \beta_{11} \cdot \text{occupied private dwelling proportion}_{it} + \beta_{12} \cdot \text{rented dwelling proportion}_{it} + \beta_{13} \cdot \text{treated}_{it} + \beta_{14} \cdot \text{period}_{it} + \beta_{15} \cdot \text{treated x period}_{it} + e_{it}\)

(Equation 7.5)

Table 7.8 presents the results of the full PSM- DiD model. The model estimates the busway increased the proportion of CBD-bound bus commuters in each CAU by 9.0 percentage points on average (p-value = 0.002). The average proportion of commuters travelling by bus to the CBD in the control group is 24%. Thus, a 9.0 percentage point increase in bus commuters to the CBD is substantial.

We also estimated a reduced version of this model, using only statistically significant control variables. Results from the reduced model (Table 7.9) show the treatment effect could be as large as 9.7%. An F-test shows the insignificant control variables within the full model do not have a joint statistical significance at the 5% level; in other words, taken together they do not add any meaningful variation to the model.

Therefore, our preferred model specification is the reduced model, which estimates the busway increased the proportion of bus commuters to the CBD in each CAU by a statistically significant 9.7 percentage points on average (p-value < 0.01).

Table 7.8 PSM- DiD regression results – full model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.738*</td>
<td>0.421</td>
</tr>
<tr>
<td>Treated</td>
<td>0.055***</td>
<td>0.020</td>
</tr>
<tr>
<td>Period</td>
<td>0.004</td>
<td>0.022</td>
</tr>
<tr>
<td>Couple with children proportion</td>
<td>-0.224</td>
<td>0.247</td>
</tr>
<tr>
<td>Full time employed over age 15 proportion</td>
<td>1.278***</td>
<td>0.258</td>
</tr>
<tr>
<td>Travel to CBD by train proportion</td>
<td>-0.973</td>
<td>0.820</td>
</tr>
<tr>
<td>Travel to CBD by private vehicle proportion</td>
<td>-0.162**</td>
<td>0.069</td>
</tr>
<tr>
<td>Mean number of household members</td>
<td>0.445***</td>
<td>0.080</td>
</tr>
<tr>
<td>Median age</td>
<td>0.010**</td>
<td>0.004</td>
</tr>
<tr>
<td>Median household income</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>No motor vehicle proportion</td>
<td>0.257</td>
<td>0.543</td>
</tr>
<tr>
<td>Occupied private dwellings separate house proportion</td>
<td>-0.137</td>
<td>0.105</td>
</tr>
<tr>
<td>Rented dwellings proportion</td>
<td>-0.545***</td>
<td>0.162</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>-1.266***</td>
<td>0.257</td>
</tr>
<tr>
<td>Two motor vehicles proportion</td>
<td>-1.652***</td>
<td>0.364</td>
</tr>
<tr>
<td>Treated x period</td>
<td>0.090***</td>
<td>0.028</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01
Table 7.9  PSM- DiD regression results – reduced model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.791**</td>
<td>0.363</td>
</tr>
<tr>
<td>Treated</td>
<td>0.059***</td>
<td>0.020</td>
</tr>
<tr>
<td>Period</td>
<td>0.017</td>
<td>0.019</td>
</tr>
<tr>
<td>Full time employed over age 15 proportion</td>
<td>1.301***</td>
<td>0.224</td>
</tr>
<tr>
<td>Travel to CBD by private vehicle proportion</td>
<td>-0.134**</td>
<td>0.067</td>
</tr>
<tr>
<td>Mean number of household members</td>
<td>0.376***</td>
<td>0.061</td>
</tr>
<tr>
<td>Median age</td>
<td>0.011***</td>
<td>0.004</td>
</tr>
<tr>
<td>Rented dwellings proportion</td>
<td>-0.407***</td>
<td>0.149</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>-1.297***</td>
<td>0.235</td>
</tr>
<tr>
<td>Two motor vehicles proportion</td>
<td>-1.606***</td>
<td>0.245</td>
</tr>
<tr>
<td>Treated x period</td>
<td>0.097***</td>
<td>0.028</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

7.3.6  AT HOP users travelling by bus to the CBD

The analysis in the previous two sections focused on the impact of the busway on travel to work, since that is the only type of travel that is available in census data. However, the busway also facilitates other types of travel. We do not have data on all types of travel by all transport modes, but we could obtain from AT a cross-section of travel patterns of HOP card users for 2016.52

The AT data contains total AT HOP trips by bus from every Auckland CAU to every other CAU, for 2016. From this data, we cannot tell which bus service was used for each trip, or the number of trips that used the busway. However, given that the busway services focus on travel to and from the CBD, we defined the outcome of interest as the proportion of bus trips to the CBD53 as a proportion of all bus trips originating from a given CAU. This analysis therefore looks at whether the busway has encouraged travel to the CBD by public transport users.

The treated group includes 25 North Shore CAUs located near the busway (as in sections 7.3.4 and 7.3.5), and the potential control group includes all other Auckland CAUs, less those located on a Gulf Harbour Island or within 2km of a busway station.

Like the other bus patronage analyses, we used PSM to improve the overlap and balance of analysis units. We used the same matching variables as used in the PSM of section 7.3.4 (see table B.1 for a full list of matching variables). Figure 7.13 shows that control unit propensity scores are sufficiently dispersed to enable some ‘unsupported’ treated units to be omitted from the matched data. A total of 24 control units and 19 treated units make up the matched dataset. Figure 7.14 provides a geographic representation of the matched CAUs.

52 HOP is Auckland Transport’s public transport electronic card payment system. It is also possible to pay cash for travel on public transport services; cash paying customers are excluded from our analysis in this section as origin-destination data for such customers is not available.

53 CAUs of Auckland Central East, Auckland Central West, and Auckland Harbourside
Unlike the above bus patronage analyses, the data is only available for a single year (2016), making a DiD model infeasible. Therefore, we employ a PSM-weighted-cross sectional model, i.e., a weighted cross-sectional regression using the propensity scores of the matched data as weights. We assigned weights of 1 and $PS/(1-PS)$ to the matched treated and control units, respectively, where $PS$ is the propensity score for a unit (as recommended by Khandker et al. 2010). With these weights, the estimated coefficient on the treatment dummy variable is an estimate of the ATET.

The functional form of this model is set out below:

\[
\begin{align*}
\mathbb{P}(\text{AHOP} | \text{CBD} & ) = \beta_0 + \beta_1 \cdot \text{Couple with children proportion}_i + \beta_2 \cdot \text{Full time employed proportion}_i + \\
&\beta_3 \cdot \text{Bicycle to work proportion}_i + \beta_4 \cdot \text{Drove to work proportion}_i + \beta_5 \cdot \text{Train to work proportion}_i + \\
&\beta_6 \cdot \text{Walked to work proportion}_i + \beta_7 \cdot \text{Median hh member count}_i + \beta_8 \cdot \text{Median age}_i + \beta_9 \cdot \text{Median hh income}_i + \\
&\beta_{10} \cdot \text{No vehicle proportion}_i + \beta_{11} \cdot \text{2 vehicles proportion}_i + \beta_{12} \cdot \text{3 or more vehicles proportion}_i + \\
&\beta_{13} \cdot \text{Occupied private dwelling proportion}_i + \beta_{14} \cdot \text{Rented dwelling proportion}_i + \beta_{15} \cdot \text{Distance to CBD}_i + \\
&\beta_{16} \cdot \text{Treated}_i + \epsilon_i
\end{align*}
\]

(Equation 7.6)

Estimates from this weighted regression are presented in Table 7.10. The busway's impact on the proportion of AT HOP trips to the CBD is represented by the coefficient on the treated variable. The model estimates that the busway induced a 10.1 percentage point increase in CBD-bound AT HOP trips ($p$-value = 0.02). For perspective, the average proportion of AT HOP bus trips to the CBD in the control group is 31% Therefore, a 10 percentage point increase in CBD-bound AT HOP trips from the busway is relatively important.
Table 7.10  PSM- cross- sectional weighted regression results – full model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.703</td>
<td>2.33</td>
</tr>
<tr>
<td>Treated</td>
<td>0.101**</td>
<td>0.04</td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>Couple with children proportion</td>
<td>0.003</td>
<td>1.26</td>
</tr>
<tr>
<td>Full time employed over age 15 proportion</td>
<td>0.763</td>
<td>1.34</td>
</tr>
<tr>
<td>Home travel to work bicycle proportion</td>
<td>18.03**</td>
<td>8.07</td>
</tr>
<tr>
<td>Home travel to work drove a private vehicle proportion</td>
<td>- 0.644</td>
<td>1.18</td>
</tr>
<tr>
<td>Home travel to work train proportion</td>
<td>145.7**</td>
<td>55.7</td>
</tr>
<tr>
<td>Home travel to work walked or jogged proportion</td>
<td>- 6.259***</td>
<td>2.05</td>
</tr>
<tr>
<td>Mean number of household members</td>
<td>0.143</td>
<td>0.45</td>
</tr>
<tr>
<td>Median age</td>
<td>0.008</td>
<td>0.03</td>
</tr>
<tr>
<td>Median household income</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>No motor vehicle proportion</td>
<td>1.210</td>
<td>2.45</td>
</tr>
<tr>
<td>Occupied private dwellings separate house proportion</td>
<td>0.101</td>
<td>0.49</td>
</tr>
<tr>
<td>Rented dwellings proportion</td>
<td>- 1.008</td>
<td>0.96</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>- 4.162*</td>
<td>1.52</td>
</tr>
<tr>
<td>Two motor vehicles proportion</td>
<td>- 0.397</td>
<td>1.63</td>
</tr>
</tbody>
</table>

* p- value < 0.10, ** p- value < 0.05, *** p- value < 0.01

Summary statistics of a reduced model (table 7.11) show a greater treatment effect of 13.1% and a stronger statistical significance (p- value < 0.01). An F- test shows the insignificant control variables in the
full model do not have a joint statistical significance at the 5% level. Therefore, our preferred model specification is the reduced model, which estimates the busway increased the proportion of AT HOP trips to the CBD by a statistically significant 13 percentage points.

Table 7.11 SM- OLS- weighted regression results - reduced model

<table>
<thead>
<tr>
<th>Outcomes variable = proportion of AT HOP users travelling by bus to CBD</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.815***</td>
<td>0.18</td>
</tr>
<tr>
<td>Treated</td>
<td>0.131***</td>
<td>0.04</td>
</tr>
<tr>
<td>Home travel to work drove a private vehicle proportion</td>
<td>24.47***</td>
<td>5.07</td>
</tr>
<tr>
<td>Home travel to work train proportion</td>
<td>109.0***</td>
<td>32.5</td>
</tr>
<tr>
<td>Home travel to work walked or jogged proportion</td>
<td>- 5.306***</td>
<td>1.27</td>
</tr>
<tr>
<td>Three or more motor vehicles proportion</td>
<td>- 3.203***</td>
<td>0.72</td>
</tr>
</tbody>
</table>

*p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

7.3.7 Satisfaction of bus passengers travelling on the Busway

This analysis looks at the influence of the busway on customer satisfaction levels measured by AT’s ongoing customer satisfaction survey. All variables in this analysis (outcomes and explanatory factors) are from AT customer satisfaction survey data provided to us.

As only a 2016 cross section of data was available, our analysis compares the difference in customer satisfaction between passengers on busway and non-busway services, without considering pre-busway satisfaction levels. This makes it difficult to be sure the estimated effects of the busway on reported customer satisfaction are not due to characteristics of busway passengers rather than the busway itself, although we have controlled for some characteristics of passengers using data reported in the survey.

AT reports various types of customer satisfaction (eg satisfaction with vehicle, travel time, staff and value for money). We focus on six different measures of customer satisfaction, as shown in table 7.12.

Table 7.12 Outcome names given to customer satisfaction measures

<table>
<thead>
<tr>
<th>Outcome name</th>
<th>Subject of AT survey question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall service</td>
<td>Overall satisfaction of service</td>
</tr>
<tr>
<td>Value for money</td>
<td>Value for money of this trip</td>
</tr>
<tr>
<td>Journey time</td>
<td>Time taken for the journey</td>
</tr>
<tr>
<td>On time</td>
<td>Vehicles arriving and departing on time</td>
</tr>
<tr>
<td>Vehicle</td>
<td>Overall satisfaction with the vehicle you are on</td>
</tr>
<tr>
<td>Stop</td>
<td>Overall satisfaction with the stop/station/wharf you got on</td>
</tr>
</tbody>
</table>

Respondents answer each question using to a 0-10 scale (0 = very dissatisfied and 10 = very satisfied). To be consistent with AT’s reported satisfaction measures, we calculate customer satisfaction by route as the proportion of responses equal to or greater than six on the 10-point scale. For example, an 80% satisfaction score corresponds to a situation where 80% of respondents gave a score of six or more.

Explanatory variables include: dummies for route mode (train, bus or ferry), the proportion of respondents by route travelling for a given purpose (education, work and personal), and the proportion of respondents by route who report using public transport because it is superior to their private alternatives.
Analysis was undertaken at the route level using 2016 data. We identify treated units (routes) by filtering for routes that are bus services, in the ‘north’ sector, but not operated by Birkenhead Transport since Birkenhead Transport mainly serves an area that is not connected to the busway. Routes with fewer than 20 responses in the customer satisfaction survey were excluded to ensure robust statistical inference. Using AT’s bus route information, we omitted four routes that do not travel via the busway or use it only over a relatively short segment, and assigned the remaining seven routes into the treatment group.54 To identify the (63) control routes, we simply excluded routes with fewer than 20 responses, and those previously selected for the treatment group. Our analysis therefore compares customer satisfaction for busway services with all other public transport services in Auckland.

As the data is a 2016 cross section, we are limited to a simple cross-sectional regression model. We use the same functional form for all six customer satisfaction outcomes, as set out below in equation 7.7 (i represents a bus route). Corresponding regression results are presented in table 7.13.

\[
\text{Customer Satisfaction}_i = \beta_0 + \beta_1.\text{treated}_i + \beta_2.\text{ferry dummy}_i + \beta_3.\text{train dummy}_i + \beta_4.\text{travelling to work proportion}_i + \beta_5.\text{travelling to education proportion}_i + \beta_6.\text{travelling for personal reasons proportion}_i + \beta_7.\text{PT option superior to private proportion}_i + e_i
\]  

(Equation 7.7)

Of the six outcomes analysed, ‘journey time’ is the only customer satisfaction measure shown to be significantly different (p-value < 0.05) on busway routes compared with non-busway routes. The treated coefficient in this model is 0.10, which implies that, on average, a greater proportion of busway users (10% more) are satisfied with their journey time (ie report a 6 or greater for overall satisfaction) than non-busway services. The lack of statistically significant (p-value > 0.05) treated coefficients for the other five models means we cannot say the busway improved satisfaction with respect to these outcomes. However, as noted above this analysis is affected by data limitations that may make it difficult to isolate the effect of the busway on satisfaction.

Table 7.13  Cross- sectional regression of customer satisfaction scores

<table>
<thead>
<tr>
<th>Outcome variable = customer satisfaction with...</th>
<th>Overall service</th>
<th>Value for money</th>
<th>Journey time</th>
<th>On time</th>
<th>Vehicle</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory variables</td>
<td>Coefficient (SE)</td>
<td>Coefficient (SE)</td>
<td>Coefficient (SE)</td>
<td>Coefficient (SE)</td>
<td>Coefficient (SE)</td>
<td>Coefficient (SE)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.43*** (0.26)</td>
<td>1.08*** (0.36)</td>
<td>1.52*** (0.32)</td>
<td>1.76*** (0.44)</td>
<td>0.89*** (0.17)</td>
<td>0.94*** (0.28)</td>
</tr>
<tr>
<td>Treated</td>
<td>0.04 (0.03)</td>
<td>0.03 (0.04)</td>
<td>0.10*** (0.04)</td>
<td>0.06 (0.05)</td>
<td>-0.01 (0.02)</td>
<td>0.02 (0.03)</td>
</tr>
<tr>
<td>Ferry</td>
<td>0.04 (0.04)</td>
<td>-0.09 (0.06)</td>
<td>0.09* (0.05)</td>
<td>0.12 (0.07)</td>
<td>-0.06** (0.03)</td>
<td>-0.03 (0.05)</td>
</tr>
<tr>
<td>Train</td>
<td>0.04 (0.03)</td>
<td>0.03 (0.05)</td>
<td>0.07* (0.04)</td>
<td>0.16*** (0.06)</td>
<td>0.00 (0.02)</td>
<td>0.03 (0.04)</td>
</tr>
<tr>
<td>Work</td>
<td>-0.55** (0.27)</td>
<td>-0.36 (0.37)</td>
<td>-0.72** (0.33)</td>
<td>-1.04*** (0.46)</td>
<td>0.09 (0.18)</td>
<td>-0.02 (0.28)</td>
</tr>
<tr>
<td>Education</td>
<td>-0.78*** (0.26)</td>
<td>-0.65* (0.37)</td>
<td>-1.00*** (0.33)</td>
<td>-1.36*** (0.45)</td>
<td>-0.09 (0.18)</td>
<td>-0.19 (0.28)</td>
</tr>
<tr>
<td>Personal</td>
<td>-0.62*** (0.30)</td>
<td>-0.24 (0.41)</td>
<td>-0.76* (0.37)</td>
<td>-1.04** (0.51)</td>
<td>0.01 (0.2)</td>
<td>-0.11 (0.32)</td>
</tr>
<tr>
<td>Better than private</td>
<td>0.03 (0.10)</td>
<td>0.17 (0.14)</td>
<td>0.12 (0.13)</td>
<td>0.09 (0.18)</td>
<td>0.11 (0.07)</td>
<td>0.08 (0.11)</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01

54 Busway routes include the 76X, 839, 858, 879, 881X, 956, and NEX.
7.3.8 Total bus patronage

Our final piece of analysis of the busway is its impact on total bus patronage in Auckland. Monthly patronage data is available from AT back to July 2005. We used this data to estimate an ITSA model to test whether the busway had an impact on the level and/or trend of total bus patronage in Auckland.

The monthly bus patronage data was converted to quarterly totals for analysis and runs from the third quarter of 2005 to the third quarter of 2016. Quarters from 2008 onwards were defined as the ‘treatment’ period when the busway was fully operational.

Bus patronage in Auckland has grown steadily over time, and to isolate the impacts of the busway, it is important to control for other factors that may have driven this trend. We obtained data on Auckland’s population, national real GDP per capita, the unemployment rate, real petrol prices, and real road passenger transport prices for use as explanatory variables. The models we tested also included a deterministic trend, a busway dummy variable, and an interaction between the busway dummy and the deterministic trend. Seasonal dummy variables were included to capture the seasonal pattern in quarterly patronage.

Our initial model was a linear regression of quarterly bus patronage on all the explanatory variables listed above. Diagnostic testing indicated strong evidence of serial correlation in the residuals of this model, so we re-estimated the model with an autoregressive error component. This model could not be estimated due to the high correlation between Auckland’s population and the deterministic trend variable. We therefore omitted the population variable, so the effect of population growth is captured by the estimated coefficient on the trend variable.

The estimated coefficients of this model are shown in table 7.14, and its functional form is:

\[
Q_{CCt} = \beta_0 + \beta_1 Q2 \text{ dummy}_t + \beta_2 Q3 \text{ dummy}_t + \beta_3 Q4 \text{ dummy}_t + \beta_4 \text{ Real GDP per capita}_t + \beta_5 \text{ Unemployment rate}_t + \beta_6 \text{ Real petrol price}_t + \beta_7 \text{ Real road passenger price index}_t + \beta_8 \text{ Time trend}_t + \beta_9 \text{ Treated}_t + \beta_{10} \text{ Treated} \times \text{ time}_t + \beta_{11} \text{ AR(1) error}_t + \epsilon_t
\]

(Equation 7.8)

The busway dummy variable and the interaction between this dummy and the time trend are significant at the 1% and 10% levels respectively, and the two coefficients are jointly significant at the 5% level.

Table 7.14 Estimated ITSA model for total quarterly bus patronage

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficient</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>22,568.0***</td>
<td>7460.1</td>
</tr>
<tr>
<td>Q2 dummy</td>
<td>745.8***</td>
<td>137.6</td>
</tr>
<tr>
<td>Q3 dummy</td>
<td>1,070.5***</td>
<td>140.7</td>
</tr>
<tr>
<td>Q4 dummy</td>
<td>-422.2***</td>
<td>112.2</td>
</tr>
<tr>
<td>Real GDP per capita</td>
<td>-1.4**</td>
<td>0.6</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>-116.0</td>
<td>105.5</td>
</tr>
<tr>
<td>Real petrol price index</td>
<td>551.1</td>
<td>755.7</td>
</tr>
<tr>
<td>Real road passenger price index</td>
<td>2514.0</td>
<td>2527.8</td>
</tr>
<tr>
<td>Time trend</td>
<td>90.5</td>
<td>56.7</td>
</tr>
<tr>
<td>Busway dummy</td>
<td>-959.4***</td>
<td>372.3</td>
</tr>
<tr>
<td>Busway dummy x time trend</td>
<td>80.5*</td>
<td>46.0</td>
</tr>
<tr>
<td>Autoregressive (1) error</td>
<td>0.3*</td>
<td>0.2</td>
</tr>
</tbody>
</table>

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01
Figure 7.15 shows the quarterly patronage levels implied by this model, with and without the busway. The model predicts that the busway increased patronage by an increasing amount over time, as shown by the difference between the two predicted lines in this figure.

Figure 7.15   Estimated impacts of the busway on total bus patronage

![Chart showing estimated impacts of the busway on total bus patronage](chart.png)

Note: that the difference between the two predictions is greater than total busway patronage in most quarters, hence these estimates of the busway impact are implausibly large.

However, in most quarters (from 2010 onwards) the model predicts the impact of the busway on total bus patronage exceeds total patronage on the busway itself. For example, in the third quarter of 2016, the model predicts the busway increased bus patronage by 2.66 million passengers, but the total number of passengers on the busway (on NEX and all other services) was only 1.22 million in that quarter. This is not a plausible result – we should expect the net impact of the busway on total patronage to be less than actual patronage on the busway since some busway passengers would have travelled on other services if the busway did not exist.

This implausible result is likely due to some factors that have affected total patronage over time being omitted from the model we have estimated. The impacts of such factors have been partly picked up by the busway variables, leading to an over-estimate of the busway’s impact. This analysis illustrates the importance of sense-checking the results of causal analysis whenever possible.

In other words, it seems likely the busway (including NEX) increased bus patronage in Auckland, but this particular model does not give a reliable estimate of the size of that effect.\(^\text{55}\)

---

\(^\text{55}\) Overseas evidence of the impact of bus rapid transit on patronage is provided by Stewart et al (2017).
7.4 Summary

Table 7.15 summarises the impacts of the busway on each of the outcomes we have analysed above. Overall, we find strong evidence that the busway:

- reduced traffic volumes (AADT) on the Harbour Bridge
- increased the proportions of North Shore residents living near to busway stations commuting to work at any location and in the CBD by bus
- increased the proportions of AT HOP users travelling from locations near to busway stations to the CBD
- increased public transport customer satisfaction with journey times.

Our analysis was constrained by data availability. It was not possible to evaluate all the stated objectives of the busway due to a lack of suitable data for some objectives. For some other objectives that we could evaluate, data constraints forced us to use less reliable methods, such as simple cross-sectional or time-series regression models. This highlights the importance of considering the data required for robust ex-post evaluation of a project’s objectives before the project is implemented.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Estimated effect</th>
<th>Evaluation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic volumes on the Harbour Bridge (AADT)</td>
<td>Strong evidence of a reduction in AADT of 6,500 vehicles/day (3.9% of estimated counterfactual AADT)</td>
<td>ITSA (Poisson model)</td>
</tr>
<tr>
<td>Travel time variability on the Northern Motorway</td>
<td>Weak evidence of a long-term increase in TTV of over 100%</td>
<td>DiD</td>
</tr>
<tr>
<td>Congestion on the Northern Motorway</td>
<td>No evidence of an effect</td>
<td>DiD</td>
</tr>
<tr>
<td>Business activity and employment on the North Shore around busway stations</td>
<td>No evidence of an effect</td>
<td>PSM-DiD</td>
</tr>
<tr>
<td>Proportion of North Shore residents living near busway stations travelling to work by bus</td>
<td>Strong evidence of an increase by 1.9 percentage points (relative to 8.7% of residents travelling to work by bus in the control group).</td>
<td>PSM-DiD</td>
</tr>
<tr>
<td>Proportion of North Shore residents living near busway stations travelling to work in the CBD by bus</td>
<td>Strong evidence of an increase by 9.7 percentage points (relative to 24% of residents travelling to work in the CBD by bus in the control group).</td>
<td>PSM-DiD</td>
</tr>
<tr>
<td>Proportion of AT HOP users travelling from locations near busway stations to the CBD</td>
<td>Strong evidence of an increase by 13 percentage points (relative to 31% of AT HOP users travelling to the CBD by bus in the control group)</td>
<td>PSM-weighted cross-sectional regression</td>
</tr>
<tr>
<td>Customer satisfaction of public transport users (various measures)</td>
<td>Strong evidence of a 10% increase in the number of people reporting high satisfaction with journey time. No evidence of effects on other customer satisfaction measures we analysed.</td>
<td>Cross-sectional regression</td>
</tr>
<tr>
<td>Total bus patronage in Auckland</td>
<td>Strong evidence of an effect but the estimated effect exceeds total busway patronage, suggesting the estimate is biased.</td>
<td>ITSA</td>
</tr>
</tbody>
</table>
8 Conclusions and recommendations

Ex-post evaluation of transport interventions often seeks to understand the effects caused by these interventions. As we have shown in this report, estimating causal effects from observational data, i.e., data that is not produced from a randomised experiment, is challenging and needs to be done with care. To do so, evaluators need to clearly define:

- the treatment itself
- the observational units across which outcomes will be measured and the analysis of treatment effects will be performed
- one or more measurable outcome(s) the treatment was expected to affect (i.e., reflecting the objectives of the treatment)
- a counterfactual or control against which the effects of the treatment will be compared
- how the effect on the outcome(s) of the treatment will be estimated relative to the counterfactual.

Of these criteria, a key challenge is establishing a robust counterfactual against which observed outcomes can be compared. Simple before-and-after comparisons can be useful as a benchmark for comparison with the results of more sophisticated methods, but such comparisons do not give reliable estimates of causal effects because using the ‘before’ observations as the counterfactual does not separate the effects of other factors that may have affected the outcome of interest at the same time as the intervention.

Even if the effects of all other factors can be controlled for by estimating a regression model (i.e., by performing an ITSA), it is difficult to generalise the estimated effects of the intervention beyond the places or groups to which it was applied. And, it is often difficult to be sure all relevant factors that may have affected outcomes over time have been accounted for. If relevant factors are omitted from the analysis, the estimated causal effects can be biased and/or inaccurate.

Comparisons of outcomes between a treatment group and a control group at a single point in time (i.e., cross-sectional analysis) can lead to estimates of causal effects that are easier to generalise beyond the treatment group, if the characteristics of the two groups are similar. The technique of PSM can be used to improve the balance of characteristics across the two groups. However, cross-sectional analysis can only account for observed factors, aside from the intervention, which may have caused different outcomes between the treatment and control groups. Often, there is a concern that other, unobserved, factors also had an influence on outcomes, and such factors can cause estimates of treatment effects to be biased and/or inaccurate.

The IV method provides a way of resolving this problem and eliminating bias created by unobserved factors but it relies on finding a suitable ‘instrument’ which can make it difficult to apply in practice. Instead, other methods such as difference-in-differences and fixed-effects panel data models are often used to control for some types of unobserved factors.

The difference-in-differences method offers a relatively simple way of producing robust estimates of causal effects. It requires data on pre- and post-intervention outcomes for a treatment group and control group, but the members of these groups do not have to be the same at both points in time. In its simplest form, a difference-in-differences estimate of a causal effect can be calculated in a spreadsheet and this will account for the effects of any factors that have a constant effect on outcomes across units in the treatment and control groups or across time. This includes the effects of some types of unobserved factors that cannot be controlled for in a cross-sectional or time series analysis.
More sophisticated difference-in-differences analysis can be done by estimating a simple regression model. This allows the effects of observed factors, not constant across units and time, to be controlled for by including them as explanatory variables in the model. Difference-in-differences analysis can also be easily combined with PSM to ensure there is a good balance of characteristics between the treatment and control groups and further increase the reliability of the estimated causal effects.

Based on the research summarised in this report, we make the following recommendations:

- All attempts at ex-post estimation of the effects caused by transport interventions must carefully consider how to define and measure an appropriate counterfactual for the intervention.
- Simple before-and-after comparisons of outcomes, or comparisons of average outcomes for a treatment and control group, are unlikely to give reliable estimates of causal effects, and these methods should be discouraged in ex-post analysis.
- If the only data available for ex-post analysis is observations of a single outcome over time, an ITS regression model should be used to estimate the effect of an intervention that occurred at a certain point in the time series. Where possible, the regression model should include other variables aside from the intervention that may have affected the outcome of interest over time. However, this method should not be used if the data is available to use other more reliable methods as it is difficult to be sure all other factors have been controlled for and it is difficult to generalise the results of such analysis.
- If the only data available for ex-post analysis is outcomes for treatment and control groups at a single point in time, cross-sectional regression analysis that includes variables measuring observed characteristics of the two groups should be used. However, the results of such analysis could be biased by the effects of unobserved factors that cannot be controlled for, and this possibility should be kept in mind when interpreting the results.
- Whenever possible, a DiD or panel fixed-effects model should be used for estimating causal effects. These methods require pre- and post-intervention data on outcomes for treatment and control groups, and are more likely to produce reliable estimates of causal effects than the other methods described.
- If possible, reporting systems should be set up so suitable data for DiD analysis is available to evaluate interventions ex-post. This requires suitable treatment and control groups to be established, and observing pre-intervention characteristics of both groups. Data on pre- and post-intervention outcomes for both groups is also required.
9 References


Bureau of Transport and Regional Economics (BTRE) (2007) Ex-post economic evaluation of national highway project. Case study 1: Wallaville Bridge, working paper 70.1, Canberra ACT.


Hamilton City Council (2013) *Speed areas and speed management policy: overview of work completed to date and the work proposed for 2013/14 financial year*.


123
Ex-post evaluation of transport interventions using causal inference methods


References


Appendix A: Basic regression models

Regression analysis attempts to understand the average relationship between a response (dependent) and one or more explanatory (independent) variables. A simple slope-intercept equation for a line is a good starting place to understanding the regression model:

\[ y = mx + c \]  
(Equation A.1)

In equation A.1, the response variable \(y\) is perfectly explained by the effect of a unit change in the explanatory variable \(m\), i.e., slope of the line) multiplied by the level of explanatory variable \(x\), plus the y-axis intercept \(c\). This deterministic relationship defines the exact relationship between variables and is illustrated in the left-hand chart of figure A.1.

However, real world relationships are not so straightforward. In reality, other factors that we cannot observe will affect \(y\). To address this problem, regression models take a probabilistic approach and assume outcomes are made up of both a deterministic (e.g., \(mx + c\)) component that is observed, and a random error \(e\) component that is not observed (see right-hand chart of figure A.1). The error component encapsulates all other factors that are not included in the model but that affect outcomes.

Through combining these two components, we arrive at the simple linear regression model:

\[ y = \beta_0 + \beta_1 x + e \]  
(Equation A.2)

Where:
- \(y\) is the actual value of the outcome/dependent variable
- \(x\) is the actual value of the independent/explanatory variable(s)
- \(\beta_0\) is a constant estimated by the model (y-axis intercept)
- \(\beta_1\) is the estimated average effect of unit change(s) in \(x\) to \(y\) (the slope)
- \(e\) is a random error term with zero mean and constant variance.

Figure A.1  Deterministic and probabilistic models
When we estimate the model using a technique such as ordinary least squares, we use data to determine appropriate values for the coefficients $\beta_0$ and $\beta_1$. The process of estimation also yields a set of residuals, which are the differences between the model’s predicted values of $y$ and the actual values of $y$. Including more explanatory variables will reduce the size of the residuals and improve the explanatory power of the model.

Certain assumptions must be met for linear regression to provide the ‘best linear unbiased estimator’ (BLUE) of a given parameter. Two key assumptions specific to the error term are:

- the expected value of error term conditional on $x$ is zero: $E(e|x) = 0$. In other words, on average, errors around the deterministic component of the model balance out (Stock and Watson 2003).
- nothing systematic has been omitted from the model, i.e., there are no factors omitted from the model that affect both the explanatory variables and outcomes, consequently $x$ and $e$ are uncorrelated (Dubin and Rivers 1989).

In a causal inference context, when a variable influencing both the treatment and the outcome is not specified in the regression equation, $X$ and $e$ are correlated. Because the error term represents the effect of all factors influencing $y$ other than those included in the model, the error term will now include the omitted yet relevant variable, which is also correlated with the treatment variable. Therefore, the regression assumption above does not hold, $E(e|x) \neq 0$ (Hardy and Bryman 2004). The implications are that regression estimates are biased, i.e., over- or under-state the true effect of explanatory variables on outcomes.

---

56 The residuals are not exactly the same thing as the error term in the regression model. The error term is more of a theoretical concept which represents a collection of everything that is not accounted for by observable variables included in the model. The residuals are numerical values that can be calculated after the model has been estimated, and reflect how well the model fits the actual observations of $Y$.

57 This condition is also known as omitted variable bias.
## Appendix B: Northern Busway evaluation

Table B.1  PSM variables used in analysis of busway impacts

<table>
<thead>
<tr>
<th>Outcome of interest</th>
<th>Proportion of people travelling by bus to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee and geographic unit count</td>
<td>any work location</td>
</tr>
<tr>
<td><strong>Outcome of interest</strong></td>
<td><strong>Proportion of people travelling by bus to...</strong></td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>Distance from CBD</td>
</tr>
<tr>
<td>Median household income</td>
<td>Median household income</td>
</tr>
<tr>
<td>Employees per km²</td>
<td>Couple with children proportion</td>
</tr>
<tr>
<td>Geographic units per km²</td>
<td>Full time employed proportion</td>
</tr>
<tr>
<td>Unemployed proportion</td>
<td>Median age</td>
</tr>
<tr>
<td>Workplace labourers proportion</td>
<td>Mean number of household members</td>
</tr>
<tr>
<td>Workplace managers proportion</td>
<td>No motor vehicle proportion</td>
</tr>
<tr>
<td>Workplace professionals proportion</td>
<td>Occupied private dwellings separate house proportion</td>
</tr>
<tr>
<td>Workplace sales workers proportion</td>
<td>Rented dwellings proportion</td>
</tr>
<tr>
<td>Workplace technicians/ trades proportion</td>
<td>Three or more motor vehicles proportion</td>
</tr>
<tr>
<td>Workplace travel to work motor cycle proportion</td>
<td>Two motor vehicles proportion</td>
</tr>
<tr>
<td>Workplace travel to work bicycle proportion</td>
<td>Home travel to work bicycle proportion</td>
</tr>
<tr>
<td>Workplace travel to work drove private vehicle proportion</td>
<td>Travel to CBD by bus proportion</td>
</tr>
<tr>
<td>Workplace travel to work public bus proportion</td>
<td>Home travel to work public bus proportion</td>
</tr>
<tr>
<td>Workplace travel to walk or jogged proportion</td>
<td>Travel to CBD by train proportion</td>
</tr>
<tr>
<td>Workplace travel to train proportion</td>
<td>Home travel to work train proportion</td>
</tr>
<tr>
<td>Workplace administrative workers proportion</td>
<td>Home travel to work walked or jogged proportion</td>
</tr>
</tbody>
</table>
 Appendix C: Hamilton safer speed areas analysis code

**R code**

```
## analysis of HCC's "Safer speed areas" (Treated streets -> 50kmh to 40kmh in 2011)
# Setup
rm(list = ls())
require(MASS)
require(AER)
require(dplyr)  # for data manipulation
require(broom)  # for summarising results
require(forecast)  # for moving average calculation
require(car)  # for F-tests

# Import data
hcc_data = read.csv("../data/hcc_data.csv",
                   stringsAsFactors = FALSE)
names(hcc_data) = tolower(names(hcc_data))
hcc_itsa_data = read.csv("../data/hcc_itsa_data.csv",
                        stringsAsFactors = FALSE)
names(hcc_itsa_data) = tolower(names(hcc_itsa_data))

# Correlation matrix
cor(hcc_data[,c(2:11)])

# Naive comparisons of means
# Post-intervention crash count: Treatment vs control group
t1 = t.test(filter(hcc_data, treated == 1)$post.crash.count,
            filter(hcc_data, treated == 0)$post.crash.count,
            var.equal = FALSE)
print(tidy(t1))

# Post- vs pre-intervention crash count for treatment group
t2 = t.test(filter(hcc_data, treated == 1)$post.crash.count,
            filter(hcc_data, treated == 1)$pre.crash.count,
            paired = TRUE)
print(tidy(t2))

# Simple regression models
# Post-intervention comparison of treatment group & control group: regression on dummy variable
r1 = lm(post.crash.count ~ treated, data = hcc_data)
print(summary(r1))
r1_poisson = glm(post.crash.count ~ treated, family = "poisson", data = hcc_data)
print(summary(r1_poisson))
r1_poisson_treatment_effect = exp(coef(r1_poisson)["(Intercept)"])
    * (exp(coef(r1_poisson)["treated"] - 1))
print(paste("Treatment effect:", r1_poisson_treatment_effect))
```
Appendix C: Hamilton Safer Speed Areas analysis code

print(dispersiontest(r1_poisson, trafo = 1))
r1_nb = glm.nb(post.crash.count ~ treated, data = hcc_data)
print(summary(r1_nb))
r1_nb_treatment_effect = exp(coef(r1_nb)["(Intercept)"]) * 
(exp(coef(r1_nb)["treated"]) - 1)

# Post-intervention comparison of treatment group & control group: regression on dummy 
and controls
r2 = lm(post.crash.count ~ treated + t.int. + cross.int. + pre.vkt + pre.crash.count, 
data = hcc_data)
print(summary(r2))
r2_poisson = glm(post.crash.count ~ treated + t.int. + cross.int. + pre.vkt + 
pre.crash.count, family = poisson(link = "identity"), data = hcc_data)
print(summary(r2_poisson))
dispersiontest(r2_poisson, trafo = 1)
r2_explanatory_not_treated = data.frame(treated = 0, 
t.int. = hcc_data$t.int.,
cross.int. = hcc_data$cross.int.,
pre.vkt = hcc_data$pre.vkt,
pre.crash.count = hcc_data$pre.crash.count)
r2_explanatory_treated = data.frame(treated = 1, 
t.int. = hcc_data$t.int.,
cross.int. = hcc_data$cross.int.,
pre.vkt = hcc_data$pre.vkt,
pre.crash.count = hcc_data$pre.crash.count)
r2_poisson_treatment_effect = mean(predict(r2_poisson, newdata = 
r2_explanatory_treated, type = "response") - 
predict(r2_poisson, newdata = 
r2_explanatory_not_treated, type = "response") )
print( summary(r2_nb))

# Post-intervention comparison of treatment group & control group: restricted model 
excluding insignificant variables
r3 = lm(post.crash.count ~ treated + pre.crash.count, data = hcc_data)
print(summary(r3))
r3_poisson = glm(post.crash.count ~ treated + pre.crash.count, family = "poisson", 
data = hcc_data)
print(summary(r3_poisson))
r3_nb = glm.nb(post.crash.count ~ treated + pre.crash.count, data = hcc_data)
print(summary(r3_nb))

# Create dataframe of treated roads only
hcc_data_treated_pre = hcc_data %>% 
filter(treated == 1) %>% 
select(name, cross.int., t.int., 
crash.count = pre.crash.count, 
vkt = pre.vkt) %>% 
mutate(period = 0)
hcc_data_treated_post = hcc_data %>% 
filter(treated == 1) %>% 
select(name, cross.int., t.int., 
crash.count = post.crash.count, 
vkt = post.vkt) %>% 
mutate(period = 1)
hcc_data_treated = bind_rows(hcc_data_treated_pre, hcc_data_treated_post)
rm(hcc_data_treated_pre, hcc_data_treated_post)

# Post- vs pre-intervention comparison for the treatment group: regression on dummy 
variable
r4 = lm(crash.count ~ period, data = hcc_data_treated)
print(summary(r4))
r4_poisson = glm(crash.count ~ period, family = "poisson", data = hcc_data_treated)
print(summary(r4_poisson))
dispersiontest(r4_poisson, trafo = 1)

# Post- vs pre-intervention comparison for the treatment group: regression on dummy variable and controls
r5 = lm(crash.count ~ period + cross.int. + t.int. + vkt, data = hcc_data_treated)
print(summary(r5))
r5_poisson = glm(crash.count ~ period + cross.int. + t.int. + vkt, family = "poisson",
data = hcc_data_treated)
print(summary(r5_poisson))
dispersiontest(r5_poisson, trafo = 1)

# ----------------------------------------------------------------------------
# ----------------------------------------------------------------------------
# Instrumental variables analysis of post-intervention crash counts
# Hausman test for endogeneity of treatment variable
# 1st stage
hausman_stage1 = glm(treated ~ pre.crash.count + pre.vkt + t.int. + cross.int. +
                     u15.pop.2006 + near.school + crashes.2000,
                     family = binomial(link = "probit"),
data = hcc_data)
print(summary(hausman_stage1))
hcc_data$hausman_stage1_residuals = residuals(hausman_stage1)

# 2nd stage
hausman_stage2 = lm(post.crash.count ~ treated + pre.crash.count + pre.vkt + t.int. +
cross.int. + hausman_stage1_residuals,
data = hcc_data)
print(summary(hausman_stage2))
coefftest(hausman_stage2, vcov = vcovHC(hausman_stage2, "HC0"))

# Two-stage least squares
hcc_data$treated_fitted = fitted(hausman_stage1)
iv1 = lm(post.crash.count ~ treated_fitted + pre.crash.count + pre.vkt + cross.int. +
t.int., data = hcc_data)
print(summary(iv1))

# Fixed effects regression models
# Without controlling for VKT
fe1 = lm(post.crash.count - pre.crash.count ~ treated, data = hcc_data)
print(summary(fe1))

# Controlling for VKT
hcc_data$diff.vkt = hcc_data$post.vkt - hcc_data$pre.vkt
fe2 = lm(post.crash.count - pre.crash.count ~ treated + diff.vkt, data = hcc_data)
print(summary(fe2))

# Difference-in-difference models
# Create "stacked" dataset
hcc_data_pre = hcc_data %>%
    select(name, treated, length, cross.int., t.int.,
crash.count = pre.crash.count,
vkt = pre.vkt) %>%
Appendix C: Hamilton Safer Speed Areas analysis code

```r
mutate(period = 0)
hcc_data_post = hcc_data %>%
  select(name, treated, length, cross.int., t.int.,
         crash.count = post.crash.count,
         vkt = post.vkt) %>%
mutate(period = 1)
hcc_data_stacked = bind_rows(hcc_data_pre, hcc_data_post)
rm(hcc_data_pre, hcc_data_post)

# Simple DiD model
did1 = lm(crash.count ~ period * treated, data = hcc_data_stacked)
print(summary(did1))

# DiD model controlling for VKT
did2 = lm(crash.count ~ period * treated + vkt, data = hcc_data_stacked)
print(summary(did2))

# Interrupted time series analysis
# Create time trend variable, quarterly dummies, and moving average
hcc_itsa_data = hcc_itsa_data %>%
  mutate(q2 = ifelse(quarter == 2, 1, 0),
         q3 = ifelse(quarter == 3, 1, 0),
         q4 = ifelse(quarter == 4, 1, 0),
         crash.count.ma = ma(crash.count, 4))

# Regressions of quarterly crash count
itsa1 = lm(crash.count ~ time.trend + treated * time.trend + q2 + q3 + q4, data =
           hcc_itsa_data)
print(summary(itsa1))
print(linearHypothesis(itsa1, c("treated = 0", "time.trend:treated = 0")))
itsa1_explanatory_not_treated = data.frame(time.trend = hcc_itsa_data$time.trend,
                                           treated = 0,
                                           q2 = hcc_itsa_data$q2,
                                           q3 = hcc_itsa_data$q3,
                                           q4 = hcc_itsa_data$q4,
                                           'time.trend:treated' = 0)

itsa1_treatment_effect = sum((fitted(itsa1) - predict(itsa1, newdata =
                               itsa1_explanatory_not_treated, type = "response"))) / 6

itsa1_poisson = glm(crash.count ~ time.trend + treated * time.trend + q2 + q3 + q4,
                    family = "poisson", data = hcc_itsa_data)
print(summary(itsa1_poisson))
print(linearHypothesis(itsa1_poisson, c("treated = 0", "time.trend:treated = 0")))
print(dispersiontest(itsa1_poisson, trafo = 1))
itsa1_poisson_treatment_effect = sum((fitted(itsa1_poisson) - predict(itsa1_poisson,
                                                                         newdata = itsa1_explanatory_not_treated, type = "response"))[48:56]) / 6

# Regression of quarterly crash count with autoregressive errors
ar_xreg = data.frame(time.trend = hcc_itsa_data$time.trend,
                      treated = hcc_itsa_data$treated,
                      time.trend.treated = hcc_itsa_data$time.trend *
                      hcc_itsa_data$treated,
                      q2 = hcc_itsa_data$q2,
                      q3 = hcc_itsa_data$q3,
                      q4 = hcc_itsa_data$q4)
itsa1_ar = arima(hcc_itsa_data$crash.count,
                 order = c(1, 0, 0),
                 xreg = ar_xreg)
itsa1_ar$p.values = (1 - pnorm(abs(itsa1_ar$coef) / sqrt(diag(itsa1_ar$var.coef))))*2
print(summary(itsa1_ar))
print(linearHypothesis(itsa1_ar, c("treated = 0", "time.trend:treated = 0")))
```

133
Ex-post evaluation of transport interventions using causal inference methods

print(itsa1_ar$p.values)
itsa1_ar_treatment_effect = sum(coef(itsa1_ar)["treated"] +
coef(itsa1_ar)["time.trend.treated"] * 48:56) / 6

# Regression of moving average of quarterly crash count
itsa2 = lm(crash.count.ma ~ time.trend + treated * time.trend, data = hcc_itsa_data)
print(summary(itsa2))
print(linearHypothesis(itsa2, c("treated = 0", "time.trend:treated = 0")))
itsa2_explanatory_not_treated = data.frame(time.trend = hcc_itsa_data$time.trend,
treated = 0,
`time.trend:treated` = 0)
itsa2_treatment_effect = sum((fitted(itsa2) - predict(itsa2, newdata =
itsa2_explanatory_not_treated, type = "response")[3:54])) / 6

# STATA code

* SETUP
* clear all
* capture log close
* set more off
* log using hcc-analysis.log, replace

* READ DATA
* insheet using ".data/hcc_data.csv", clear

* NAIVE COMPARISONS OF MEANS
* Post-intervention crash count: Treatment vs control group
ttest postcrashcount, by(treated) unequal

* Post- vs pre-intervention crash count for treatment group
ttest postcrashcount == precrashcount if treated == 1

* SIMPLE REGRESSION MODELS
* Post-intervention comparison of treatment group & control group: regression on dummy variable
regress postcrashcount treated // Linear regression
poisson postcrashcount treated // Poisson regression
nbreg postcrashcount treated // Negative binomial regression

* Post-intervention comparison of treatment group & control group: regression on dummy and controls
regress postcrashcount treated tint crossint prevkt precrashcount // Linear regression
poisson postcrashcount treated tint crossint prevkt precrashcount // Poisson regression
nbreg postcrashcount treated tint crossint prevkt precrashcount // Negative binomial regression

* Post-intervention comparison of treatment group & control group: reduced models excluding insignificant controls
regress postcrashcount treated precrashcount // Linear regression
poisson postcrashcount treated precrashcount // Poisson regression
nbreg postcrashcount treated precrashcount // Negative binomial regression
Appendix C: Hamilton Safer Speed Areas analysis code

* Post- vs pre-intervention comparison for the treatment group: regression on dummy variable
insheet using "../data/hcc_data_treated.csv", clear
regress crashcount period // Linear regression
poisson crashcount period // Poisson regression
nbreg crashcount period // Negative binomial regression

* Post- vs pre-intervention comparison for the treatment group: regression on dummy variable and controls
regress crashcount period crossint tint vkt // Linear regression
poisson crashcount period crossint tint vkt // Poisson regression
nbreg crashcount period crossint tint vkt // Negative binomial regression

* ----------------------------------------------------------------------------

* ----------------------------------------------------------------------------

* Instrumental variables analysis of post-intervention crash counts

* Hausman test for endogeneity of treatment variable
  * 1st stage
reglogit treated precrashcount prevkt crossint tint u15pop2006 nearschool crashes2000
predict hausman_stage1_residuals, resid
predict treated_fitted

* 2nd stage
regress postcrashcount treated precrashcount prevkt crossint tint
hausman_stage1_residuals, robust

* Two-stage least squares
treatreg postcrashcount precrashcount prevkt crossint tint, treat(treated = u15pop2006
nearschool crashes2000) twostep

* Negative binomial version of second stage
nbreg postcrashcount treated_fitted precrashcount prevkt crossint tint

* ---------------------------------------------------------------------------

* Fixed-effects regression models

* Load panel data and configure panel
insheet using "../data/hcc_data_panel.csv", clear
encode name, gen(roadid)
xtset roadid period

* Without controlling for VKT
xtreg crashcount period treated, fe

* Controlling for VKT
xtreg crashcount period treated vkt, fe

* ---------------------------------------------------------------------------

* Difference-in-difference models
* Note this also requires the panel dataset loaded in the previous example

* Simple DiD model
regress crashcount period treated treatedperiod

* DiD model controlling for VKT
regress crashcount period treated treatedperiod vkt

* ---------------------------------------------------------------------------
Ex-post evaluation of transport interventions using causal inference methods

* Interrupted time series analysis

* Load data and create variables
insheet using "../data/hcc_itsa_data.csv", clear
tsset timetrend
gen treatedtimetrend = treated * timetrend
tssmooth ma crashcountma = crashcount, window(4) replace

* Regressions of quarterly crash count
xi: regress crashcount timetrend treated treatedtimetrend i.quarter // Linear regression
test treated = treatedtimetrend = 0 // F-test of joint significance
xi: poisson crashcount timetrend treated treatedtimetrend i.quarter // Poisson regression
test treated = treatedtimetrend = 0 // F-test of joint significance
xi: nbreg crashcount timetrend treated treatedtimetrend i.quarter // Negative binomial regression
test treated = treatedtimetrend = 0 // F-test of joint significance

* Regression of quarterly crash count with AR error
xi: arima crashcount timetrend treated treatedtimetrend i.quarter, ar(1)
test treated = treatedtimetrend = 0 // F-test of joint significance

* Regression of moving average of quarterly crash count
regress crashcountma timetrend treated treatedtimetrend
test treated = treatedtimetrend = 0 // F-test of joint significance

* Regression of moving average of quarterly crash count with autoregressive errors
arima crashcountma timetrend treated treatedtimetrend, ar(1)
test treated = treatedtimetrend = 0 // F-test of joint significance

* Clean up
log close
set more on
Appendix D: Northern Busway analysis code

R Code

## Analysis of Auckland's Northern Busway

# Setup
rm(list = ls())
require(MASS)
require(AER)
library(forecast)  # for time series charts
require(dplyr)     # for data manipulation
require(broom)     # for summarising results
require(car)       # for F-tests
require(MatchIt)   # for PSM analysis

# Outcome: Change in employment and business geographic unit counts
# Model: PSM-DiD

# Import data
business_data = read.csv("data/business-data.csv", stringsAsFactors = FALSE)

# PSM
business_matching = matchit(treated ~ dist_cbd +
        EC.2006_density +
        GC.2006_density +
        median_household_income_2006 +
        unemployed_over_age_15_proportion_2006 +
        workplace_travel_to_work_bicycle_propportion_2006 +
        workplace_travel_to_work_drove_a_private_vehicle_propportion_2006 +
        workplace_travel_to_work_motor_cycle_or_power_cycle_propportion_2006 +
        workplace_travel_to_work_public_bus_propportion_2006 +
        workplace_travel_to_work_train_propportion_2006 +
        workplace_travel_to_work_walked_or_jogged_propportion_2006 +
        workplace_clerical_and_administrative_workers_over_age_15_propportion_2006 +
        workplace_labourers_over_age_15_propportion_2006 +
        workplace_managers_over_age_15_propportion_2006 +
        workplace_professionals_over_age_15_propportion_2006 +
        workplace_sales_workers_over_age_15_propportion_2006 +
        workplace_technicians_and_trades_workers_over_age_15_propportion_2006,
        data = business_data %>% select(code, contains("2006"), treated, cau_size, dist_cbd) %>% na.omit,
        method = "nearest", ratio = 2)
summary(business_matching)
plot(business_matching, type = "jitter")
business_matched_data = match.data(business_matching)

# Difference-in-difference with PSM
# Create "stacked" dataset
business_matched_data_2006 = business_matched_data %>%
    select(-distance, -weights) %>%
    mutate(period = 0)
business_matched_data_2013 = filter(business_data, code in%
    business_matched_data$code) %>%
    select(code, contains("2013"), treated, cau_size, dist_cbd, -au2013_label) %>%
    mutate(period = 1)
names(business_matched_data_2006) = gsub("2006", ",",
names(business_matched_data_2006))
names(business_matched_data_2013) = gsub("2013", ",",
names(business_matched_data_2013))
names(business_matched_data_2006) = gsub("tion_", "tion",
names(business_matched_data_2006))
names(business_matched_data_2013) = gsub("tion_", "tion",
names(business_matched_data_2013))
business_did_data = bind_rows(business_matched_data_2006, business_matched_data_2013)

# Full DiD model on matched data, outcome = employee count
EC_did_psm_full = lm(EC. ~ period * treated +
workplace_labourers_over_age_15_proportion + workplace_machinery_operators_and_drivers_over_age_15_proportion + workplace_managers_over_age_15_proportion + workplace_professionals_over_age_15_proportion + workplace_sales_workers_over_age_15_proportion + workplace_technicians_and_trades_workers_over_age_15_proportion +
workplace_travel_to_work_bicycle_proportion + workplace_travel_to_work_drove_a_private_vehicle_proportion + workplace_travel_to_work_motor_cycle_or_power_cycle_proportion +
workplace_travel_to_work_other_proportion + workplace_travel_to_work_passenger_in_a_private_or_company_vehicle_proportion + workplace_travel_to_work_public_bus_proportion + workplace_travel_to_work_train_proportion,
data = business_did_data)
print(summary(EC_did_psm_full))

# Reduced model DiD model on matched data, outcome = employee count
EC_did_psm_red = lm(EC. ~ period * treated +
workplace_travel_to_work_drove_a_private_vehicle_proportion + workplace_travel_to_work_train_proportion,
data = business_did_data)
print(summary(EC_did_psm_red))

# F-test for reduced vs full model
anova(EC_did_psm_red, EC_did_psm_full)

# Full DiD on matched data, outcome = geo unit count of businesses
GC_did_psm_full = lm(GC. ~ period * treated +
workplace_labourers_over_age_15_proportion + workplace_machinery_operators_and_drivers_over_age_15_proportion + workplace_managers_over_age_15_proportion + workplace_professionals_over_age_15_proportion + workplace_sales_workers_over_age_15_proportion + workplace_technicians_and_trades_workers_over_age_15_proportion +
workplace_travel_to_work_bicycle_proportion + workplace_travel_to_work_drove_a_private_vehicle_proportion + workplace_travel_to_work_motor_cycle_or_power_cycle_proportion +
workplace_travel_to_work_other_proportion + workplace_travel_to_work_passenger_in_a_private_or_company_vehicle_proportion + workplace_travel_to_work_public_bus_proportion + workplace_travel_to_work_train_proportion,
data = business_did_data)
print(summary(GC_did_psm_full))

# Reduced model DiD model on matched data, outcome = geo unit count
Appendix D: Northern Busway analysis code

```
GC_did_psm_red = lm(GC. ~ period * treated +
    workplace_travel_to_work_drove_a_private_vehicle_proportion +
    workplace_sales_workers_over_age_15_proportion,
    data = business_did_data)
print(summary(GC_did_psm_red))
# F-test for reduced vs full model
anova(GC_did_psm_red, GC_did_psm_full)
# -----------------------------------------------------------------------------
# Outcome: Proportion of people commuting to any work location by bus
# Model: PSM-DiD

# Import data
bus_commuter_data_06 = read.csv("data/bus-commuter-data-06.csv", stringsAsFactors =
    FALSE)
bus_commuter_data_13 = read.csv("data/bus-commuter-data-13.csv", stringsAsFactors =
    FALSE)

# PSM
bus_commuter_matching = matchit(treated ~ dist_cbd +
    couple_with_children_proportion +
    full_time_employed_over_age_15_proportion +
    home_travel_to_work_bicycle_proportion +
    home_travel_to_work_drove_a_private_vehicle_proportion +
    home_travel_to_work_public_bus_proportion +
    home_travel_to_work_train_proportion +
    home_travel_to_work_walked_or_jogged_proportion +
    mean_number_of_household_members +
    median_age +
    median_household_income +
    no_motor_vehicle_proportion +
    occupied_private_dwellings_separate_house_proportion +
    rented_dwellings_proportion +
    three_or_more_motor_vehicles_proportion +
    two_motor_vehicles_proportion,
    data = na.omit(bus_commuter_data_06), method =
    "nearest", ratio = 2, discard = "both")
bus_commuter_matched_data = match.data(bus_commuter_matching)
print(summary(bus_commuter_matching))
plot(bus_commuter_matching, type = "jitter")

# Difference-in-difference with matched data
# Create DiD dataset
bus_commuter_did_data = bind_rows(filter(bus_commuter_data_06, code %in%
    bus_commuter_matched_data$code),
    filter(bus_commuter_data_13, code %in%
    bus_commuter_matched_data$code))

# Full DiD model
bus_commuter_did_psm_full = lm(home_travel_to_work_public_bus_proportion ~ treated *
    period +
    couple_with_children_proportion +
    full_time_employed_over_age_15_proportion +
    home_travel_to_work_bicycle_proportion +
    home_travel_to_work_drove_a_private_vehicle_proportion +
    home_travel_to_work_train_proportion +
    home_travel_to_work_walked_or_jogged_proportion +
    mean_number_of_household_members +
    median_age +
    median_household_income +
```
Ex-post evaluation of transport interventions using causal inference methods

\[
\begin{align*}
\text{no}_{\text{motor vehicle proportion}} + \\
\text{occupied}_{\text{private dwellings separate house proportion}} + \\
\text{rented}_{\text{dwellings proportion}} + \\
\text{three or more}_{\text{motor vehicles proportion}} + \\
\text{two}_{\text{motor vehicles proportion}}, \\
\text{data} = \text{bus_commuter_did_data})
\end{align*}
\]

print(summary(bus_commuter_did_psm_full))

# Reduced DiD model
bus_commuter_did_psm_red = lm(home_travel_to_work_public_bus_proportion ~ treated * period +
    couple_with_children_proportion + 
    full_time_employed_over_age_15_proportion + 
    home_travel_to_work_drove_a_private_vehicle_proportion 
    +
    home_travel_to_work_walked_or_jogged_proportion + 
    mean_number_of_household_members + 
    no_{motor vehicle proportion} + 
    three_or_more_{motor vehicles proportion} + 
    two_{motor vehicles proportion}, 
    data = bus_commuter_did_data)

print(summary(bus_commuter_did_psm_red))

# F-test for reduced vs full model
anova(bus_commuter_did_psm_red, bus_commuter_did_psm_full)

# Outcome: Proportion of people commuting to CBD area units by bus  
# Model: PSM-DiD

# Import data  
cbd_commuter_data_06 = read.csv("data/cbd-commuter-data-06.csv", stringsAsFactors = FALSE)  

# PSM  
cbd_commuter_matching = matchit(treated ~ dist_cbd +
    couple_with_children_proportion + 
    full_time_employed_over_age_15_proportion + 
    bus_to_cbd_prop + 
    train_to_cbd_prop + 
    private_to_cbd_prop + 
    mean_number_of_household_members + 
    median_age + 
    median_household_income + 
    no_{motor vehicle proportion} + 
    occupied_private_dwellings_separate_house_proportion + 
    rented_dwellings_proportion + 
    three_or_more_{motor vehicles proportion} + 
    two_{motor vehicles proportion}, 
    data = cbd_commuter_data_06 %>% na.omit , method = "nearest", ratio = 2)

cbd_commuter_matched_data = match.data(cbd_commuter_matching)
print(summary(cbd_commuter_matching))
plot(cbd_commuter_matching, type = "jitter")

# Difference-in-difference with matched data  
# Create DiD dataset  
cbd_commuter_did_data = bind_rows(filter(cbd_commuter_data_06, code %in%
    cbd_commuter_matched_data$code),
    filter(cbd_commuter_data_13, code %in%
    cbd_commuter_matched_data$code),
    data = bus_commuter_did_data)

print(summary(cbd_commuter_did_data))
plot(cbd_commuter_did_data, type = "jitter")

# F-test for reduced vs full model
anova(cbd_commuter_did_psm_red, cbd_commuter_did_psm_full)
Appendix D: Northern Busway analysis code

```r
filter(cbd_commuter_data_13, code %in% 
cbd_commuter_matched_data$code))

# DiD with matched data
cbd_commuter_did_psm_full = lm(bus_to_cbd_prop ~ treated * period +
couple_with_children_proportion +
full_time_employed_over_age_15_proportion +
train_to_cbd_prop +
private_to_cbd_prop +
mean_number_of_household_members +
median_age +
median_household_income +
no_motor_vehicle_proportion +
occupied_private_dwellings_separate_house_proportion +
rented_dwellings_proportion +
three_or_more_motor_vehicles_proportion +
two_motor_vehicles_proportion,
data = cbd_commuter_did_data)
print(summary(cbd_commuter_did_psm_full))

# Reduced DiD model with matched data
cbd_commuter_did_psm_red = lm(bus_to_cbd_prop ~ treated * period +
full_time_employed_over_age_15_proportion +
private_to_cbd_prop +
mean_number_of_household_members +
median_age +
rented_dwellings_proportion +
three_or_more_motor_vehicles_proportion +
two_motor_vehicles_proportion,
data = cbd_commuter_did_data)
print(summary(cbd_commuter_did_psm_red))

# F-test
anova(cbd_commuter_did_psm_full, cbd_commuter_did_psm_red)

# Outcome: Proportion AT HOP users travelling to the CBD by bus
# Model: PSM-OLS (weighted cross-sectional OLS)

# Import data
hop_data = read.csv("data/hop-data.csv", stringsAsFactors = FALSE)

# PSM
hop_matching = matchit(treated ~ dist_cbd +
couple_with_children_proportion +
full_time_employed_over_age_15_proportion +
home_travel_to_work_bicycle_proportion +
home_travel_to_work_drove_a_private_vehicle_proportion +
home_travel_to_work_public_bus_proportion +
home_travel_to_work_train_proportion +
home_travel_to_work_walked_or_jogged_proportion +
mean_number_of_household_members +
median_age +
median_household_income +
no_motor_vehicle_proportion +
occupied_private_dwellings_separate_house_proportion +
rented_dwellings_proportion +
three_or_more_motor_vehicles_proportion +
two_motor_vehicles_proportion,
data = hop_data, method = "nearest", ratio = 2, discard = "both")
hop_matched_data = match.data(hop_matching)
print(summary(hop_matching))
```
plot(hop_matching, type = "jitter")

# Weighted OLS on matched data
# Weights: Treated variables * 1, Control variables * PS/(1-PS)
hop_matched_data = hop_matched_data %>%
  select(-weights) %>%
  mutate(w1 = ifelse(treated == 1, 1, (distance / (1 - distance))))

hop_ols_psm_full_w = lm(at_prop_cbd_16 ~ treated +
  dist_cbd +
  couple_with_children_proportion +
  full_time-employed_over_age_15_proportion +
  home_travel_to_work_bicycle_proportion +
  home_travel_to_work_drove_a_private_vehicle_proportion +
  home_travel_to_work_train_proportion +
  home_travel_to_work_walked_or_jogged_proportion +
  mean_number_of_household_members +
  median_age +
  median_household_income +
  no_motor_vehicle_proportion +
  occupied_private_dwellings_separate_house_proportion +
  rented_dwellings_proportion +
  three_or_more_motor_vehicles_proportion +
  two_motor_vehicles_proportion,
  data = hop_matched_data, weights = w1)

print(summary(hop_ols_psm_full_w))

# Reduced model
hop_ols_psm_red_w = lm(at_prop_cbd_16 ~ treated +
  home_travel_to_work_bicycle_proportion +
  home_travel_to_work_train_proportion +
  home_travel_to_work_walked_or_jogged_proportion +
  three_or_more_motor_vehicles_proportion,
  data = hop_matched_data, weights = w1)

print(summary(hop_ols_psm_red_w))

# F-test: reduced model vs full model
anova(hop_ols_psm_full_w, hop_ols_psm_red_w)

# Outcome: Customer satisfaction by public transport route
# Model: OLS regression

# Import data
customer_satisfaction = read.csv("data/customer-satisfaction-data.csv",
  stringsAsFactors = FALSE)

# OLS Regressions
customer_satisfaction_model = lm(general_rating ~ treated + ferry + train + work +
  educ + personal + better.than.private,
  data = customer_satisfaction)
summary(customer_satisfaction_model)
customer_satisfaction_value_model = lm(value_rating ~ treated + ferry + train + work +
  educ + personal + better.than.private,
  data = customer_satisfaction)
summary(customer_satisfaction_value_model)
customer_satisfaction_journey_time_model = lm(journey_time_rating ~ treated + ferry +
  train + work + educ + personal + better.than.private,
  data = customer_satisfaction)
summary(customer_satisfaction_journey_time_model)
customer_satisfaction_ontime_model = lm(ontime_rating ~ treated + ferry + train + work +
  educ + personal + better.than.private,
  data = customer_satisfaction)
summary(customer_satisfaction_ontime_model)
customer_satisfaction_vehicle_model = lm(vehicle_rating ~ treated + ferry + train + work + educ + personal + better.than.private, data = customer_satisfaction)
summary(customer_satisfaction_vehicle_model)
customer_satisfaction_stops_model = lm(stop_rating ~ treated + ferry + train + work + educ + personal + better.than.private, data = customer_satisfaction)
summary(customer_satisfaction_stops_model)

# ----------------------------------------------------------------------------
# ----------------------------------------------------------------------------
# Outcome: Harbour Bridge AADT
# Model: ITSA

# Import data
harbour_bridge_data = read.csv("data/harbour-bridge-data.csv", stringsAsFactors = FALSE)

# ITSA basic
bridge_aadt_itsa1 = lm(aadt ~ treated * time.trend + gdp.real + vpt.construction, data = harbour_bridge_data)
print(summary(bridge_aadt_itsa1))
print(linearHypothesis(bridge_aadt_itsa1, c("treated = 0", "treated:time.trend = 0")))

# Test serial correlation
print(bgtest(bridge_aadt_itsa1))
tsdisplay(bridge_aadt_itsa1$residuals)

# ITSA Poisson
bridge_aadt_itsa1_poisson = glm(aadt ~ treated * time.trend + gdp.real + vpt.construction, family = "poisson", data = harbour_bridge_data)
print(summary(bridge_aadt_itsa1_poisson))
print(linearHypothesis(bridge_aadt_itsa1_poisson, c("treated = 0", "treated:time.trend = 0")))
print(dispersiontest(bridge_aadt_itsa1_poisson, trafo = 1))

# Calculate ITSA Poisson treatment effect
harbour_bridge_counterfactual = harbour_bridge_data %>%
  select(-treated) %>%
  mutate(treated = 0)
bridge_aadt_diff = exp(predict(bridge_aadt_itsa1_poisson)) -
  exp(predict(bridge_aadt_itsa1_poisson, harbour_bridge_counterfactual))
print(mean(bridge_aadt_diff[8:15]))
print(mean(bridge_aadt_diff[8:15]) / mean(exp(predict(bridge_aadt_itsa1_poisson, harbour_bridge_counterfactual)[8:15])))

# ----------------------------------------------------------------------------
# ----------------------------------------------------------------------------
# Outcome: Total bus patronage for Auckland City
# Model: ITSA

# Import data
bus_patronage_data = read.csv("data/total-bus-patronage-data.csv", stringsAsFactors = FALSE)

# ITSA model
bus_patronage_itsa1 = lm(bus_patronage ~ factor(quarter) + akl_pop + real_gdp_pc + unempl_rate + cpi_petro1_real + cpi_road_real + time_trend * busway, data = bus_patronage_data)
print(summary(bus_patronage_itsa1))

# Test serial correlation
print(bgtest(bus_patronage_itsa1))
tsdisplay(bus_patronage_itsa1$residuals)
# Ex-post evaluation of transport interventions using causal inference methods

# Outcome: Travel time variability and Congestion
# Model: DiD estimation

# Import data
ttv_variability_data = read.csv("data/travel-time-variability-data.csv", stringsAsFactors = FALSE)

## Outcome = travel time variability = (t_max - t_min) / t_avg
## 05-09
ttv_did_05_09 = lm(var_t ~ treated * period, data = ttv_variability_data %>% filter(year == 2005 | year == 2009))
summary(ttv_did_05_09)
## 05-12
ttv_did_05_12 = lm(var_t ~ treated * period, data = ttv_variability_data %>% filter(year == 2005 | year == 2012))
summary(ttv_did_05_12)

## Outcome = congestion index (minutes of delay per kilometre of travel, given a segment's speed limit)
## 05-09
cgi_did_05_09 = lm(cgi ~ treated * period, data = ttv_variability_data %>% filter(year == 2005 | year == 2009))
summary(cgi_did_05_09)
## 05-12
cgi_did_05_12 = lm(cgi ~ treated * period, data = ttv_variability_data %>% filter(year == 2005 | year == 2012))
summary(cgi_did_05_12)

# STATA Code

* Northern Busway case study analysis

* Note the psmatch2 and ftest packages are required
* These can be installed with:
* `ssc install psmatch2`
* `ssc install ftest`

* Setup
  * clear all
capture log close
  * log using busway-analysis.log, replace

* Outcomes: Change in employment count and change in geographic units
* Model: PSM-DID

* Import data
  * Import delimited ".data/business-data.csv", clear

* Drop unnescessary variables
drop ec2000_density-ec2005_density ec2007_density-ec2012_density ec2013_density-ec2015_density
* Calculate change in employment count and geographic units (2013 vs 2006)
gen ec_change = ec2013 - ec2006
gen gc_change = gc2013 - gc2006

* PSM
#delimit ;
psmatch2 treated
dist_cbd
ec2006_density
gc2006_density
median_household_income_2006
unemp_proportion_2006
wwt_wt_bicycle_proportion_2006
wwt_drove_a_priv_veh_propn_2006
wwt_motor_cycle_proportion_2006
wwt_public_bus_proportion_2006
wwt_train_propportion_2006
wwt_walked_or_jogged_propn_2006
clerical_and_admin_propn_2006
labourer_propn_2006
managers_propn_2006
professionals_propn_2006
sales_workers_propn_2006
technicians_trades_propn_2006,
norepl ;
#delimit cr

* Create dataset for DiD using matched data
drop if missing(_weight) // Unmatched observations have missing _weight
#delimit ;
reshape long ec
gc
couple_with_children_propn_
couple_wo_children_propn_
fte_over_age_15_proportion_
mean_no_of hh members_
median_age_
median_household_income_
no_motor_vehicle_propotion_
not_in_labour_force_propn_
opd_separate_house_propn_
one_motor_veh_propn_
one_parent_w children_propn_
ppte_over_age_15_proportion_
rented_dwellings_propportion_
student_propportion_
three_or_more_mtr_veh_propn_
two_mtr_veh_propn_
unemp_proportion_
usually_resident_population_
clerical_and_admin_propn_
comm_personal_propn_
labourer_propn_
machinery_op_propn_
managers_propn_
professionals_propn_
sales_workers_propn_
technicians_trades_propn_
wwt_wt_bicycle_proportion_
wwt_did_not_work_propn_
wwt_company_vehicle_propn_
wwt_drove_a_priv_veh_propn_
wwt_motor_cycle_proportion_
wwt_other_proportion_
Ex-post evaluation of transport interventions using causal inference methods

```stata
wttw_passenger_proportion_  
wttw_public_bus_proportion_  
wttw_train_proportion_  
wttw_walked_or_jogged_propn_  
wttw_worked_at_home_propn_  
, i(_id) j(year) ;
#delimit cr

gen period = 0
replace period = 1 if year == 2013

#delimit cr

* Full DiD for employment count using matched observations
regress ec period treated period_x_treated
labourer_propn_  
machinery_op_propn_  
managers_propn_  
professionals_propn_  
sales_workers_propn_  
technicians_trades_propn_  
wttw_bicycle_propportion_  
wttw_drove_a_priv_veh_propn_  
wttw_motor_cycle_propportion_  
wttw_other_propportion_  
wttw_passenger_propportion_  
wttw_public_bus_propportion_  
wttw_train_propportion_ ;
#delimit cr

est sto m_full_ec // Save estimates for F-test

* Reduced DiD for employment count using matched observations
regress ec period treated period_x_treated
machinery_op_propn_  
wttw_drove_a_priv_veh_propn_ ;
#delimit cr

est sto m_reduced_ec // Save estimates for F-test

* F-test to compare full and reduced models for employment count
ftest m_full_ec m_reduced_ec

* Full DiD for geographic unit count using matched observations
regress gc period treated period_x_treated
labourer_propn_  
machinery_op_propn_  
managers_propn_  
professionals_propn_  
sales_workers_propn_  
technicians_trades_propn_  
wttw_bicycle_propportion_  
wttw_drove_a_priv_veh_propn_  
wttw_motor_cycle_propportion_  
wttw_other_propportion_  
wttw_passenger_propportion_  
wttw_public_bus_propportion_  
wttw_train_propportion_ ;
#delimit cr

est sto m_full_gc

* Reduced DiD for geographic unit count using matched observations
regress gc period treated period_x_treated
machinery_op_propn_ ;
#delimit cr
```

146
Appendix D: Northern Busway analysis code

est sto m_reduced_gc
* F-test to compare full and reduced models for geographic unit count
ftest m_full_gc m_reduced_gc
* -----------------------------------------------

* -----------------------------------------------
* Outcome: Proportion of people commuting to any work location by bus
* Model: PSM-DiD

* Import data
import delimited "./data/bus-commuter-data.csv", clear

* PSM
#delimit ;
psmatch2 treated
couple_with_children_propn2006
fte_over_age_15_proportion2006
httwp_bicycle_proportion2006
httwp_drove_a_priv_veh_proportion2006
httwp_public_bus_proportion2006
httwp_train_proportion2006
httwp_walked_or_jogged_proportion2006
mean_number_of_hh_members2006
median_age2006
median_household_income2006
no_motor_vehicle_proportion2006
opd_separate_house_proportion2006
rented_dwelling_proportion2006
three_or_more_mtr_veh_proportion2006
two_motor_vehicles_proportion2006
, norepl ;
#delimit cr

* Create dataset for DiD using matched data
drop if missing(_weight)  // Unmatched observations have missing _weight
#delimit ;
reshape long couple_with_children_propn
fte_over_age_15_proportion
httwp_bicycle_proportion
httwp_drove_a_priv_veh_proportion
httwp_public_bus_proportion
httwp_train_proportion
httwp_walked_or_jogged_proportion
mean_number_of_hh_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_proportion
rented_dwelling_proportion
three_or_more_mtr_veh_proportion
two_motor_vehicles, i(_id) j(year) ;
#delimit cr
gen period = 0
replace period = 1 if year == 2013

* Full DiD model using matched data
#delimit ;
regress httwp_public_bus_proportion period treated period_x_treated
couple_with_children_propn
fte_over_age_15_proportion
httwp_bicycle_proportion
Ex-post evaluation of transport interventions using causal inference methods

httw_drove_a_priv_veh_propn
httw_train_proportion
httw_walked_or_jogged_propn
mean_number_of_hh_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_propn
rented_dwelling_proportion
three_or_more_mtr_veh_propn
two_motor_vehicles_propn;
#delimit cr
est sto m_full_bus_commuting
* Reduced DiD model using matched data
#delimit;
regress httw_public_bus_proportion period treated period_x_treated
httw_drove_a_priv_veh_propn
mean_number_of_hh_members
median_age
median_household_income
no_motor_vehicle_proportion
three_or_more_mtr_veh_propn
two_motor_vehicles_propn;
#delimit cr
est sto m_reduced_bus_commuting
* F-test to compare full and reduced models
ftest m_full_bus_commuting m_reduced_bus_commuting
*----------------------------------------------------------------------------
*----------------------------------------------------------------------------
* Outcome: Proportion of people commuting to CBD area units by bus
* Model: PSM-DiD
* Import data
import delimited "./data/cbd-commuter-data.csv", clear
* PSM
#delimit;
psmatch2 treated

dist_cbd
couple_with_children_proportion2006
bus_to_cbd_prop2006
train_to_cbd_prop2006
private_to_cbd_prop2006
mean_number_of_hh_members2006
median_age2006
median_household_income2006
no_motor_vehicle_proportion2006
opd_separate_house_prop2006
rented_dwelling_prop2006
three_or_more_mtr_veh_prop2006
two_motor_vehicles_prop2006
, norepl;
#delimit cr
* Create dataset for DiD using matched data
drop if missing(_weight) // Unmatched observations have missing _weight
#delimit;
reshape long bus_to_cbd_prop
train_to_cbd_prop
private_to_cbd_prop
couple_with_children_proportion2006

Appendix D: Northern Busway analysis code

fte_over_age_15_proportion
httw_bicycle_proportion
httw_drove_a_priv_veh_propn
httw_public_bus_proportion
httw_train_proportion
httw_walked_or_jogged_propn
mean_number_of_hh_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_propn
rented_dwellings_proporation
three_or_more_mtr_veh_propn
two_motor_vehicles_propn
, i(_id) j(year) ;
#delimit cr
gen period = 0
replace period = 1 if year == 2013
gen period_x_treated = period * treated
* Full DiD model using matched data
#delimit ;
regress bus_to_cbd_prop period treated period_x_treated
couple_with_children_propn
fte_over_age_15_proportion
train_to_cbd_prop
private_to_cbd_prop
mean_number_of_hh_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_propn
rented_dwellings_proporation
three_or_more_mtr_veh_propn
two_motor_vehicles_propn ;
#delimit cr
est sto m_full_bus_cbd
* Reduced DiD model using matched data
#delimit ;
regress bus_to_cbd_prop period treated period_x_treated
fte_over_age_15_proportion
train_to_cbd_prop
private_to_cbd_prop
median_household_income
three_or_more_mtr_veh_propn
two_motor_vehicles_propn ;
#delimit cr
est sto m_reduced_bus_cbd
* F-test to compare full and reduced models
ftest m_full_bus_cbd m_reduced_bus_cbd
* -----------------------------------------------------------------------------
* -----------------------------------------------------------------------------
* Outcome: Proportion of AT HOP users travelling to the CBD by bus
* Model: PSM-OLS (weighted cross-sectional OLS)
* Import data
import delimited "./data/hop-data.csv", clear
* PSM
#delimit ;
psmatch2 treated

149
Ex-post evaluation of transport interventions using causal inference methods

```
dist_cbd
couple_with_children_proportion
full_time_empl_over_15_propn
httw_bicycle_proportion
httw_drove_priv_veh_proportion
httw_public_bus_proportion
httw_train_proportion
httw_walked_or_jogged_proportion
mean_number_of_household_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_proportion
rented_dwellings_proportion
three_or_more_motor_veh_propn
two_motor_vehicles_proportion
, norepl ;
#delimit cr
* Create matched dataset and add weights
drop if missing(_weight) // Unmatched observations have missing _weight
gen pw = 1
replace pw = _pscore / (1 - _pscore) if treated == 0
* Weighted OLS on matched data: full model
#delimit ;
regress at_prop_cbd_16
   treated
dist_cbd
couple_with_children_proportion
full_time_empl_over_15_propn
httw_bicycle_proportion
httw_drove_priv_veh_proportion
httw_train_proportion
httw_walked_or_jogged_proportion
mean_number_of_household_members
median_age
median_household_income
no_motor_vehicle_proportion
opd_separate_house_proportion
rented_dwellings_proportion
three_or_more_motor_veh_propn
two_motor_vehicles_proportion
[aweight = pw] ;
#delimit cr
est sto m_full_hop_cbd
* Weighted OLS on matched data: reduced model
#delimit ;
regress at_prop_cbd_16
   treated
   httw_bicycle_proportion
   httw_train_proportion
   httw_walked_or_jogged_proportion
   three_or_more_motor_veh_propn
[aweight = pw] ;
#delimit cr
est sto m_reduced_hop_cbd
* F-test to compare full and reduced models
ftest m_full_hop_cbd m_reduced_hop_cbd
* ---------------------------------------------------------------

* ---------------------------------------------------------------
```

* Outcome: Customer satisfaction by public transport route (various measures)
  * Model: OLS regression

* Import data
  import delimited "./data/customer-satisfaction-data.csv", clear

* Regression models for satisfaction measures
  regress general_rating treated ferry train work educ personal betterthanprivate
  regress value_rating treated ferry train work educ personal betterthanprivate
  regress journey_time_rating treated ferry train work educ personal betterthanprivate
  regress ontime_rating treated ferry train work educ personal betterthanprivate
  regress vehicle_rating treated ferry train work educ personal betterthanprivate
  regress stop_rating treated ferry train work educ personal betterthanprivate

* Outcome: Harbour Bridge AADT
  * Model: ITSA

* Import data
  import delimited "./data/harbour-bridge-data.csv", clear
  gen treated_trend = treated * timetrend
  tsset timetrend

* Basic ITSA model
  regress aadt timetrend treated treated_trend gdpreal vptconstruction

* Test serial correlation
  estat bgodfrey

* Poisson ITSA model
  poisson aadt timetrend treated treated_trend gdpreal vptconstruction

* Calculate treatment effect from Poisson model
  predict aadt_factual if year > 2007
  replace treated = 0 if year > 2007
  replace treated_trend = 0 if year > 2007
  predict aadt_counterfactual if year > 2007
gen treatment_effect = aadt_factual - aadt_counterfactual
  mean treatment_effect

* Outcome: Total bus patronage for Auckland City
  * Model: ITSA

* Import data
  import delimited "./data/total-bus-patronage-data.csv", clear
  gen time_trend_busway = time_trend * busway
  tsset time_trend

* Basic ITSA model
  xi: regress bus_patronage i.quarter akl_pop real_gdp_pc unempl_rate cpi_petrol_real
cpi_road_real time_trend busway time_trend_busway

* Test serial correlation
  estat bgodfrey

*完毕*
Ex-post evaluation of transport interventions using causal inference methods

* Outcome: Travel time variability and congestion
* Model: DiD

* Import data
import delimited "./data/travel-time-variability-data.csv", clear
gen period_treated = period * treated

* Travel time variability models
* 2009 vs 2005
regress var_t period treated period_treated if year == 2005 | year == 2009
* 2012 vs 2005
regress var_t period treated period_treated if year == 2005 | year == 2012
* 2009 vs 2005
regress cgi period treated period_treated if year == 2005 | year == 2009
* 2012 vs 2005
regress cgi period treated period_treated if year == 2005 | year == 2012

* Clean up
capture log close
set more on
## Appendix E: Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2SLS</td>
<td>two stage least squares</td>
</tr>
<tr>
<td>AADT</td>
<td>annual average daily traffic</td>
</tr>
<tr>
<td>ALPURT</td>
<td>Albany-Puhoi realignment</td>
</tr>
<tr>
<td>APO</td>
<td>average potential outcomes</td>
</tr>
<tr>
<td>AR</td>
<td>autoregressive</td>
</tr>
<tr>
<td>ARIMA</td>
<td>autoregressive integrated moving average</td>
</tr>
<tr>
<td>AT</td>
<td>Auckland Transport</td>
</tr>
<tr>
<td>ATE</td>
<td>average treatment effect</td>
</tr>
<tr>
<td>ATET or ATT</td>
<td>average treatment effect on the treated units</td>
</tr>
<tr>
<td>AT HOP</td>
<td>Auckland Transport’s electronic payment card for public transport services</td>
</tr>
<tr>
<td>BLUE</td>
<td>best linear unbiased estimator</td>
</tr>
<tr>
<td>BTRE</td>
<td>Bureau of Infrastructure, Transport and Regional Economics</td>
</tr>
<tr>
<td>CAU</td>
<td>census area unit</td>
</tr>
<tr>
<td>CBA</td>
<td>cost-benefit analysis</td>
</tr>
<tr>
<td>CBD</td>
<td>central business district</td>
</tr>
<tr>
<td>CGI</td>
<td>congestion index</td>
</tr>
<tr>
<td>DiD</td>
<td>difference-in-differences</td>
</tr>
<tr>
<td>EB</td>
<td>Empirical Bayes</td>
</tr>
<tr>
<td>EEM</td>
<td>Economic evaluation manual</td>
</tr>
<tr>
<td>EPRR</td>
<td>enhanced post-implementation review</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>FB</td>
<td>full Bayes</td>
</tr>
<tr>
<td>GLM</td>
<td>generalised linear models</td>
</tr>
<tr>
<td>GLMM</td>
<td>generalised linear mixed models</td>
</tr>
<tr>
<td>GPS</td>
<td>generalised propensity score</td>
</tr>
<tr>
<td>HHC</td>
<td>Hamilton City Council</td>
</tr>
<tr>
<td>ITS</td>
<td>interrupted time series</td>
</tr>
<tr>
<td>ITSA</td>
<td>interrupted time series analysis</td>
</tr>
<tr>
<td>IV</td>
<td>instrumental variables</td>
</tr>
<tr>
<td>KSI</td>
<td>killed or seriously injured</td>
</tr>
<tr>
<td>LCC</td>
<td>London congestion charges</td>
</tr>
<tr>
<td>LEZ</td>
<td>low emission zone</td>
</tr>
<tr>
<td>LNMS</td>
<td>local network management schemes</td>
</tr>
<tr>
<td>LOTI</td>
<td>Loi d’Orientation des Transports Intérieurs</td>
</tr>
<tr>
<td>ML</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>NEX</td>
<td>northern express</td>
</tr>
<tr>
<td>NGTSN</td>
<td>National guidelines for transport system management</td>
</tr>
<tr>
<td>NNM</td>
<td>nearest neighbour matching</td>
</tr>
</tbody>
</table>
Ex-post evaluation of transport interventions using causal inference methods

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPRA</td>
<td>Norwegian Public Roads Administration</td>
</tr>
<tr>
<td>O-D</td>
<td>over-dimension vehicles</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>OLS</td>
<td>ordinary least squares</td>
</tr>
<tr>
<td>PIR</td>
<td>post-implementation review</td>
</tr>
<tr>
<td>POPE</td>
<td>post-opening project evaluation</td>
</tr>
<tr>
<td>PSM</td>
<td>propensity score matching</td>
</tr>
<tr>
<td>PT</td>
<td>public transport</td>
</tr>
<tr>
<td>RA</td>
<td>regression adjustment</td>
</tr>
<tr>
<td>RD</td>
<td>regression discontinuity</td>
</tr>
<tr>
<td>RTM</td>
<td>regression to the mean</td>
</tr>
<tr>
<td>SH</td>
<td>state highway</td>
</tr>
<tr>
<td>SPF</td>
<td>safety performance function</td>
</tr>
<tr>
<td>SUTVA</td>
<td>stable unit treatment value assumption</td>
</tr>
<tr>
<td>TTV</td>
<td>travel time variability</td>
</tr>
<tr>
<td>VKT</td>
<td>vehicle kilometres travelled</td>
</tr>
</tbody>
</table>