Abstract

This thesis involves three essays that explore the theory and application of econometric analysis to labour market interventions. One essay is methodological, and two essays are applications. The first essay contributes to the literature on inference with data sets containing within-cluster correlation. The essay highlights a problem with current practices when the number of clusters is 11 or fewer. Current practices can result in p-values that are not point identified but are instead p-value intervals. The chapter provides Monte Carlo evidence to support a proposed solution to this problem. The second essay analyzes a labour market intervention within Canada—the Youth Hires program—which aimed to reduce youth unemployment. We find evidence that the program was able to increase employment among the targeted group. However, the impacts are only present for males, and we find evidence of displacement effects amongst the non-targeted group. The third essay examines a set of Graduate Retention Programs that several Canadian provinces offer. These programs are aimed at mitigating future skill shortages. Once the solution proposed in the first essay is applied, I find little evidence of the effectiveness of these programs in attracting or retaining recent graduates.
Co-authorship

Chapter 3 of this thesis was co-authored with Professor Arthur Sweetman at McMaster University, and Assistant Professor Casey Warman at Dalhousie University.
To my parents for all their love and support.
I would foremost like to thank my supervisors Steven Lehrer and James G. MacKinnon for their continued support and patience throughout the writing of this thesis. This thesis would not exist without their tireless efforts. I am also indebted to Susumu Imai, Arthur Sweetman, and Casey Warman for endless encouragement and inspiration over the course of my studies.

I am extremely grateful for the stimulating and nurturing environment that exists within the Queen’s Economics Department. Thanks are due to conference participants and discussants at several conferences, most notably the Canadian Economics Association Annual Meetings in 2011 and 2012 and the Canadian Econometric Study Group. I am very appreciative of my fellow doctoral colleagues for many captivating conversations over the years. Special thanks are due to Patrick Alexander, Andrea Craig, Steven Kivinen, Afrasiab Mirza, Karl Skogstad, and Derek Stacey.

I would like to thank my extended family for their lifelong support. I am grateful to my parents and sister for their love and encouragement.

Finally, I’d like to thank my friends (especially Chris Coholan, Dana Knarr, and Adam Pachnik) for their continued willingness to put up with me. Their eagerness to engage in pursuits both intellectual and otherwise have made the doctoral process much more pleasant.
# Table of Contents

Abstract i

Co-authorship ii

Acknowledgments iv

Table of Contents v

List of Tables vii

List of Figures ix

Chapter 1: Introduction ........................................... 1

Chapter 2: Reworking Wild Bootstrap Based Inference for Clustered Errors ........................................... 5

2.1 Introduction .................................................... 6

2.2 Background on Methods to Deal With Within Cluster Correlation ..................................................... 9

2.3 Alternative Bootstrap Methods ................................. 17

2.4 Monte Carlo Evidence .......................................... 22

2.5 Conclusion ..................................................... 26
List of Tables

2.1 Design of Monte Carlo Simulations ............................. 31
2.2 Results from Monte Carlo Study with Different Numbers of Clusters:
   Replicating Results in Cameron, Gelbach, and Miller (2008) ..... 31
2.3 Results from Monte Carlo Study with Different Numbers of Clusters:
   Small Number of Clusters Simulation ............................. 32
2.4 Results from Monte Carlo Study with Different Numbers of Clusters:
   Larger Number of Clusters Simulation ............................. 32

3.2 Difference-in-Differences Employment Regressions .................. 60
3.3 Difference-in-Differences Estimates for 18-24 Year Olds, 25-30 as Comparison Group ........................................ 61
3.4 Difference-in-Differences Coefficient Estimates with 31-35 Year Olds as the Comparison Group ................................ 62
3.5 Falsification Test Using Data From 2002-2005 .................... 63

4.1 Summary of the various Graduate Retention Programs ............ 95
4.2 Variable Means from All Data Sets ............................... 96
4.3 Comparison of P-values for Full Sample Estimates ................ 97
4.4 LFS and SLID Estimates for the Atlantic Provinces ............... 98
List of Figures

2.1 Estimated Differences From Three Different P-values . . . . . . . . . . . 28
2.2 Histogram of 50,000 Monte Carlo P-values: Rademacher Distribution 29
2.3 Histogram of 50,000 Monte Carlo P-values: Enumerated Wild Bootstrap 29
2.4 Histogram of 50,000 Monte Carlo P-values: 6-point Distribution . . . 30

3.1 Weeks Employed by Age Group . . . . . . . . . . . . . . . . . . . . . 64
3.2 Weeks Not in the Labour Force by Age Group . . . . . . . . . . . . . 65
3.3 Employment Rate by Age Group . . . . . . . . . . . . . . . . . . . . . 66
Chapter 1

Introduction

While taxes and regulations may be set in an attempt to optimize the labour market, the status quo sometimes leaves something to be desired. Governments often intervene in labour markets to achieve desired ends. Broad goals usually involve broad instruments, while narrow targets usually call for more precise instruments. Econometric analysis of these interventions requires careful consideration as standard econometric tools are likely to be inappropriate. When these interventions occur within Canada, the analysis is further complicated by the small number of provinces. This dissertation analyzes two such targeted labour market interventions. This thesis also contributes to the literature on inference with data sets containing within-cluster correlation. In both cases, the implications of the analysis hinge on proper inference: ignoring the within-cluster correlation or the limitations of current techniques results in spurious estimates of statistical significance.

Chapter Two extends the literature on inference with clustered errors. The presence of clustered errors results in many problems for inference, especially prevalent when analyzing the effectiveness of public policy. Labour market interventions are
often evaluated using a difference-in-differences (DiD) methodology. Although estimation is straightforward, inference is more difficult as traditional OLS, robust, and cluster-robust standard errors are often too small when data sets contain few clusters. Following the work of Bertrand, Duflo and Mullainathan (2004) this problem has received considerable attention. A common method for overcoming this problem is the use of the wild cluster bootstrap proposed by Cameron, Gelbach and Miller (2008). This chapter highlights a problem with this technique when there are very few clusters, specifically fewer than 11. The problem arises since p-values are not point identified. It occurs as a result of there being too few unique bootstrap t-statistics. Several solutions to overcome this problem are proposed, and Monte Carlo evidence is presented which supports a six-point bootstrap weight distribution as the preferred solution. This new distribution allows for more reliable inference when using DiD in the case of few clusters.

Chapter Three examines a labour market intervention that was aimed at reducing youth unemployment in Canada in 1999 and 2000. In those years the federal government rebated 100% of employment insurance premiums that employers paid on insurable earnings for employees aged 18 to 24. The premium rate in this time period was just under 4%, so the Youth Hires program effectively offered a subsidy on workers in the targeted age range. Employers could receive this subsidy by increasing the number of youths employed, their hours worked, or their wages. A difference in differences methodology is used to compare the impacts of the targeted group against the impacts of those slightly outside the subsidy window. Analysis from two datasets shows economically and statistically significant employment impacts for the targeted
groups. The increase in employment leaves unemployment unchanged as the pro-
grams appear to lead youths to join the labour force. Interestingly, these effects are
only found for men; no program impacts were found for women. We also find evidence
of displacement, as those just too old for the subsidy were negatively impacted by
the program.

Chapter Four examines another labour market intervention aimed at youth. The
programs being analyzed are a set of Graduate Retention tax credits currently offered
by four Canadian provinces. These programs are aimed at reducing future skill short-
ages and offer generous financial incentives for recent college and university graduates
to reside in province. The policies change the relative costs of both obtaining an edu-
cation and residing in different provinces. The chapter presents the first econometric
analysis of the effectiveness of these programs and uses a difference in differences ap-
proach to examine the effects in the Atlantic provinces. After analyzing several data
sets, I find no compelling evidence that the programs have achieved their primary
goal of attracting or retaining recent graduates. I do find slight evidence that the
programs discourage current post secondary students from dropping out. The lack of
effectiveness exists despite the programs having considerable costs, suggesting consid-
erable uptake. The result of this analysis hinges on the procedure for inference, and
this chapter also presents the first application of the bootstrap weight distribution
proposed in chapter two.

Econometric analysis of labour market interventions is possible, but careful con-
sideration needs to be taken in conducting inference. Labour markets are comprised
of many individual actors, so analysis of interventions is best done with individual
level data. When data sets contain large numbers of individuals sorted into groups,
standard practices of analysis can be inappropriate. Appropriate analysis is easily obtained, though it is computationally expensive.

Labour market interventions can be effective, but the details of the interventionist policy need to be selected diligently. This thesis examines two interventions, one which appears to be successful and one which appears to be unsuccessful. Neither intervention is ideally designed. The Youth Hires program subsidizes individuals for whom there are close substitutes in the labour market. This feature results in displacement effects which an ideal intervention would avoid. The Youth Hires subsidy was only available for marginal insurable earnings, aligning the benefits with the goal. The Graduate Retention Programs appear ineffective. Despite their ineffectiveness, the various credits are claimed by many people each year, suggesting that the programs are not targeting the marginal person. Ideal interventionist policies are beyond the scope of this dissertation, though careful econometric analysis of interventions should aid in our understanding of ideal policies.
Chapter 2

Reworking Wild Bootstrap Based Inference for Clustered Errors

Many empirical projects are well suited to incorporating a linear difference-in-differences research design. While estimation is straightforward, reliable inference can be a challenge. Past research has not only demonstrated that estimated standard errors are biased dramatically downwards in models possessing a group clustered design, but has also suggested a number of bootstrap-based improvements to the inference procedure. In this chapter, I first demonstrate using Monte Carlo experiments, that these bootstrap-based procedures and traditional cluster-robust standard errors perform poorly in situations with fewer than eleven clusters - a setting faced in many empirical applications. I subsequently introduce two easy-to-implement alternative procedures that involve the wild bootstrap. Further Monte Carlo simulations provide evidence that the use of a 6-point distribution with the wild bootstrap can improve the reliability of inference.
2.1 Introduction

Difference-in-differences (DiD) estimators have a great deal of appeal, as there are often policy changes that affect a subset of the population. The presence of two groups allows us to make inferences about the causal effects of a policy change. The appropriateness of DiD estimators depends on a few critical assumptions being satisfied, beyond having a treatment and a control group for the estimates. The first assumption is that there is common support amongst the two groups. Common support requires that the composition of both groups—in terms of both observable and unobservable characteristics—be similar. The second assumption of common or parallel trends requires that each of the groups had a similar trend in the dependent variable before the policy change.\(^1\) The common trend assumption implies that in the absence of the change, both the treatment and control groups would have followed along the same trends as before the change. Any change that is observed in the treatment group, differenced by the change in the control group, is then attributed to the policy change. For more details on program evaluation see DiNardo and Lee (2011).\(^2\)

Beyond worrying about the identifying assumptions for DiD, recent research has asked the question: how reliable are the inferences made with DiD? In answering this question the literature has shown that a fundamental problem with difference-in-differences arises from the use of data with clustered errors. Since ignoring clustered errors leads to very unreliable inference, corrections for clustered errors have become

\(^1\)Abadie (2005) and Athey and Imbens (2006) relax this assumption.

\(^2\)Arguably, the most well-known application of DiD estimators is Card and Krueger (1994), which examined the impact of increasing the minimum wage on employment in the fast food industry. Other well-cited DiD applications have involved analyzing changes in tax laws on health insurance Gruber and Poterba (1994) and changes to the Earned Income Tax Credit on labour supply Eissa and Liebman (1996). Overviews of difference-in-differences estimators are provided in Meyer (1995) and Angrist and Pischke (2008).
commonplace in empirical work. Although asymptotic corrections work well in many cases, recent studies have suggested that estimating cluster-robust variance-estimator (CRVE) standard errors leads to biased inference when the number of clusters is small.

Donald and Lang (2007) first showed the unreliability of DiD estimators in the case when there are few groups and when some variables are fixed within groups. The authors also used Monte Carlo work that estimated rejection frequencies are too high. The authors propose a two-step method to estimate the significance of DiD coefficients. Bertrand, Duflo and Mullainathan (2004) (BDM) show that in Monte Carlo simulations to show that DiD coefficients are estimated to be significant at the 5% level, 45% of the time. They suggest that the over-rejection is largely driven by the serial correlation in their data. The authors propose a number of methods to correct this problem, including a block bootstrap procedure. Conley and Taber (2011) argue that point estimates within DiD frameworks are not consistent because variables for policy interventions are often invariant over time for a given group. They propose a method of inference which relies on information contained in the control groups. Finally, Cameron, Gelbach and Miller (2008) (CGM) propose a wild bootstrap-based procedure extending the work of BDM.

Empirical researchers, following CGM, have frequently used wild cluster bootstrap-generated p-values for improved inference. However, this chapter demonstrates that when the number of clusters is quite small the procedure for inference is noisy and imprecise as estimated ‘p-values’ are intervals rather point estimates. As a result the standard 2-point wild cluster bootstrap is not appropriate when there are few clusters, with the technique becoming less appropriate when there are fewer than eleven

---

3A seminal paper on estimating clustered errors, Rogers (1994), has over 1700 citations according to Google Scholar as of June 2013.
4As of June 2013 this article has been cited over 417 times according to Google Scholar.
clusters. There are many real world problems where data sets contain eleven or fewer clusters. For example, policy analysts in Canada often exploit variation across ten provinces, while policy analysts in Australia often examine eight states. Alternatively, clustering is often performed in the time dimension, and it is common to have fewer than eleven time periods in panel data. This is particularly true in finance, following the suggestion of Thompson (2011) that if in a data set the number of firms greatly exceeds the number of time periods, then much of the bias is eliminated through clustering by year.

This chapter proposes two procedures when the sample is collected from a small number of clusters, considering both enumerating the bootstrap samples and new bootstrap weight distributions. Enumeration involves systematically calculating all of the possible bootstrap samples, and their associated t-statistics. The enumeration procedure has the benefit of being invariant to resampling variance, but it is limited in the precision of the calculated p-values when the number of clusters is quite small. Expanding the 2-point wild cluster bootstrap to either a 4-point or a 6-point distribution allows for an approximate test for significance. The 4-point and 6-point wild cluster bootstraps have resampling variability, but more precise p-values can be determined. The proposed distributions appear to work well even in the case of five clusters, when the conventional 2-point wild cluster bootstrap is most inappropriate.

The organization of this chapter is as follows: Section 2.2 provides a background on the challenge of clustered errors in empirical research and current strategies to deal with them. Specifically, the limitations of the 2-point wild cluster bootstrap are identified and examined. Alternative bootstrap methods to account for the small cluster problem are discussed in section 2.3, with an enumeration technique and the
aforementioned new bootstrap weight distributions considered. Section 2.4 discusses the design and results of Monte Carlo simulations. The results expose the limitations of existing techniques when properly calculated, and favor a new 6-point distribution. Section 2.5 concludes.

2.2 Background on Methods to Deal With Within Cluster Correlation

Consider a standard two-period linear difference-in-differences model such as:

\[ Y_{igt} = \beta_0 + \beta_1 \ast treat_g + \beta_2 \ast post_t + \beta_3 \ast treat_g \ast post_t + X_{igt} \gamma + u_{igt}. \]  (2.1)

Here \( Y_{igt} \) is an observation for person \( i \) in group \( g \) and time \( t \), \( treat_g \) is a dummy variable for whether the observation is in the treatment group, and \( post_t \) is a dummy variable for whether the observation is in time period after the treatment occurred. Neither group was treated in the pre-period. The \( treat_g \ast post_t \) variable is an interaction of the two indicator variables. It is an indicator for those individuals in the treated group in the treated period. The coefficient on this term can be interpreted as the difference-in-differences estimate, which can be viewed as a causal parameter. \( X_{igt} \) is a vector of other independent variables. It is quite easy to extend this setup to multiple periods.

Models like (2.1) are quite common in empirical work, though many papers have shown a problem with using conventional OLS or heteroskedasticity-consistent standard errors for inference when data are of a grouped or clustered nature. A data
set can be considered clustered when there is an underlying natural grouping of the observations. Sometimes these groups are based on methodology, as in data on many students in several classrooms within a particular school. More often the grouping is geographic as there are data on many individuals residing within a given state. The problem is most severe when estimating the impact of a common group variable, such as a treatment variable, on individual level outcomes. The first paper that identified this problem is Kloek (1981), though the problem was popularized by Moulton (1990) and Rogers (1994). The problem was considered in the DiD context by Bertrand, Duflo and Mullainathan (2004) as well as Donald and Lang (2007). For a detailed survey of the issues related to clustered data see Cameron and Miller (2010).\footnote{Another paper that deals with some of the issues of clustered data is Ibragimov and Muller (2010). However, their paper proposes an estimation technique which requires estimating a t-statistic for each cluster separately. While this technique works well in many situations, it does not work at all when there is a binary independent variable which is invariant within a cluster. As this is often the case with DiD estimates, no Monte Carlo simulations testing their technique will be performed in this chapter.}

The estimates of the $\beta$ coefficients are unaffected by the clustered nature of the data and can be obtained using the OLS estimator. The issue with clustered data is that the estimated error terms, $\hat{u}_{igt}$, can no longer be assumed to be i.i.d. Although the errors are independently distributed across clusters, the errors are correlated within clusters. Expressed formally, clustered data results in $E[u_g] = 0$, $E[u_g'u'_g] = \Sigma_g$, $E[u_g'u'_h] = 0$ for cluster $h \neq g$. Given that the i.i.d. assumption is violated, the standard OLS variance matrix is an inappropriate estimate of the variance. The Cluster Robust Variance Estimator (CRVE) was developed in response to the need to correct for within cluster correlation. The standard Cluster Robust Variance Estimate
The CRVE estimate takes a familiar sandwich form, though the \( \hat{u}_g \) terms can be non-standard and need to be estimated from the data. In the simplest case, the OLS residuals are used with \( \hat{u}_g = y_g - X_g \hat{\beta} \). In other cases, the expression in equation (2.2), \( \sum_{g=1}^{G} X_g \hat{u}_g \hat{u}_g' X_g' \), is replaced by \( \sum_{g=1}^{G} \tilde{U}_g \tilde{U}_g' \). Software packages have different routines for estimating \( \tilde{U}_g \). For example, Stata uses:

\[
\tilde{U}_g = \sum_{i=1}^{N_g} \hat{u}_{ig} \begin{pmatrix} 1 \\ X_g \end{pmatrix},
\]

where \( \hat{u}_{ig} \) is the OLS residual for individual i in group g. CRVE controls for both error heteroskedasticity and quite general correlation and heteroskedasticity within clusters.

White (1984) shows the consistency of this estimator based on three assumptions:

A1. The number of clusters, G, goes to infinity.

A2. The degree of within-cluster correlation is constant across clusters.

A3. Each cluster contains an equal number of observations. Several authors have previously studied the finite sample properties of the estimator when A1 is not satisfied. Carter, Schnepele and Steigerwald (2012) relax both assumptions A2 and A3 and derive a new asymptotic distribution for the test statistic. MacKinnon and Webb (2013) study the finite sample properties when A3 is violated. Simulation results from Bertrand, Duflo and Mullainathan (BDM), Cameron, Gelbach and Miller (CGM), and
those presented in this chapter, show that the rejection rates based on OLS standard errors are almost an order of magnitude greater than those based on CRVE. In my own simulations with 30 clusters (discussed at length later in the chapter and shown in table 2.2), the estimated 5% rejection rate for OLS is 49.9% whereas it is only 8% with CRVE standard errors. In simulations with 5-clusters, the rejection rate is 47% for OLS and 21% for CRVE.\footnote{On a historical note, it was Moulton (1990) that pointed out just how large the rejection rates were for OLS, and BDM who pointed out that CRVE is still a great distance from the desired size of 5% when there are few clusters.}

2.2.1 Should we use the Wild Bootstrap to Conduct Inference in DiD Models?

Although CRVE is a substantial improvement over OLS in the presence of grouped data, it is not without its weaknesses. If one uses CRVE with $\hat{U}_g = \hat{u}_g$ it is biased, as $E[\hat{U}_g \hat{U}_g^\prime] \neq \Sigma_g = E[u_g u_g^\prime]$. The bias depends on the form of $\Sigma_g$ but will usually be downward, which results in coefficients being estimated as significant too often. It is important to note that while the CRVE estimates of standard errors are biased, they are less biased than the OLS standard errors.

After presenting Monte Carlo evidence of the over-rejection problem when using standard CRVE techniques, BDM propose a bootstrap procedure as a means of improving the size of the tests. In particular BDM suggest block bootstrapping, where they resample blocks of all observations from a given state. Cameron, Gelbach and Miller (2008) perform additional Monte Carlo experiments and find that when the number of clusters is small, e.g. fewer than 30, the rejection rate of the block
bootstrap method proposed by BDM is too large.\textsuperscript{7} CGM investigate several alternative bootstrap methods for improved inference and argue that the ‘Wild Cluster bootstrap-t’ method is preferred.

The wild cluster bootstrap is preferable to the block bootstrap in several ways. Each bootstrap sample has the same number of observations, equal to the original sample size, while the block bootstrap generates samples of unequal size. Additionally, every observation in the data set is in every bootstrap sample. This is an important characteristic when identification may be coming from only a few observations, such as when a certain policy is operating in a particular state for only a few years, as pointed out by Conley and Taber (2011). Finally, the wild bootstrap preserves the structure of the error correlation within clusters. The structure is preserved as every residual within a cluster is multiplied by the same weight. CGM present Monte Carlo simulations as evidence that the wild cluster bootstrap-t technique allows for valid inference with as few as five clusters. As this chapter will discuss, unfortunately, there is a problem with the rejection rates they calculate to justify that claim. This problem is a result of an insufficient number of unique bootstrap samples.

Imagine we are interested in calculating a wild cluster bootstrap-t p-value for $\beta_3$ in equation (2.1). We can construct the p-value by first estimating the t-statistic, $\tilde{t}$, in the original sample using cluster-robust standard errors. We then re-estimate the equation by imposing the null hypothesis, to obtain the restricted estimates $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{u}_{igt}$, etc. Then $B$ iterations, or bootstraps, are performed. In each iteration a bootstrap

\textsuperscript{7}The over rejection is a result of the BDM technique using OLS standard errors in generating bootstrap t-statistics, rather than CRVE standard errors, and a result of less-than-desirable features of the block bootstrap.
sample is generated from the bootstrap data generating process

\[ y^*_{igt} = \tilde{\beta}_0 + \tilde{\beta}_1^{* \cdot \text{treat}_g} + \tilde{\beta}_2^{* \cdot \text{post}_t} + \tilde{\beta}_3^{* \cdot \text{treat}_g \cdot \text{post}_t} + X_{igt} \tilde{\gamma} + \tilde{u}_{igt} v_g, \]  

(2.3)

where the \( i^{th} \) residual in time \( t \) from group \( g \), \( \tilde{u}_{igt} \), is transformed by the bootstrap weight \( v_g \). The difference between the wild cluster bootstrap and the conventional wild bootstrap is that under the former the same \( v_g \) is applied to all observations within the same cluster, while the conventional wild bootstrap applies a \( v_{igt} \) to each observation. The bootstrap weight can take many forms, as will be discussed later. In each iteration, a bootstrap t-statistic, \( t^*_j \) is generated using cluster-robust standard errors. After \( B \) iterations the bootstrap p-value is then calculated by:

\[ \hat{p}^*(\hat{t}) = 2 \min \left( \frac{1}{B} \sum_{j=1}^{B} I(t^*_j \leq \hat{t}), \frac{1}{B} \sum_{j=1}^{B} I(t^*_j > \hat{t}) \right), \]  

(2.4)

where \( I(.) \) is the standard indicator function.\(^9\)

This procedure is based on the assumption that a given number of bootstrap samples, \( B \), are taken from an extremely large pool of potential bootstrap samples. This means that a set of bootstrap samples are drawn that will contain very few, if any, repeated samples. Suppose we are concerned about the significance of our estimated \( \hat{\beta} \) by examining our t-statistic, \( \hat{t} \), and we have generated a vector of 999 bootstrap t-statistics, \( t^* = t^*_1, ..., t^*_{999} \). If we observe that our estimated t-statistic falls between the 90th and 91st bootstrap t-statistic, then we would say that the p-value

\(^8\)In general the bootstrap DGP should impose the null hypothesis, which in this case would mean setting \( \tilde{\beta}_3 = 0 \).

\(^9\)These p-values are equal tail p-values, while the enumeration p-values are symmetric p-values calculated by \( \frac{1}{B} \sum_{j=1}^{B} I(|t^*_j| > |\hat{t}|) \).
associated with this t-statistic is 0.180.\(^\text{10}\)

With few clusters the number of unique potential bootstrap samples is rather small. The bootstrap samples depend on the choice of a bootstrap weight distribution. In the literature two well-known distributions are the Rademacher and the Mammen distributions, both of which contain only two points. With these distributions, \(v_g\) from equation (2.3) is set to one of two values with a given probability, \(p\). The Rademacher distribution is defined as:

\[
v_g = \pm 1 \text{ with probability } 0.5. \quad (2.5)
\]

The Mammen distribution is defined as:

\[
v_g = -\frac{\sqrt{5} - 1}{2} \text{ w.p. } p = \frac{\sqrt{5} + 1}{2\sqrt{5}} \quad \text{and} \quad v_g = \frac{\sqrt{5} + 1}{2} \text{ w.p. } 1 - p.
\]

Accordingly, there are only \(2^G\) possible bootstrap samples, where \(G\) is the number of groups. When \(G = 5\) there are only 32 possible bootstrap samples. Cameron, Gelbach and Miller (CGM) recommend using the wild cluster bootstrap-t technique with Rademacher weights. Thus the 32 bootstrap samples yield 32 distinct t-statistics. However, the set of unique absolute value t-statistics is only \(2^{G-1}\) or 16 in the five cluster case. A proof of this result is provided in Appendix A. When \(G = 5\) and \(B = 399\), by sampling with replacement you are choosing 399 elements from a set of 32. This is not a problem when \(G\) is large as you will obtain a vector of mostly unique t-statistics, but when \(G\) is small it is quite problematic.

The CGM procedure inaccurately treats the \(B\) t-statistics as \(B\) unique values;

\(^{10}\)Recall that the equal tail p-value is the result of a two-sided test for t-statistics, so \(\hat{p}^* = 2\min\left(\frac{90}{999}, 1 - \frac{90}{999}\right)\).
however, in the small cluster case, the majority of bootstrap t-statistics are not unique. Having many repeated t-statistics leaves open the possibility that \( \hat{t} \) will be found multiple times in this vector.\textsuperscript{11} We should thus regard unique t-statistics as a signal as to the significance of \( \hat{\beta} \), and repeated t-statistics as noise which interferes with our ability to make inferences about \( \hat{\beta} \). When \( 2^G \) is small we cannot perform conventional inference. This problem comes as a result of the inability to point-identify where \( \hat{t} \) falls within the sorted vector of bootstrap t-statistics. If \( \hat{t} \) is found multiple times within the vector, then the ‘p-value’ would not be a point but would instead be an interval from the first occurrence of \( \hat{t} \) to the last occurrence of \( \hat{t} \). Returning to the above example, if we have \( B = 999 \) and 31 of those bootstrap samples result in \( t^* = \hat{t} \) then \( \hat{t} \) would appear in the sorted vector 31 times. For example, if \( \hat{t} = t^*_{70}, ..., t^*_{100} \), then the ‘p-value’ would be the interval from 0.140-0.200. Figure 2.1 plots the observed spread between the first occurrence p-value and the last occurrence p-value across clusters from Monte Carlo simulations using the Rademacher distribution with 999 bootstraps. The figure shows that the p-values occupy a wide interval when the number of clusters is small. This wide interval makes it quite difficult to assess significance at conventional levels. It is not until there are more than eleven clusters that these intervals are quite small.

Calculating additional bootstrap t-statistics will not shrink the width of these intervals. Calculating an infinite number of bootstraps will lead to results equivalent to those of the enumeration p-values, discussed in section 2.3. One way to generate more unique t-statistics is to increase the number of clusters, though in empirical work the number of clusters will be determined by the data.

When using the Rademacher distribution, one of the possible bootstrap t-statistics,\textsuperscript{11}There will also likely be repetitions of other t-statistics in the vector of bootstrap t-statistics.
$t_j^*$ is the original estimate of the t-statistic, $\hat{t}$.\textsuperscript{12} When $2^G$ is small, this will be observed, almost surely. As there are only $2^G$ possible t-statistics, estimating the p-value depends on identifying where $\hat{t}$ lies in the vector of sorted t-statistics. The $2^G$ possible t-statistics map into a small number of p-values. If inference is done correctly, we should only observe that small number of p-values across Monte Carlo simulations. CGM instead chose to estimate the p-value as being the center of this range. Figure 2.2 shows a histogram of 50,000 p-values based on the CGM method for 2-point wild cluster bootstrap-t. If the CGM procedure worked appropriately, only 16 unique p-values would be observed.\textsuperscript{13} However, the histogram is quite smooth and shows that numerous p-values were calculated. These additional p-values are a result of the noise from their estimates.\textsuperscript{14} The noise leads to improper inference and makes the 2-point wild bootstrap inappropriate in cases with few clusters.

### 2.3 Alternative Bootstrap Methods

The first technique considered in this chapter for improving inference is that of enumeration. The above mentioned issues are a result of using a 2-point distribution in general, and the Rademacher distribution in particular. Inference using a 2-point distribution will be limited in the few cluster case on account of the interval identified p-values. However, if one is convinced that the Rademacher distribution is ideal, then enumeration is the correct way to conduct analysis.\textsuperscript{15} The procedure for estimating a

\textsuperscript{12}This occurs when $v_g = 1, \forall g$.

\textsuperscript{13}The fact there are 16 p-values is a result of $\hat{t} = t_2^*$ and $\hat{t} = t_{30}^*$ resulting in the same p-value in a two tail test.

\textsuperscript{14}The noise comes in part from repeated $\hat{t}$, but also from repeated $t_1^*, t_2^*, \text{etc.}$ The number of $\hat{t}$ repetitions changes from one replication to another, thus the variability of p-values in their technique.

\textsuperscript{15}This procedure was alluded to in Efron’s seminal bootstrap paper in 1979 and mentioned in Davidson and Flachaire (2008) specifically in the context of the (non-cluster) wild bootstrap.
p-value is quite similar to the wild cluster bootstrap procedure discussed above. The main difference is that with the wild bootstrap $v_g$ is picked at random from the distribution, while with enumeration $v_g$ is selected methodically. Given the small number of possible bootstrap samples when the number of clusters is small, it is feasible to calculate all possible t-statistics. When all possible test statistics are calculated, it is referred to as full enumeration; when a subset of these test statistics is calculated, it is referred to as partial enumeration.

Under full enumeration, the resulting p-values do not depend on resampling variation. In conventional bootstrap procedures the results will depend in part on the set of samples drawn, and thus are subject to resampling variation. This is not the case with full enumeration as all samples have been drawn. When the number of clusters is large, it is infeasible to calculate all possible t-statistics; however, partial enumeration is possible and will result in a sample of bootstrap t-statistics without any repetitions. The main benefit of enumeration is that you get a sample of t-statistics with no repetition, though there can be a small benefit in terms of computing time.

The resulting p-value of this procedure is different than a conventional p-value. For instance when $G = 5$ there are only $2^{G-1}$ unique t-statistics in absolute value. If $|\hat{t}| = |t_{*2}|$, the p-value is equal to $\frac{2}{16}$, which tells us something about the statistical significance of $\hat{\beta}$. We have to be careful not to think about this p-value as 0.125, as doing so can lead one to incorrectly infer that the observed p-value is drawn from a continuous distribution. In this case the p-value is $\frac{2}{16}$, but it could have alternatively been $\frac{1}{16}$ or $\frac{3}{16}$, and is drawn from a discrete distribution with the p-value $\in \{\frac{1}{16}, \ldots, \frac{16}{16}\}$. The issue here is that conventional significance levels that applied econometricians work with are not as meaningful. The p-value of $\frac{2}{16}$ spans the space
from 0.0625 – 0.125 and so straddles the 10% level. Perhaps we are best to remain agnostic about whether this observed p-value is significant at the 10% level, thus the recommendation of reporting enumerated p-values as fractions as opposed to decimals to highlight the distinction.

Although enumeration has much to its credit, its advantages are largely confined to cases with small G. After G is sufficiently large, say 12, the computational limitations of full enumeration necessitate partial enumeration, which is very similar to the wild cluster bootstrap. Figure 2.3 shows the histogram of Monte Carlo p-values using the enumeration method for the 5-cluster case. This figure is similar to figure 2.2 though here it is easy to see that only the p-values associated with the 16 unique bootstrap t-statistics have been calculated. Using this technique results in inference being based on the data and the properties of the bootstrap weighting distribution, and not on resampling noise.

2.3.1 Adding Points to the Bootstrap Weight Distributions

Enumeration will generate unique t-statistics, and thus is more precise than the conventional wild bootstrap procedure. However, the limitation of having only $2^{G-1}$ t-statistics from which to conduct inference leaves much to be desired. It is possible to find variations to the wild bootstrap technique which expand the number of points in the weight distribution, $v_g$ in equation (2.3), used to generate bootstrap samples. Following Davidson and Flachaire (2008), who show that the Rademacher distributions has better finite sample properties than the Mammen distribution, I look for variations of the Rademacher distribution.
The first four moments of the ‘ideal’ distribution would be 0,1,1,1. It is however not possible to satisfy all of these moment conditions.\(^\text{16}\) The Rademacher and Mammen distributions differ in the moment conditions that they satisfy. Both distributions have a mean of zero and a variance of one. The Mammen distribution has a third moment equal to one, but a fourth moment of two. The Rademacher distribution has a third moment of zero and a fourth moment of one. The candidate distribution will expand the Rademacher distribution, imposing symmetry. Like the Rademacher, the candidate distributions ignore the third moment. The candidate 4-point distribution I consider is:

\[
v_g = -\sqrt{\frac{3}{2}}, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{3}{2}}, \text{ w.p. } \frac{1}{4}.
\]

(2.6)

In addition to imposing symmetry the 6-point distribution will impose a restriction that two of the points are 1 and –1. The imposition of symmetry means that the third moment will be 0. The ideal distribution will then have the 6-points of the form 

\(-A, -1, -B, B, 1, A\) each selected with equal probability. The first four moments of this symmetric 6-point distribution would have to be 0,1,0,1 to match the Rademacher moments. This also is impossible. Any symmetric equal probability distribution will automatically satisfy the first and third moment restrictions. It is then a matter of trying to satisfy the second and fourth moment conditions. Rearranging these moment conditions results in the following equation: 

\[A^2 + B^2 + 1^2 = A^4 + B^4 + 1^4.\]

This is only satisfied when A and B are 0, 1, or –1, which does not result in a 6-point distribution. Thus it is not possible to have a distribution of the form 

\(-A, -1, -B, B, 1, A\) with the first four moments of 0, 1, 0, 1. The candidate 6-point distribution I consider is:

\(^{16}\)I thank Professor James MacKinnon and Professor Russell Davidson for bringing this to my attention.
\[ v_g = -\sqrt{\frac{3}{2}}, -\sqrt{\frac{2}{2}}, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{2}{2}}, \sqrt{\frac{3}{2}} \quad w.p. \quad \frac{1}{6}. \quad (2.7) \]

The fourth moments of these distributions are \( \frac{5}{4} \) for the 4-point, and \( \frac{7}{6} \) for the 6-point. There exists the temptation to add additional points to the distribution to increase the potential number of bootstrap samples. There are two concerns about doing so. The first is that the ideal weights will be distinct from one another, as the weights 0.99 and 1.01 will result in very similar bootstrap samples and very similar bootstrap t-statistics, \( t_j^* \). The second is that given the desire to have a distribution with distinct weights, mean zero, and variance one, the inclusion of additional weights will often increase the fourth moment. As a limiting case I consider using the normal distribution to generate weights for the bootstrap sample where \( v_g \sim N(0, 1) \). Drawing from the normal would allow for infinite possible bootstrap samples. Mammen (1993) considered the distribution \( v_g = \frac{u_i}{\sqrt{2} + \frac{1}{2}(w_i^2 - 1)} \), where \( u_i \) and \( w_i \) are draws from the normal distribution.\(^{17}\)

The main benefit of adding additional points to the bootstrap weight distribution is that the number of potential bootstrap samples increases exponentially. For instance, with the 2-point distribution the number of bootstrap samples is \( 2^G \), but with the 4-point distribution it increases to \( 4^G \), and to \( 6^G \) for the 6-point distribution. So in the case of five clusters, the number of potential bootstrap samples increases from 32 to 1024 to 7776. It should be noted that the unique number of absolute value t-statistics is less than \( 4^G \) or \( 6^G \). The proposed distributions are also symmetric, and so have the same feature as the Rademacher distribution. As a result, the number of

\(^{17}\)Mammen (1993) also considered another more complicated distribution. These two distributions are ignored in this chapter since simulation results in MacKinnon (2012) show them to be inferior to the Normal distribution.
unique absolute value t-statistics is $\frac{G}{2}$ for the 4-point and $\frac{6G}{2}$ for the 6-point. The high number of possible bootstrap samples should give us confidence that the inferences made using the 6-point distribution are based primarily on the estimated t-statistics, and not on noise introduced from resampling as is the case when using the 2-point distribution. Figure 2.4 shows a histogram of 50,000 p-values based on the 6-point wild bootstrap method. In contrast to figure 2.2, the smoothness seen in this figure comes from the great number of unique and correctly calculated p-values.

2.4 Monte Carlo Evidence

2.4.1 Description of Simulations

To enhance the comparability of the simulations, I follow the simulation procedure in section IV.A of Cameron, Gelbach and Miller (2008). Data are generated using

\[ y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \]

or

\[ y_{ig} = \beta_0 + \beta_1 (z_g + z_{ig}) + (\epsilon_g + \epsilon_{ig}). \] (2.8)

With $z_g, z_{ig}, \epsilon_g, \epsilon_{ig}$ each an independent draw from $N(0,1)$. We can think of $z_g$ as a group specific component of $x_{ig}$ and $\epsilon_g$ as the group level error. The presence of $\epsilon_g$ introduces correlation amongst the error terms. Alternatively, $z_{ig}$ is the idiosyncratic component of $x_{ig}$, while $\epsilon_{ig}$ is the idiosyncratic component of the error term.

The number of observations per group, $N_g$, is set to 30 for all simulations. I perform 50,000 replications, and each of the bootstrap exercises uses 399 bootstraps.
In generating \( y_{ig} \), I set \( \beta_1 = 1 \) and test the hypothesis that \( \beta_1 = 1 \). Following common practice, the null hypothesis is imposed in the bootstrap replications. The rejection rates are estimated across replications as

\[
\hat{\alpha} = \frac{1}{R} \sum_{j=1}^{R} I(p_j^* \leq 0.05),
\]

where \( R \) is the number of replications, and \( p_j^* \) is the bootstrap p-value from the \( j^{th} \) replication. This \( \hat{\alpha} \) is then compared to the true size of the test which is given by \( \alpha = 0.05 \).

In total eight different rejection rates are compared, across a variety of asymptotic and bootstrap methodologies. A description of the simulations can be found in table 2.1. Designs 1-3 use asymptotic procedures for generating p-values, while designs 4-8 use bootstrap procedures. Design 1 uses t-statistics which are calculated using OLS standard errors and are assumed to follow a normal distribution. As the OLS standard errors ignore the clustered nature of the data these rejection rates should be rather high, as was pointed out by Moulton (1990). Design 2 uses the CRVE standard errors of equation (2.2), and the t-statistics are assumed to be distributed normally. Design 3 uses the same t-statistics as in design 2, but the distribution of the t-statistics is assumed to follow a t-distribution with \( G-1 \) degrees of freedom, where \( G \) is the number of groups.

Designs 4-8 employ the wild cluster bootstrap-t procedure as discussed above, but differ in which bootstrap weight distribution is used. Design 4 generates p-values using

\[\text{The code I used for performing the bootstrap simulations is based off the code provided by Douglas Miller, which can be found at: http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/bs_example.do I thank the author for making his code publicly available.}\]

\[\text{This distribution is both recommended by Donald and Lang (2007) and is the default Stata uses with the cluster command.}\]
the wild cluster bootstrap with $v_g$ drawn from the 2-point Rademacher distribution described above in equation (2.5), this is the test that was recommended by CGM.\textsuperscript{20} Design 5 generates p-values with $v_g$ drawn from $N(0, 1)$. Design 6 uses $v_g$ drawn from the 4-point distribution that was proposed above in equation (2.6). Design 7 uses $v_g$ drawn from the 6-point distribution that was proposed above in equation (2.7). Finally, design 8 generates p-values by enumerating the Rademacher wild bootstrap t-statistics. When $G \leq 11$ full enumeration is used and all t-statistics are calculated. When $G > 11$ partial enumeration is used and a unique set of t-statistics is calculated. The results of the Monte Carlo experiments are discussed below.

\subsection*{2.4.2 Simulation Results}

Table 2.2 replicates Cameron, Gelbach and Miller (2008) by performing tests 1-5.\textsuperscript{21} The table shows the severe problem of ignoring the clustered nature of the data, as the test using OLS standard errors gives rejection rates of close to 50%. Clustering the standard errors and performing inference-based tests 2 and 3 works much better. Assuming that the t-statistics are normally distributed is rather problematic when there are very few clusters. The assumption that the t-statistics follow a t-distribution with $G - 1$ degrees of freedom goes a long way in correcting the size of the test, but the rejection rate is still too large when $G$ is very small. The rejection rates for the wild cluster bootstrap-t with Rademacher weights look deceivingly nice. As explained

\textsuperscript{20}The values reported are slightly different than the values reported by CGM. They are different because different random numbers were used, but more importantly because CGM use the average value at which $\hat{t}$ matches the bootstrap t-statistics, while I use the max value at which this occurs. The difference is negligible when $2^G$ is large, but significant when $2^G$ is small, see figure 2.1. The difference is largely irrelevant as neither rejection rate is correct in the small G case.

\textsuperscript{21}In all of the tables, the simulation standard error is not shown to save space. The standard error is $s_{\hat{\alpha}} = \sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{R-1}}$ with R being the number of replications, and $\hat{\alpha}$ being the estimated rejection rate. Given the observed rejection rates the largest standard error is 0.0022 and the smallest is 0.0008.
above the results for $G = 5$ and $G = 10$ should not be trusted as they are based on a very noisy vector of t-statistics, but the results for $G \geq 15$ do not suffer from this problem. A histogram of the 5-cluster wild bootstrap p-values can be seen in figure 2.2. The under-rejection in the table is evidenced by the under-concentration of p-values in the far left of the figure.

Table 2.3 shows the results of simulations in which the number of clusters is small. The wild bootstrap with Normal weights does fairly well, though it seems to be outperformed in most cases by the wild bootstrap with either the 4-point or 6-point distribution. Both the 4-point and 6-point distribution work well, though in all cases the 6-point distribution seems to outperform the 4-point distribution. Note that when $G = 5$ the rejection rate is 0.070, which is still noticeably above 0.05, but better than the $T(G - 1)$ rate of 0.100. This over-rejection can also be seen in figure 2.4, as evidenced by the over-concentration of p-values in the left tail of the histogram.

As mentioned previously, the enumerated p-values are not point identified and are instead identified by an interval. Two rejection frequencies are calculated for these p-values, one using the lower bound, and one using the upper bound. The wide differences in these two rejection frequencies are to be expected, as was shown in figure 2.1. In the 5-cluster case the upper bound never rejects at the 5% level as $\frac{1}{16}$ is above that threshold. The lower bound rejects far too often. This is particularly interesting since both of these rejection frequencies come from the same estimated t-statistic. The upper bound rejection frequency and lower bound rejection frequencies converge as $G$ increases. For the 5-cluster case, a histogram of the 16 t-statistics is shown in

---

22 The results for the wild cluster bootstrap with Rademacher weights are not presented in this table since the p-values are not correctly calculated in the 5-10 cluster range.

23 The normal distribution is not ideal since the fourth moment is too large; see MacKinnon (2012).
figure 2.3. The over-concentration of p-values in the left of the figure corresponds
with the result in the table that the enumeration technique rejects too often when
using the lower bound of the interval. Even with 10 clusters the enumeration rejection
frequencies are higher than those from the 6-point distribution.

Table 2.4 shows the results when the number of clusters ranges from 5 to 30. The
partial enumeration method which is used for \( G \geq 15 \) works well in the case
of \( G=15 \) and \( G=20 \). The results are not presented for \( G \geq 25 \), since they should
be equivalent to the results from using the 2-point distribution. The 4-point and
6-point distributions seem to work equally well, though in a few cases the 6-point
distribution outperforms the 4-point distribution. In general, the various bootstrap
methods work better than the analytic \( T(G - 1) \) method, test 3. While the 6-point
distribution dominates the 2-point distribution in most cases, the 2-point distribution
seems to do slightly better when \( G \) is equal to 15 or 20, though it slightly under-rejects
compared to the 6-point bootstrap when \( G \) is larger. Given the small range in which
the 6-point distribution is inferior to the 2-point distribution, and the problems with
the 2-point distribution in the case of few clusters, the 6-point distribution is generally
preferable.

\section{2.5 Conclusion}

While difference-in-differences estimators are widely used to evaluate policy changes,
care must be taken in performing inference. This is particularly true when individual-
level data are used, and when the data are grouped or clustered. I evaluate the
performance of several inference procedures in Monte Carlo simulations and confirm
the findings of Cameron, Gelbach and Miller (2008). In this chapter I show that
substantial improvements can be made to the inference procedure when the researcher faces few clusters. The few cluster concern is quite common in practice, as many data sets have fewer than eleven clusters. The issue with the conventional wild bootstrap procedure is that it uses a 2-point weight distribution. The small number of weights leads to p-values not being point identified when there are few clusters. Using a 6-point weight distribution solves this problem and works equally well when there are more clusters.
Figure 2.1: Estimated Differences From Three Different P-values

Notes: A is the difference between the max p-value and the mean p-value. B is the difference between the max p-value and the min p-value.
Figure 2.2: Histogram of 50,000 Monte Carlo P-values: Rademacher Distribution

Figure 2.3: Histogram of 50,000 Monte Carlo P-values: Enumerated Wild Bootstrap
Figure 2.4: Histogram of 50,000 Monte Carlo P-values: 6-point Distribution
Table 2.1: Design of Monte Carlo Simulations

<table>
<thead>
<tr>
<th>#</th>
<th>Design</th>
<th>Standard Error</th>
<th>t-statistic distributed as</th>
<th>Bootstrap Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OLS</td>
<td>OLS</td>
<td>N(0,1)</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Cluster ∼ N</td>
<td>CRVE</td>
<td>N(0,1)</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Cluster ∼ T</td>
<td>CRVE</td>
<td>T(G-1)</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Wild Cluster - Rademacher</td>
<td>CRVE</td>
<td>-</td>
<td>2-point - rademacher</td>
</tr>
<tr>
<td>5</td>
<td>Wild Cluster - Normal</td>
<td>CRVE</td>
<td>-</td>
<td>~ N(0,1)</td>
</tr>
<tr>
<td>6</td>
<td>Wild Cluster - 4-point</td>
<td>CRVE</td>
<td>-</td>
<td>4-point equation (2.6)</td>
</tr>
<tr>
<td>7</td>
<td>Wild Cluster - 6-point</td>
<td>CRVE</td>
<td>-</td>
<td>6-point equation (2.7)</td>
</tr>
<tr>
<td>8</td>
<td>Enumeration - Rademacher</td>
<td>CRVE</td>
<td>-</td>
<td>2-point - rademacher</td>
</tr>
</tbody>
</table>

Table 2.2: Results from Monte Carlo Study with Different Numbers of Clusters: Replicating Results in Cameron, Gelbach, and Miller (2008)

<table>
<thead>
<tr>
<th>Number of Groups (G)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 OLS ∼ N(0,1)</td>
<td>0.468</td>
<td>0.486</td>
<td>0.493</td>
<td>0.494</td>
<td>0.489</td>
<td>0.499</td>
</tr>
<tr>
<td>2 CRVE ∼ N(0,1)</td>
<td>0.211</td>
<td>0.133</td>
<td>0.108</td>
<td>0.094</td>
<td>0.084</td>
<td>0.080</td>
</tr>
<tr>
<td>3 CRVE ∼ T(G − 1)</td>
<td>0.100</td>
<td>0.090</td>
<td>0.081</td>
<td>0.075</td>
<td>0.070</td>
<td>0.069</td>
</tr>
<tr>
<td>4 Wild 2pt BS</td>
<td>*0.037</td>
<td>*0.054</td>
<td>0.050</td>
<td>0.050</td>
<td>0.047</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Notes: Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). * - estimate is not accurately calculated. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00084 and the largest standard error is .00224.
### Table 2.3: Results from Monte Carlo Study with Different Numbers of Clusters: Small Number of Clusters Simulation

<table>
<thead>
<tr>
<th>Number of Groups (G)</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRVE $\sim T(G - 1)$</td>
<td>0.100</td>
<td>0.100</td>
<td>0.094</td>
<td>0.096</td>
<td>0.088</td>
<td>0.090</td>
</tr>
<tr>
<td>Wild $N(0, 1)$ BS</td>
<td>0.072</td>
<td>0.070</td>
<td>0.072</td>
<td>0.072</td>
<td>0.071</td>
<td>0.069</td>
</tr>
<tr>
<td>Wild 4pt BS</td>
<td>0.070</td>
<td>0.069</td>
<td>0.064</td>
<td>0.062</td>
<td>0.059</td>
<td>0.057</td>
</tr>
<tr>
<td><strong>Wild 6pt BS</strong></td>
<td><strong>0.070</strong></td>
<td><strong>0.067</strong></td>
<td><strong>0.063</strong></td>
<td><strong>0.061</strong></td>
<td><strong>0.057</strong></td>
<td><strong>0.056</strong></td>
</tr>
<tr>
<td>Enum. Lower Bound</td>
<td>0.118</td>
<td>0.095</td>
<td>0.084</td>
<td>0.068</td>
<td>0.062</td>
<td>0.060</td>
</tr>
<tr>
<td>Enum. Upper Bound</td>
<td>0.000</td>
<td>0.059</td>
<td>0.067</td>
<td>0.061</td>
<td>0.058</td>
<td>0.058</td>
</tr>
</tbody>
</table>

**Notes:** Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). Preferred procedure is presented in **bold**. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00000 and the largest standard error is .00144.

### Table 2.4: Results from Monte Carlo Study with Different Numbers of Clusters: Larger Number of Clusters Simulation

<table>
<thead>
<tr>
<th>Number of Groups (G)</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRVE $\sim T(G - 1)$</td>
<td>0.100</td>
<td>0.090</td>
<td>0.081</td>
<td>0.075</td>
<td>0.070</td>
<td>0.069</td>
</tr>
<tr>
<td>Wild 2pt BS</td>
<td><em>0.037</em></td>
<td><em>0.054</em></td>
<td>0.050</td>
<td>0.050</td>
<td>0.047</td>
<td>0.048</td>
</tr>
<tr>
<td>Wild $N(0, 1)$ BS</td>
<td>0.072</td>
<td>0.069</td>
<td>0.065</td>
<td>0.063</td>
<td>0.059</td>
<td>0.059</td>
</tr>
<tr>
<td>Wild 4pt BS</td>
<td>0.070</td>
<td>0.057</td>
<td>0.054</td>
<td>0.052</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td><strong>Wild 6pt BS</strong></td>
<td><strong>0.070</strong></td>
<td><strong>0.056</strong></td>
<td><strong>0.052</strong></td>
<td><strong>0.052</strong></td>
<td><strong>0.049</strong></td>
<td><strong>0.049</strong></td>
</tr>
<tr>
<td>Enum. Lower Bound</td>
<td>0.118</td>
<td>0.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enum. 2pt BS</td>
<td></td>
<td></td>
<td>0.052</td>
<td>0.051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enum. Upper Bound</td>
<td>0.000</td>
<td>0.058</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). * - estimate is not accurately calculated. Preferred procedure is presented in **bold**. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00084 and the largest standard error is .00144.
Chapter 3

How Targeted is Targeted Tax Relief?

Evidence from the Unemployment Insurance Youth Hires Program

Targeted employment subsidy programs are commonly employed by governments. This study examines one such initiative that rebated unemployment insurance premiums to employers with net increases in insurable earnings for youth aged 18 to 24. In each of two datasets, statistically and economically significant impacts on employment are observed for the targeted age group relative to older age groups. However, neither dataset exhibits a concurrent change in aggregate unemployment; instead there is a reduction in those not in the labor force. Oddly, no program impacts are
observed for females and all of the effects involve only males. Notably, evidence of displace- ment – substitution away from slightly older non-subsidized workers towards the younger subsidized group – is observed. These spillovers suggest that the aggregate impact of the program is less than that observed for the targeted group.

3.1 Introduction

In the wake of a recession, governments are often interested in stimulating employment growth with programs targeting particular locations, industrial sectors or demographic groups. One approach of potential interest in developed economies involves tax or social insurance premium rebates. The US Hiring Incentives to Restore Employment (HIRE) Act of 2010 is one such effort. It exempted employers from paying their share of Social Security, that is Old-Age, Survivors, and Disability Insurance (OASDI) taxes, for each hire who had been employed for 40 hours or less during the preceding 60 days. Since, in 2012, employers and employees each paid 6.2% of the employees annual earnings below $110,100, this amounted to a substantial subsidy for employers to hire unemployed or underemployed persons. Employers were also eligible for a $1000 retention credit for each of those new workers retained for at least one year. To prevent employers from laying off existing workers to make room for the new employees, the subsidy was calculated as a function of the difference between the relevant quarter’s wage base and that four quarters earlier, with the amount capped for each employee. It was therefore a credit for net new hires. Testifying before the Senate Committee on the Budget, Elmendorf (Congressional Budget Office, 2011) argued that such programs, which subsidize employers as a function of payroll growth, have the greatest effects on employment per dollar expended among
a range of policies they considered. Another targeted US federal employment subsidy is the Empowerment Zone program studied by Busso, Gregory and Kline (2013), but in this case the targeting is based on geographic “place” rather than unemployment status. Despite fears that this program would distort economic markets as a result of geographic displacement by firms and workers in response to the subsidy, they find positive benefits with only modest distortions.

In the same vein as the US programs, Canada’s government pursued two initiatives involving unemployment insurance (UI – renamed employment insurance or EI since 1996) following the recession of the 1990s, with one targeting youth and the other small firms. It is also currently pursuing a strategy targeting small businesses.¹ This chapter examines the labour market impacts of one of these short term programs that targeted youth unemployment. In 1999 and 2000 this program – “Youth Hires” – rebated any increase in aggregate UI premiums paid by firms for workers aged 18 to 24 that were in excess of the 1998 premiums they paid for that age group. While most economists believe that the relative inelasticity of the labour supply curve implies that changes in payroll taxes are passed on to workers through adjustments in wage rates in the long run, with minimal ensuing employment effects, there may be scope for a short term program to affect employment levels during a period of slack labour demand.² However, a program that targets a particular identifiable group may induce substitution towards the subsidized workers (i.e., displacement of close substitutes) and the program’s aggregate impact may differ from that experienced by the targeted


²A recent study by Owyang et al. (2013) presents evidence that government spending multipliers in Canada are much higher during periods of high unemployment than in periods of low unemployment.
group. We look for evidence of such effects.

A large research literature looks at optimal UI benefit rates (e.g., Chetty 2006, 2008), and the labour supply effects of UI, especially with respect to benefit duration (e.g., Card, Chetty and Weber (2007a,b)); for an overview see Krueger and Meyer (2002)). Of particular relevance to this study, much research addresses how workers and firms tailor their behavior to the parameters of the UI system, which is an important phenomenon in Canada where UI is not experience rated (see, e.g., Green and Riddell 1997; Green and Sargent 1998; Kuhn and Sweetman 1998). Kuhn and Riddell (2010) find appreciable long run responses to UI’s parameters in both nations in their contrast of adjacent regions in the US and Canada.

A related literature examines active labour market programs. Andersen and Svarer (2012) review a series of randomized experiments which occurred in Denmark. The experiments largely took the form of increased job counseling and had different target groups, including one targeting youth unemployment. The authors conclude that many of the interventions pass a cost benefit analysis, as the interventions often increased the employment rates of the targeted groups. A large scale program aimed at reducing youth unemployment was introduced in the UK in 1998. This New Deal for Young People is studied by De Giorgi (2005). The program was aimed at individuals 18 to 24 years of age, who had been unemployed for six months or more. The author uses a regression discontinuity design, testing the impact of the program for those just below the age cutoff versus those just above the age cutoff. The intervention, which in this case was national in scope, is estimated to increase the likelihood of being employed 18 months later by 6-7%. However, studies like these should be interpreted with some skepticism as recent work by Crepon et al. (2012) argues that
these types of studies do not consider general equilibrium effects. The authors study a set of randomized experiments in France, which aimed to reduce unemployment among educated youths. The authors show that there are significant displacement effects and that these are in general more significant for males than for females. Khan and Lehrer (2013) find evidence of displayment effects in a randomized experiment aimed at increasing the size of individuals’ social networks. Both Andersen and Svarer (2012) and De Giorgi (2005) mention the possibility of the control groups being disadvantaged by the policy changes, but their estimation strategy does not separately identify such displacement effects.

However, despite targeted programs being common, we are aware of relatively few studies evaluating employer-side targeted employment stimulus programs using social program premium rebates such as the Youth Hires program. A notable exception is Hernanz, Jimeno and Kugler (2003) who examine a 1997 reduction in Spanish payroll taxes and dismissal costs for permanent contract employees. Spain’s reforms reduced dismissal costs by 25% and payroll taxes by 40%. As in this chapter, the authors exploit differences in tax reductions for different age groups. They compare outcomes of those aged 20-29 against those of 30-39 year olds and find significant increases in the probability of being employed amongst the young treated population. O’Leary, Decke and Wandner (2005) study the related concept of targeted bonuses for UI recipients. They observe that only targeted bonuses are cost effective in the context of the US UI system and illustrate the importance of the choice of the target population for cost-effectiveness.

The long tradition of targeted subsidies and credits is evidenced by the US New Jobs Tax Credit (NJTC) studied by Perloff and Wachter (1979), Bishop and Haveman
(1979), and Bishop (1981), and the Targeted Jobs Tax Credit examined by Hollenbeck and Willke (1991). Britain’s Working Families’ Tax Credit, evaluated by Francesconi and van der Klaauw (2007), is an example of worker-side targeting. Some of these subsidies can be quite large. Bishop (1981) examines the NJTC, which offered a tax credit of 50% on the first $4200 of wages per employee for increases in employment in excess of 2%, with a cap of $100,000 per firm that effectively favored smaller firms. About $2.4 billion in credits were claimed in 1977, and $4.5 billion in 1978. Overall it is estimated that 28% of employers claimed a credit through the program in 1978. More recently, a broadly similar proposal was put forward by Bartik and Bishop (2009). In contrast to these US examples, the Canadian approach did not require vouchers nor did it target individuals, other than by age, so there is no stigmatization. In fact, the administration of the Canadian model was through employers, and the workers in question need not even have been aware of the UI premium rebate.

We do not attempt to estimate the impact of the Youth Hires program on aggregate job creation. Rather, using a difference-in-differences framework, we focus on the precursors to this by attempting to ascertain if there are any impacts on the targeted age group, and any displacement effects on those a little older. Classic work on displacement effects is the exploration of the UI bonus experiments by Davidson and Woodbury (1993). Their study showed the power of formal modeling in identifying general equilibrium effects and found that displacement undid a modest but nontrivial proportion of the program’s benefits. However identification relied on specific economic theories and functional form assumptions. Dahlberg and Forslund (2005) is a more recent examination of displacement from wage subsidies and training exploiting variation across municipalities in Sweden. They find substantial displacement effects.
from subsidies. Understanding the magnitude of any displacement effects in different contexts is fundamental to the evaluation of many labor market interventions.

Additionally, in our estimation context, there are well known problems of inference given that the source of randomization, the policy change, is at the aggregate level. We explore the sensitivity of our results to alternative methods for dealing with this issue, and for most of the analysis employ the Cameron, Gelbach and Miller (2008) approach, which has desirable finite sample properties that address problems associated with over-rejecting the null hypothesis. Overall, we observe a modest although discernible impact of Youth Hires in that it increases employment for the targeted 18-24 age group, but we also find what appears to be a concurrent employment decrease for those 25 to 29 suggesting that some substitution or displacement is occurring.

The paper is structured so that the next section provides the institutional background. Section 3.3 then describes the two independent data sets employed, defines two comparison groups that have different strengths and weaknesses, and presents descriptive statistics and an initial graphical analysis. Following that, section 3.4 addresses the econometric methodology with a focus on issues of interpretation and inference that are relevant in this context, and section 3.5 presents the empirical results where similar findings from both datasets add to our confidence in the analysis. The final section summarizes and interprets the findings.
3.2 Institutional Background Regarding UI and Youth Hires Program

Legally, the incidence of Canadian UI premiums is partitioned across employers and employees with employers paying 1.4 times the employee rate, which varies from year to year. This system is also notably different than the American program in that it operates nationally, and premiums are set by the federal government and are not experience rated against the history of either the employer or the employee.\(^3\) Premiums are invariant across regions and, since the premium rebate affected employers in all regions equally, we estimate the impact at the national level. Youth Hires was announced in the federal budget on February 24, 1998 and was described as being a temporary measure in place in 1999 and 2000 to address high youth unemployment rates. For workers who were aged 18-24 at any point during each calendar year, any premiums paid in 1999 and 2000 in excess of the firm’s 1998 premiums were refunded to the employer. Employer premium rates in 1998, 1999 and 2000 were respectively 3.78\%, 3.57\% and 3.36\% of insurable earnings with the maximum insurable earnings fixed at $39,000 in nominal terms.

One idiosyncrasy is that the premium rate decline implied that a firm’s aggregate UI insurable payroll for those in the relevant age group had to increase by, for example, 0.21\% in 1999 before the firm was entitled to the first dollar of rebate. Further, although the intention of the program was to increase youth employment by using premiums paid as the benchmark, employers had several margins on which they could

\(^3\)For a short period starting in 1997, the benefit rate was experience rated on the employee side. It decreased as the number of weeks of benefit receipt in the previous five years increased.
adjust. Employers were eligible for the credit if they increased insurable earnings sufficiently by any combination of increasing wages below the cap (including regular annual increases), the number of young workers employed, and/or hours per year for existing young employees. However, firms received no credit towards the rebate for any individual worker’s annual earnings that were in excess of the maximum insurable $39,000 limit. For more information on this program see Canada Employment Insurance Commission (1999, 2000, 2001).

Employers’ information sets are an extremely important determinant underlying any behavioral change they might undertake. If they are unaware of the program, then it only operates through easing the budget constraint on expanding firms and not through the behavioral change required to target youth; this also affects the timing of any effect. Clearly, in this case the government was interested in behavioral change since the goal of the program was to target unemployment among a specific age group. In addition to discussions of the program in the media and mailings to human resource departments or others in firms paying UI premiums, this program had the advantage of following on the heels of a similar program, the New Hires Program, that operated in 1997 and 1998. The earlier program refunded UI premiums associated with net job growth in small businesses.4

One of the criticisms of the earlier program was that many small firms were not aware of its existence. Also, it required an application to receive the refund that many small businesses found difficult (Canada Employment Insurance Commission, 2000). By contrast, Youth Hires was more broadly known and the premium rebate was presented as being automatic and without administrative burden, thereby making it more

---

4The program entitled firms with EI premiums of up to $60,000 to a full rebate on additional hires in 1997. It is broadly similar to the current program. Unfortunately, we are unable to examine the New Hires Program due to data limitations.
attractive. The program refunded over $400 million in premiums to approximately 295,000 firms (Canada Employment Insurance Commission, 1999 to 2002).\(^5\)

An important limitation to our analysis is the very substantial reform associated with the move from the UI to EI system, which was phased in during the six months ending January 1, 1997. This limits our ‘before’ period to two years for difference-in-differences analyses, and also limits any ‘falsification’ exercises in the pre-program period. One particularly relevant element of the reform for youth is that prior to the reform UI did not cover part-time jobs, defined as below both 15 hours per week and an earnings threshold, whereas EI premiums are paid from the first hour of work. Friesen (2002) finds a modest shift away from part-time, and towards full-time, employment following the move to EI and the associated introduction of EI premiums for part-time employment.\(^6\)

Given the nature of the Youth Hires program, we would not necessarily expect its introduction and termination to have equal and opposite impacts. If firms react to the incentive and hire new young workers, they must incur at least some training and other fixed hiring costs and, therefore, may continue to employ these workers after the rebate period expires. Of course, job mobility rates are quite high for young workers; thus while any impact may continue beyond the program’s life it will attenuate over time. In this vein, one group that will need special attention are those who are age 24 in the first year of the program but too old to be subsidized in its second year. We address this group in the empirical specification.

\(^5\)Given the roughly 10:1 ratio between the sizes of the Canadian and US economies and the exchange rates in effect at that time, this would have been equivalent to total US program expenditures of approximately $US5.9 billion across the two years in $1999.

\(^6\)Workers with very low annual earnings – far too low to qualify for benefits – have the employee share of their premiums refunded through the tax system. However, no such refunds are made to employers.
3.3 Data and Descriptive Statistics

We analyze individuals residing in Canadian provinces using the master files of Statistics Canada’s Survey of Labour and Income Dynamics (SLID) and Labour Force Survey (LFS). The SLID is a rotating panel that contains roughly 60,000 individuals in each wave, with overlapping waves starting every three years and lasting for six years; each individual’s annual labor market outcomes are detailed. In contrast, the LFS is comparable to the US Current Population Survey and interviews roughly 54,000 households comprising about 100,000 individuals and capturing labour market information on the week that contains the 15th of each month. For both datasets survey weights are used throughout.

Given that the EI reforms make it difficult to use data before 1997, and that hiring and training costs suggest the effects of the program are likely to continue beyond its termination, the bulk of the analysis focuses on 1997-2000, that is, the two years before, and the two years of, the program’s implementation.

Two comparison groups, with different strengths, are employed. The data for analysis are initially restricted to those aged 18-30, but the age range is then expanded to those aged 18-35. The advantage of having a comparison group aged 25 to 30 is that this age group operates in labor markets that are more similar to those for 18 to 24 year olds and therefore makes a useful comparison group. But, the same characteristics that make them a good comparison group also means that they are reasonably close substitutes in hiring/employment for the targeted younger workers, and may be negatively affected by Youth Hires with displacement occurring. In contrast, 18-24 year olds are less substitutable for those aged 30-35, which is beneficial in the latter group serving as a comparison group, but the older group also is less
likely to have a common trend in employment in the absence of the program, making it a slightly less satisfactory comparison group. Using both allows two perspectives on the policy change. The possibility of including individuals younger than age 18 was explored, but not pursued given the very large share still in high school, making them less comparable since they have different labour market dynamics.

Any significant impact of Youth Hires could affect variables for individuals treated such as the likelihood of being employed, or hours or weeks worked. However, as seen for minimum wages (e.g., Landon 1997; Neumark and Wascher 2004, while government policy may be motivated by unemployed youth aged 18-24 who are out of school, post-secondary (or even high school) attendance may also be affected for this age group, so we also investigate that outcome.

Summary Statistics

Table 3.1 contains mean values and sample sizes of dependent variables to be used in similarly specified regressions. These are presented by age group for the two years prior to, and the two years of, the Youth Hires program. In the upper panel the first three variables are from the SLID and are counts of annual weeks of employment, unemployment and not in the labor force status. These variables are mutually exclusive and sum to the number of weeks in the year. Next are three annual indicator (0/1) variables that are not mutually exclusive. The first is equal to one if the individual was employed at any point in the year, and zero otherwise. In a different vein, the second variable of this set measures the fraction of individuals who were not employed in the year although they sought employment (or were unemployed) at some point in the year. Similarly, the ‘not in the labor force’ indicator is set to one if the person is
out of the labor market at any point in the year. Total hours worked at all jobs in the year is next, and following it are the natural logarithm of total annual income and the hours-weighted average hourly wage across all jobs. Both of the earnings measures are deflated to 1999 dollars. The new job variable indicates whether an individual started with a new employer in the reference year, and the full-time indicator is set to one if an individual’s primary job was full-time. If the person was a full-time student at some point in the year the student variable is set to one.

In the lower panel of Table 3.1, the same statistics are presented for variables from the LFS. All variables in the LFS refer to the reference week. The LFS binary variables for employed, unemployed and not in the labour force are mutually exclusive and exhaustive. Total weekly hours worked is for all jobs in the reference week. Weekly income includes all income earned in those jobs, and is converted to the natural log of 1999 dollars; the hourly wage is likewise converted and is a weighted average for all jobs worked. New job is defined only for those who are currently working and is set to 1 if an individual started a new job in the reference week. Finally, student is a variable which indicates whether the individual was a full-time student in the reference week.

**Graphical Analysis**

Plots for three different variables are provided to illustrate the time trend in relevant dependent variables in the years of, and surrounding, the Youth Hires program. For various age groups in the SLID, Figure 3.1 shows the trends in annual total weeks employed, which is a central variable given the aims of the program. In the first year of the program, 1999, there are opposite effects for those treated by and those
excluded from the program. In 1999 we can see a sharp year over year increase in weeks employed by those aged 22-24. This contrasts with a slight decline by those 25-27 and 28-30. At the same time the weeks worked by those 18-21 increased in line with a trend experienced throughout 1997-2002. The sharp increase for those 22-24, coupled with the slight decline for those 25-27, is what one would expect to see if the program was effective in stimulating employment for the targeted group and simultaneously generating a modest amount of substitution/displacement. However, in contrast to this effect seen in the first year of the program, there is no obvious bump in the second year.

Weeks not in the labor force, also from the SLID, is presented in Figure 3.2 and a conceptually similar pattern is obvious. Of particular note, especially in the first year of the program, is the increase in weeks not in the labour force for those 25-27 coincident with a decrease in the weeks out of the labour force for those aged 22-24. Recalling that employers were eligible for the credits if they hired those 18-24 in 1999 or 2000, it appears as though workers of the younger age group were brought into the labor force in 1999 while those just excluded from (too old for) the program were slightly displaced. But, although there may be some ongoing effect, no additional increment is apparent for the second year of the program.\footnote{In discussions with stakeholders regarding Youth Hires it was suggested that some employers were initially drawn to the rebate, but then realized that the rebate was not sufficient given the productivity differences across the age groups in question. However, this is purely speculative.}

Figure 3.3 uses LFS data to plot the employment rate over time for the various age groups, in which we see a comparatively large increase in the employment rate for those aged 18-21 in the first year of the program. Here the other age groups also see increases – though not as large – in their employment rates, which is to be expected as general economic conditions were improving. Although we do not want
to draw too many conclusions at this stage of the analysis, these graphs support
the idea that employers were preferentially hiring those subsidized by the program.
Moreover, although noticeable, the magnitude of the aggregate affect is modest in all
three graphs. Clearly, there are a large number of employers who are increasing the
size of their workforce as a result of macroeconomic trends and for whom this rebate
is a windfall gain.

3.4 Econometric Approach

We employ a framework that, in terms of the equations estimated, looks like a linear
difference-in-differences (DiD) specification. However, for the initial set of models
estimated – and perhaps for all of them – the results do not have the usual interpre-
tation as the causal impact of the treatment on the treated. Both theory and the
graphical analysis suggest that the common trend assumption required to identify
such a parameter is not satisfied (see, e.g., DiNardo and Lee, 2011) given that the
program potentially has both direct causal impacts on the targeted age groups, and
indirect causal impacts on workers slightly older than the programs maximum age.
That is, it seems plausible that the 25 to 30 age group, which is too old for Youth
Hires, is displaced by the program. In this situation, the DiD coefficient is perhaps
best interpreted as the causal change in the gap between the treatment and comp-
parison groups across the policy periods assuming that they would otherwise have a
common trend, and not as the impact of the policy change on the treatment group.

Beyond identification, inference using the DiD specification with a policy change
at the aggregate level can be problematic as demonstrated by Bertrand, Duflo and
Mullainathan (2004) who find, among other issues, that the standard approach which
relies on the asymptotic properties of the cluster-robust variance estimator does not function well when there is a small number of clusters. In this case there are only 16 clusters since we take each annual birth cohort as the basic unit affected by the policy change.

Cameron, Gelbach, and Miller (2008 – CGM hereafter) based on both econometric theory and Monte Carlo evidence argue that wild cluster bootstrap-t techniques work well even when the number of clusters is small; we employ this approach. This technique is the subject of Chapter 2 of this thesis. The proposed 6-point distribution is not used here as the Rademacher distribution works well with 16 clusters. In accord with Donald and Lang (2007), CGM’s Monte-Carlo simulations also suggest that over-rejection is less severe if we assume that the t-statistics follow a distribution with G-2 degrees of freedom, with G being the number of clusters and 2 being the number of within-cluster parameters estimated. Initially, we explore alternative approaches to inference and observe some variation. However, for the vast majority of the analysis we present only results from our preferred method of inference, which is to generate bootstrap samples using the wild cluster bootstrap-t technique with the null hypothesis imposed. This allows us to generate an empirical distribution of t-statistics – which allows for asymptotic refinement since the t-statistic is asymptotically pivotal – from which a p-value for the test statistic can be obtained.\(^8\) A downside is that this approach bypasses the generation of standard errors, which Angrist and Pischke (2008) argue some economists like to observe.

The first specification we estimate employs data from 1997-2000 and uses those aged 18-24 as the treated group, with 25-30 year-olds as the comparison group, and

\(^8\)We thank Cameron, Gelbach and Miller for making their code available.
is specified in equation (3.1).

\[ Y_{it} = \beta_0 + \beta_{YH} Y_{it} + \beta_{99} \text{Only1999}_{it} + \beta_A \text{Age}_{it} \]

\[ + \beta_B \text{YearBorn}_i + \beta_Y \text{Year}_t + [\beta_c \text{Controls}_{it}] + \epsilon_{it} \] (3.1)

In this equation \( Y_{it} \) represents a labor market variable of interest; \( YH \) is the Youth Hires indicator which is set to one if individual \( i \) is of an age to be affected by the program in a year, \( t \), when it is operating; and \( \text{Only1999} \) is an indicator set equal to one for individuals who qualify for the subsidy in the first year of the program, but not the second. \( \text{Age} \), \( \text{YearBorn} \) and \( \text{Year} \) are all vectors comprising full sets of indicator variables that respectively represent the individual’s age (measured in years as of year \( t \)) and birth year, and the calendar year in question. This represents an effort to flexibly control for any background effects that may influence the coefficient of interest. The vector of variables identified as Controls are in brackets to indicate they are included in some, but not all, specifications. For both datasets, the control variables are province of residence as well as an indicator for urban residence, while the SLID regressions additionally control for race and immigrant status. Models estimated using the LFS, but not the SLID, also include a full set of months indicators. The \( \beta \)s are vectors of coefficients to be estimated.

In all cases, the equations are estimated using ordinary least squares (OLS). Hence some specifications are linear probability models. Clustering is on the individual’s birth year since that allows a longitudinal dimension; \( \epsilon_{it} \) is allowed to be arbitrarily correlated within clusters, but is assumed to be independent across birth cohorts. In some specifications employing the SLID data, the error term is decomposed to include
an individual fixed effect recognizing that individuals are in the data for (up to, since there is some attrition) six years.

The coefficient $\beta_\text{YH}$ is the DiD variable of interest and, as mentioned, its estimate will conflate any positive impact on those in the treatment group with any negative impact those in the comparison group may experience in the years when the program is operating. We are agnostic as to the expected sign of $\beta_{99}$ since it will hinge on the impact of the program in 1999 and the degree of labor market attachment in the subsequent year. We do not report the coefficients for $\beta_{99}$ in the text, though in general the coefficients are of the same sign, smaller in magnitude and of lesser significance than the coefficients for $\beta_\text{YH}$.

A subsequent specification is estimated using data on individuals aged 18-35. It is not shown since it is similar to equation (3.1) except that the $YH$ indicator is interacted with a set of indicators for those in the 18-21, 22-24, 25-27, and 28-30 age groups. Individuals aged 31 to 35 serve as the omitted comparison group. Plausibly, this comparison group is not (or is minimally) affected by the Youth Hires program, so substitution/displacement is minimized. However, it is less credible that this older age group would have a similar trajectory across time as that of the treated age groups in the absence of the policy change. That is, the common trend assumption is less credible given the larger gap in age and the well-known differences across the business cycle in rates of unemployment, job turnover and the like with age. Nevertheless, any difference in trends may not be that large over a short period, and this model is estimated since it allows heterogeneity across age groups during the time of the program to be observed. Additionally, the estimates for this group are shown using only the 1998-1999 data to capture the initial impact of the policy. The pattern of
results was generally of the same when estimated using the 1997-2000 data.

An attempt was made to conduct a three period analysis of the program, trying to
determine the outcomes of the targeted group before, during and after the program. However, this was frustrated by the lack of a clear comparison group in the “after”
period. The issue is that the individuals treated in 1999 and 2000 would be those aged 20-26 in 2002, but that age range would contain both treated and untreated
individuals in the year 2000.

To further test the robustness of our research design, we conduct a series of falsi-
fication exercises\(^9\) using data from 2002 to 2005 (i.e., leaving a two-year gap after the
end of the program in case there are any “knock on” effects, but being reluctant to ex-
tend too far from the policy change given the possibility of other age-specific changes
derived from the education system). In general, it would be preferable to conduct a
falsification exercise using a period prior to the program. Unfortunately, as discussed,
the significant reforms in 1996-1997 render any pre-period analysis problematic.

3.5 Regression Analysis

Various approaches of inference for equation (3.1) using three key dependent variables,
all measures of employment which is the central policy variable for Youth Hires, are
presented in Table 3.2. The first two regressions use SLID data, and the third uses
LFS data. For each dependent variable there are two OLS specifications one with a
minimal set of covariates, and the other with a full set of controls. For the data from
the SLID there is also a specification including both individual fixed effects and a full
set of controls. In all cases, but particularly for the OLS regressions which are less
\(^9\)See DiNardo and Lee (2011) regarding the benefits of falsification tests.
time-consuming to bootstrap, an extremely large number of bootstrap replications are employed to increase the precision with which we can estimate p-values.

**Comparison Group Aged 25-30**

Coefficients are presented in the first line and show sensible modest increases among the targeted group, relative to the slightly older one, associated with the program: approximately two to two and a half extra weeks of employment in the year, and an increase of about 3.5% to 4% in the employment rate as measured in the SLID or just over 1% as measured using the LFS. When clustering is ignored, but the standard errors are adjusted for heteroscedasticity of unknown form, the p-values in the SLID are larger than those observed in any other test, but when individual fixed-effects are employed (with the associated clustering on the individual) the p-value seems inappropriately small. In the subsequent rows a series of p-values are presented using Stata’s “cluster” command, but choosing different degrees of freedom. Given the 16 clusters in this dataset the degrees of freedom adjustment makes a difference, but it is relatively modest. Two implementations of the wild cluster bootstrap-t are also undertaken – the first without, and the second with, the null hypothesis imposed. For the OLS models, the p-values increase slightly, but still mostly indicate statistical significance at conventional levels. For the fixed effect model, the p-values actually decreased slightly. In accord with the evidence in CGM and Davidson and MacKinnon (1999), we take the wild bootstrap with the null imposed as our preferred approach to inference. It is reassuring, however, to see that there are not enormous differences in inference across the last three approaches, which are arguably superior to the others. Although to save space we do not present the information, we observe
similar patterns for the other dependent variables that we tested where the coefficients were statistically significant.

In terms of the substantive results, they almost everywhere indicate statistical significance at conventional levels with the least statistically significant p-value being 12%. This provides robust evidence – based on alternative approaches to inference and three variables from two datasets – that Youth Hires had a causal effect increasing employment for the targeted age group relative to those slightly older.

Results from the specification in equation (3.1) for a wide range of dependent variables are presented in Table 3.3. In the upper panel are those from the SLID, while those from the LFS are in the lower one. Among the dependent variables from each dataset, those at the top of each panel are alternative measures related to employment, unemployment and out of the labor market status. The SLID provides two measures of each, whereas the LFS only has one. Employment is the only variable for which there is a strong prior expectation regarding the sign of the coefficient if the program is functioning as intended. Although some government planners might also have expected unemployment to decrease, it is well known by labor economists that the unemployment rate may increase in the early part of the expansionary period of a business cycle as the economy improves, since discouraged workers shift their labor market status from out of the labor market to unemployed as they begin active job search.

Those dependent variables in the lower half of each panel represent ancillary features of the labor market that may be affected by the policy change, but we are agnostic regarding the expected sign of the coefficient since theory suggests that there might be opposing effects in operation. For example, average hours of work could
increase if any additional employment results from increasing the hours of part-time workers, potentially making them full-time, or could decrease if additional part-time youth are added to the labor force. Additionally, as with minimum-wage legislation, the additional opportunities for employment could draw youth out of school so that the percentage who are full-time students might decline, but there need not be an effect on this dimension. These dependent variables are included to improve our understanding of the program’s impact.

The equations estimated in this table are also divided along gender lines, and this highlights a very interesting finding. Essentially none of the coefficients, in both datasets, is statistically significant for females. For some reason the entire policy response to the Youth Hires program appears to be concentrated among the males. Or, alternatively, the response is more muted for females and is statistically insignificant given the limited precision in estimating the coefficients that is feasible with only 16 birth cohorts (degrees of freedom).

Looking first at the coefficient estimates in the upper half of the table for each dataset, the positive effect on employment seen in Table 3.2 is repeated in Table 3.3.\textsuperscript{10} Moreover, both datasets are consistent in finding that there is no statistically significant change in the unemployment rate associated with the program; rather, there is a reduction in the various measures of ‘not in the labour force’. Overall, and this is a central finding of the analysis of this program, there is consistent evidence that the program served to increase the relative employment rate of the targeted group compared to those slightly older, but the aggregate effect was to draw youth

\textsuperscript{10}It is important to note the differences in the p-values for the employment regressions in Table 3.2 and Table 3.3 result from the Table 3.2 values being generated using 9999 bootstraps, while the Table 3.3 values are generated using 1499 bootstraps. Even with 1499 bootstraps the estimated p-values are not yet perfectly stable.
into the labor force.

Hours of work are not statistically significantly affected by the policy change in either dataset, and the estimated coefficients are of opposite signs across the datasets. Wages and/or annual earnings also appear to be largely unaffected, although the point estimates are mostly negative and one of them is statistically significant. Similarly, the results are mixed for the incidence of new jobs, but there appears to be a small decrease in the LFS, and also a small decrease in the incidence of full-time employment for youth. Finally, there is no evidence that this policy change is inducing students to leave their studies using this comparison group.

**Comparison Group Aged 31-35**

Those aged 31-35 serve as the comparison group in the regressions in Table 3.4 where they are compared to those subsidized by the premium rebate, as well as to those aged 25-30 who were previously used as the comparison group. Further, those 18-24 and 25-30 are each subdivided into two smaller age groups to highlight any patterns across age. The very first row, looking at annual weeks employed in the SLID, tells an interesting story. Relative to the 31-35 age group, the younger two groups have point estimates that show appreciable relative increases in their weeks of work, with that for the youngest group being statistically significant. In contrast, the coefficients for the two age groups just outside the age cutoff for the EI premium rebate have negative coefficients, with that for the older of these two being statistically significant. Akin to the informal analysis of Figure 3.1, this suggests displacement/substitution is likely to have occurred as a result of Youth Hires. Clearly, understanding the spillovers from this program is important for understanding its aggregate impact. For the LFS,
the coefficients on employment tell a similar story, but the pattern is not quite as extreme in that neither of the coefficients for the two older age groups is statistically significant and one of them is actually positive although close to zero. Although the coefficients change slightly, the remainder of the coefficients on labor force status variables largely support an interpretation suggesting that the policy change increased employment among the targeted age groups. In terms of the magnitude of the effects, they are appreciable, but not enormous, which accords with the magnitude of the subsidy associated with Youth Hires.

By breaking the subsidy-eligible group into an older and younger half, Table 3.4 also makes obvious the finding that the effects of the program appear to be larger for the 18-21 age group than the 22-24 one, which differs slightly from the informal graphical analysis. Also, some of the coefficients in the bottom half of each dataset’s panel that were not statistically significant in Table 3.3 are significant in Table 3.4. In particular, there is some evidence in this specification that students were drawn out of school as a result of the subsidy to employment targeted at their age group. To compare our findings with those of Crepon et al. (2012) that the displacement effects are more significant for men than for women, we estimated a version of the model presented in Table 3.4 separately for each gender. The estimates have been omitted in the interest of space but they also find that the displacement effects are significant for men, but not for women. This is consistent with the results in Table 3.3 where essentially none of the impacts is significant for women.
Falsification Tests

Although the timing of the data for the falsification tests presented in Table 3.5 is not ideal, the results provide some support for the analysis. Of the 63 regression coefficients estimated, seven of them are statistically significant at the 10% level, with most of these being significant between 5% and 10%. This is well within the range of what one would expect if random numbers were being regressed to generate the coefficients in question. Importantly, all of the employment related coefficients have point estimates very close to zero and none of them is statistically significant. Also, the significant coefficients in Table 3.5 do not accord with a pattern that is easy to interpret as being consistent with an alternative interpretation of the programs impacts. Overall, there does not appear to be evidence to undermine the conclusions in this analysis.

3.6 Discussion and Conclusion

We examine the effectiveness of a stimulus program in Canada that was designed to temporarily target high youth unemployment. The Youth Hires program subsidized employers to hire youth between the ages of 18 and 24 by rebating UI premiums for net new insured employment. Overall, we believe the evidence supports the conclusion that this program served to increase employment among the subsidized population by about one or two weeks per year on average, or one or two percentage points, relative to older individuals. Oddly, it appears that the effect of the program was

\[11\] Of course, it also reminds us that results in the earlier tables are subject to both type I and type II errors; although the earlier tables have, proportionately, far more coefficients that are statistically significant and typically with smaller p-values.
predominantly, if not entirely, experienced by males. For some reason female employment appears not to have been much affected. In interpreting these results it is worth remembering that the value of the annual rebate for net new youth employment was around 3.5% of earnings below $39,000 per worker/year, so an impact larger than that observed is not expected and a substantial percentage of the total subsidy payment can be thought of as a windfall gain for employers.

Youth Hires also appears to have had an impact, albeit smaller, in reducing the labor market outcomes of those slightly too old to be eligible for the EI premium rebate. This substitution/displacement effect points to the trade-offs ubiquitous in social policy development, especially when programs seek to assist targeted groups. Any future cost-benefit analysis of the program needs to take into account the likelihood that some of the costs are borne by those aged 25 to 30 (and perhaps others) via substitution/displacement. Of course, an analysis such as this cannot answer the broader general equilibrium question about the number of jobs produced in, or the benefits accruing to, the economy as a whole as a result of the program.

Results from this evaluation may be relevant in particular to the US HIRE legislation targeting new hires meeting certain (low) hours of work criteria in the 60 days preceding the subsidized employment. Those just above the threshold may be somewhat displaced. However, the findings also point to the benefit of having a threshold that includes those with low but positive hours of pre-subsidy work since this likely has less displacement than would a threshold of zero hours (true unemployment). Also, in many ways our results are similar to those found by Busso, Gregory and Kline (2013). There is likely some distortion resulting from the policy, but not enough to overturn the intended impact of the program.
<table>
<thead>
<tr>
<th>SLD</th>
<th>18-24 (Treated)</th>
<th>25-30 (Comparison)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before (97-98)</td>
<td>During (99-00)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>N</td>
</tr>
<tr>
<td>SLID</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annual Weeks Employed</td>
<td>30.01</td>
</tr>
<tr>
<td></td>
<td>Annual Weeks Unemployed</td>
<td>4.36</td>
</tr>
<tr>
<td></td>
<td>Annual Wks Not Lbr Force</td>
<td>18.63</td>
</tr>
<tr>
<td></td>
<td>E Any Time in Year</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>U Any Time in Year</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>N Any Time in Year</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Total Annual Hours</td>
<td>8.96</td>
</tr>
<tr>
<td></td>
<td>ln(Annual Income)</td>
<td>8.89</td>
</tr>
<tr>
<td></td>
<td>ln(Average Wage)</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td>New Job</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Mostly Full-time work</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Full-time Student in Year</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>LFS Employed</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Unemployed</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Not in Labor Force</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Total Weekly Hours Worked</td>
<td>30.29</td>
</tr>
<tr>
<td></td>
<td>ln(Weekly Income)</td>
<td>5.49</td>
</tr>
<tr>
<td></td>
<td>ln(Wage)</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td>New Job</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Table 3.2: Difference-in-Differences Employment Regressions

<table>
<thead>
<tr>
<th></th>
<th>SLID</th>
<th>LFS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weeks Employed</td>
<td>Employed</td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>OLS</td>
</tr>
<tr>
<td>Coefficient</td>
<td>2.409</td>
<td>2.320</td>
</tr>
<tr>
<td>Hetero. consistent std err</td>
<td>1.307</td>
<td>1.274</td>
</tr>
<tr>
<td>Clustered std err</td>
<td>0.730</td>
<td>0.784</td>
</tr>
<tr>
<td>t-stat hetero</td>
<td>1.844</td>
<td>1.821</td>
</tr>
<tr>
<td>t-stat cluster</td>
<td>3.299</td>
<td>2.957</td>
</tr>
<tr>
<td>p-value hetero df=N-k</td>
<td>0.065</td>
<td>0.069</td>
</tr>
<tr>
<td>p-value cluster df=N-k</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>p-value cluster df=G-1</td>
<td>0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>p-value cluster df=G-2</td>
<td>0.005</td>
<td>0.010</td>
</tr>
<tr>
<td>p-value wild bootstrap</td>
<td>0.020</td>
<td>0.052</td>
</tr>
<tr>
<td>p-value wild bootstrap null imposed</td>
<td>0.009</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Bootstrap replications | 9999   | 9999  | 1499  | 9999   | 9999  | 1499  | 9999   | 9999  |
Number of clusters     | 16     | 16    | 16    | 16     | 16    | 16    | 16     | 16    |
Min cluster size        | 928    | 928   | 928   | 928    | 928   | 928   | 20,452 | 20,452|
Full set of controls    | No     | Yes   | Yes   | No     | Yes   | Yes   | No     | Yes   |

Notes: Variables included in the regressions with controls are: province of residence and urban residence for both datasets, visible minority and immigrant status for SLID. The ‘hetero. consistent std error’ for the FE regressions and the associated t-stat and p-value are estimated clustering on the individual, whereas the estimates for the ‘hetero consistent std err’ off the OLS regressions are not clustered.
Table 3.3: Difference-in-Differences Estimates for 18-24 Year Olds, 25-30 as Comparison Group

<table>
<thead>
<tr>
<th>SLID</th>
<th>All Coeff</th>
<th>Wild Coeff</th>
<th>Female Coeff</th>
<th>Wild Coeff</th>
<th>Male Coeff</th>
<th>Wild Coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
</tr>
<tr>
<td>Wild</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Weeks Employed</td>
<td>2.320</td>
<td>0.019</td>
<td>1.028</td>
<td>0.504</td>
<td>3.602</td>
<td>0.007</td>
</tr>
<tr>
<td>Annual Weeks Unemployed</td>
<td>0.059</td>
<td>0.851</td>
<td>0.714</td>
<td>0.209</td>
<td>-0.562</td>
<td>0.545</td>
</tr>
<tr>
<td>Annual Wks Not Lbr Force</td>
<td>-2.379</td>
<td>0.008</td>
<td>-1.742</td>
<td>0.355</td>
<td>-3.040</td>
<td>0.007</td>
</tr>
<tr>
<td>E Any Time in Year</td>
<td>0.036</td>
<td>0.113</td>
<td>0.026</td>
<td>0.559</td>
<td>0.048</td>
<td>0.031</td>
</tr>
<tr>
<td>U Any Time in Year</td>
<td>-0.005</td>
<td>0.736</td>
<td>0.024</td>
<td>0.248</td>
<td>-0.034</td>
<td>0.171</td>
</tr>
<tr>
<td>N Any Time in Year</td>
<td>-0.055</td>
<td>0.016</td>
<td>-0.011</td>
<td>0.677</td>
<td>-0.100</td>
<td>0.037</td>
</tr>
<tr>
<td>Total Annual Hours</td>
<td>22.888</td>
<td>0.480</td>
<td>14.974</td>
<td>0.713</td>
<td>34.657</td>
<td>0.345</td>
</tr>
<tr>
<td>ln(Annual Income)</td>
<td>-0.073</td>
<td>0.108</td>
<td>-0.060</td>
<td>0.481</td>
<td>-0.090</td>
<td>0.056</td>
</tr>
<tr>
<td>ln(Average Wage)</td>
<td>-0.038</td>
<td>0.017</td>
<td>-0.049</td>
<td>0.185</td>
<td>-0.027</td>
<td>0.337</td>
</tr>
<tr>
<td>New Job</td>
<td>-0.006</td>
<td>0.880</td>
<td>0.050</td>
<td>0.292</td>
<td>-0.049</td>
<td>0.393</td>
</tr>
<tr>
<td>Mostly Full-time work</td>
<td>-0.042</td>
<td>0.021</td>
<td>-0.002</td>
<td>0.972</td>
<td>-0.077</td>
<td>0.024</td>
</tr>
<tr>
<td>Full-time Student in Year</td>
<td>-0.028</td>
<td>0.248</td>
<td>-0.013</td>
<td>0.591</td>
<td>-0.044</td>
<td>0.152</td>
</tr>
<tr>
<td>LFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>0.012</td>
<td>0.056</td>
<td>0.002</td>
<td>0.847</td>
<td>0.022</td>
<td>0.056</td>
</tr>
<tr>
<td>Unemployed</td>
<td>0.004</td>
<td>0.440</td>
<td>0.005</td>
<td>0.539</td>
<td>0.002</td>
<td>0.795</td>
</tr>
<tr>
<td>Not in Labor Force</td>
<td>-0.015</td>
<td>0.007</td>
<td>-0.007</td>
<td>0.524</td>
<td>-0.024</td>
<td>0.011</td>
</tr>
<tr>
<td>Total Weekly Hours Worked</td>
<td>-0.113</td>
<td>0.605</td>
<td>-0.309</td>
<td>0.269</td>
<td>0.109</td>
<td>0.807</td>
</tr>
<tr>
<td>ln(Weekly Income)</td>
<td>-0.003</td>
<td>0.701</td>
<td>0.004</td>
<td>0.787</td>
<td>-0.008</td>
<td>0.557</td>
</tr>
<tr>
<td>ln(Wage)</td>
<td>-0.002</td>
<td>0.759</td>
<td>0.001</td>
<td>0.916</td>
<td>-0.004</td>
<td>0.788</td>
</tr>
<tr>
<td>New Job</td>
<td>-0.028</td>
<td>0.007</td>
<td>-0.033</td>
<td>0.001</td>
<td>-0.023</td>
<td>0.100</td>
</tr>
<tr>
<td>Student</td>
<td>0.005</td>
<td>0.520</td>
<td>0.013</td>
<td>0.184</td>
<td>-0.002</td>
<td>0.916</td>
</tr>
</tbody>
</table>

Notes: Wild p-values based on 1499 bootstrap replications. All regressions have the full set of control variables listed in table 3.2.
### Table 3.4: Difference-in-Differences Coefficient Estimates with 31-35 Year Olds as the Comparison Group

<table>
<thead>
<tr>
<th></th>
<th>18-21</th>
<th>22-24</th>
<th>25-27</th>
<th>28-30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wild</td>
<td>Wild</td>
<td>Wild</td>
<td>Wild</td>
</tr>
<tr>
<td></td>
<td>Coeff</td>
<td>p-value</td>
<td>Coeff</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>SLID</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Weeks Employed</td>
<td>3.196</td>
<td>0.020</td>
<td>2.953</td>
<td>0.476</td>
</tr>
<tr>
<td>Annual Weeks Unemployed</td>
<td>1.256</td>
<td>0.160</td>
<td>0.437</td>
<td>0.175</td>
</tr>
<tr>
<td>Annual Wks Not Lbr Force</td>
<td>-4.452</td>
<td>0.008</td>
<td>-3.390</td>
<td>0.440</td>
</tr>
<tr>
<td>E Any Time in Year</td>
<td>0.053</td>
<td>0.117</td>
<td>0.023</td>
<td>0.472</td>
</tr>
<tr>
<td>U Any Time in Year</td>
<td>-0.019</td>
<td>0.253</td>
<td>0.018</td>
<td>0.009</td>
</tr>
<tr>
<td>N Any Time in Year</td>
<td>-0.059</td>
<td>0.009</td>
<td>-0.081</td>
<td>0.272</td>
</tr>
<tr>
<td>Total Annual Hours</td>
<td>130.932</td>
<td>0.027</td>
<td>105.237</td>
<td>0.333</td>
</tr>
<tr>
<td>ln(Annual Income)</td>
<td>0.212</td>
<td>0.008</td>
<td>0.131</td>
<td>0.196</td>
</tr>
<tr>
<td>ln(Average Wage)</td>
<td>0.021</td>
<td>0.348</td>
<td>0.030</td>
<td>0.268</td>
</tr>
<tr>
<td>New Job</td>
<td>-0.019</td>
<td>0.325</td>
<td>-0.078</td>
<td>0.148</td>
</tr>
<tr>
<td>Mostly Full-time work</td>
<td>0.078</td>
<td>0.032</td>
<td>0.018</td>
<td>0.707</td>
</tr>
<tr>
<td>Full-time Student in Year</td>
<td>-0.095</td>
<td>0.007</td>
<td>-0.104</td>
<td>0.035</td>
</tr>
<tr>
<td><strong>LFS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>0.061</td>
<td>0.008</td>
<td>0.034</td>
<td>0.008</td>
</tr>
<tr>
<td>Unemployed</td>
<td>-0.007</td>
<td>0.372</td>
<td>0.000</td>
<td>0.691</td>
</tr>
<tr>
<td>Not in Labor Force</td>
<td>-0.053</td>
<td>0.008</td>
<td>-0.033</td>
<td>0.024</td>
</tr>
<tr>
<td>Total Weekly Hours Worked</td>
<td>3.106</td>
<td>0.005</td>
<td>2.090</td>
<td>0.004</td>
</tr>
<tr>
<td>ln(Weekly Income)</td>
<td>0.225</td>
<td>0.016</td>
<td>0.148</td>
<td>0.001</td>
</tr>
<tr>
<td>ln(Wage)</td>
<td>0.078</td>
<td>0.004</td>
<td>0.065</td>
<td>0.035</td>
</tr>
<tr>
<td>New Job</td>
<td>-0.020</td>
<td>0.009</td>
<td>-0.009</td>
<td>0.431</td>
</tr>
<tr>
<td>Student</td>
<td>-0.087</td>
<td>0.011</td>
<td>-0.052</td>
<td>0.023</td>
</tr>
</tbody>
</table>

**Notes:** Wild p-values based on 1499 bootstrap replications. All regressions have the full set of control variables listed in table 3.2.
Table 3.5: Falsification Test Using Data From 2002-2005

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Hetero</th>
<th>Wild</th>
<th>Female</th>
<th>Hetero</th>
<th>Wild</th>
<th>Male</th>
<th>Hetero</th>
<th>Wild</th>
<th>Wild</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef</td>
<td>p-value</td>
<td>p-value</td>
<td>Coef</td>
<td>p-value</td>
<td>p-value</td>
<td>Coef</td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
</tr>
<tr>
<td>SLID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Weeks Employed</td>
<td>-0.090</td>
<td>0.945</td>
<td>0.979</td>
<td>1.186</td>
<td>0.520</td>
<td>0.388</td>
<td>-1.336</td>
<td>0.450</td>
<td>0.439</td>
<td></td>
</tr>
<tr>
<td>Annual Weeks Unemployed</td>
<td>0.526</td>
<td>0.423</td>
<td>0.329</td>
<td>0.895</td>
<td>0.264</td>
<td>0.335</td>
<td>0.111</td>
<td>0.914</td>
<td>0.855</td>
<td></td>
</tr>
<tr>
<td>Annual Wks Not Lbr Force</td>
<td>-0.437</td>
<td>0.713</td>
<td>0.687</td>
<td>-2.082</td>
<td>0.236</td>
<td>0.316</td>
<td>1.224</td>
<td>0.435</td>
<td>0.363</td>
<td></td>
</tr>
<tr>
<td>E Any Time in Year</td>
<td>-0.022</td>
<td>0.360</td>
<td>0.387</td>
<td>-0.003</td>
<td>0.922</td>
<td>0.951</td>
<td>-0.038</td>
<td>0.231</td>
<td>0.241</td>
<td></td>
</tr>
<tr>
<td>U Any Time in Year</td>
<td>0.021</td>
<td>0.126</td>
<td>0.051</td>
<td>0.015</td>
<td>0.409</td>
<td>0.547</td>
<td>0.026</td>
<td>0.205</td>
<td>0.167</td>
<td></td>
</tr>
<tr>
<td>N Any Time in Year</td>
<td>-0.025</td>
<td>0.386</td>
<td>0.585</td>
<td>-0.078</td>
<td>0.061</td>
<td>0.093</td>
<td>0.030</td>
<td>0.450</td>
<td>0.487</td>
<td></td>
</tr>
<tr>
<td>Total Annual Hours</td>
<td>9.816</td>
<td>0.851</td>
<td>0.759</td>
<td>34.133</td>
<td>0.634</td>
<td>0.383</td>
<td>-14.198</td>
<td>0.851</td>
<td>0.803</td>
<td></td>
</tr>
<tr>
<td>ln(Average Wage)</td>
<td>0.021</td>
<td>0.742</td>
<td>0.772</td>
<td>0.016</td>
<td>0.867</td>
<td>0.836</td>
<td>0.032</td>
<td>0.718</td>
<td>0.721</td>
<td></td>
</tr>
<tr>
<td>New Job</td>
<td>-0.035</td>
<td>0.281</td>
<td>0.299</td>
<td>-0.042</td>
<td>0.356</td>
<td>0.271</td>
<td>-0.029</td>
<td>0.525</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>Mostly Full-time work</td>
<td>-0.004</td>
<td>0.870</td>
<td>0.813</td>
<td>-0.030</td>
<td>0.469</td>
<td>0.152</td>
<td>0.022</td>
<td>0.525</td>
<td>0.431</td>
<td></td>
</tr>
<tr>
<td>Full-time Student in Year</td>
<td>0.008</td>
<td>0.777</td>
<td>0.757</td>
<td>0.038</td>
<td>0.321</td>
<td>0.073</td>
<td>-0.022</td>
<td>0.585</td>
<td>0.557</td>
<td></td>
</tr>
<tr>
<td>LFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>0.003</td>
<td>0.634</td>
<td>0.805</td>
<td>-0.013</td>
<td>0.101</td>
<td>0.137</td>
<td>0.016</td>
<td>0.029</td>
<td>0.239</td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>-0.009</td>
<td>0.010</td>
<td>0.088</td>
<td>0.001</td>
<td>0.748</td>
<td>0.789</td>
<td>-0.018</td>
<td>0.000</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>Not in Labor Force</td>
<td>0.006</td>
<td>0.198</td>
<td>0.368</td>
<td>0.011</td>
<td>0.111</td>
<td>0.083</td>
<td>0.002</td>
<td>0.765</td>
<td>0.843</td>
<td></td>
</tr>
<tr>
<td>Total Weekly Hours Worked</td>
<td>0.233</td>
<td>0.275</td>
<td>0.488</td>
<td>0.839</td>
<td>0.006</td>
<td>0.112</td>
<td>-0.337</td>
<td>0.256</td>
<td>0.507</td>
<td></td>
</tr>
<tr>
<td>ln(Weekly Income)</td>
<td>-0.007</td>
<td>0.458</td>
<td>0.684</td>
<td>0.012</td>
<td>0.373</td>
<td>0.588</td>
<td>-0.026</td>
<td>0.031</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>ln(Wage)</td>
<td>-0.016</td>
<td>0.003</td>
<td>0.208</td>
<td>-0.013</td>
<td>0.075</td>
<td>0.353</td>
<td>-0.019</td>
<td>0.009</td>
<td>0.131</td>
<td></td>
</tr>
<tr>
<td>New Job</td>
<td>0.004</td>
<td>0.233</td>
<td>0.560</td>
<td>-0.002</td>
<td>0.730</td>
<td>0.837</td>
<td>0.009</td>
<td>0.048</td>
<td>0.339</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>0.015</td>
<td>0.003</td>
<td>0.105</td>
<td>0.003</td>
<td>0.672</td>
<td>0.756</td>
<td>0.026</td>
<td>0.000</td>
<td>0.023</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Wild p-values based on 1499 bootstrap replications. All regressions have the full set of control variables listed in table 3.2.
Figure 3.1: Weeks Employed by Age Group

Figure 3.2: Weeks Not in the Labour Force by Age Group

Figure 3.3: Employment Rate by Age Group

Chapter 4

On the Effectiveness of Provincial Graduate Retention Programs

Many countries are facing the prospect of skill shortages in coming decades. The combination of aging baby-boomers and increasing demands for a well-educated labour force may necessitate new policies to increase aggregate education levels. Within Canada, several provinces have started to offer generous tax credits in the form of Graduate Retention Programs to encourage the settlement of recent graduates. The nature of these programs represents an intervention in migration and educational attainment decisions, which are analyzed in this chapter. This chapter provides the first analysis of the effectiveness of these costly programs. I find no compelling evidence that these programs have attracted more graduates to the offering provinces, although I do find some evidence that these programs dissuade current post-secondary students from dropping out. This chapter is also the first to implement the wild cluster bootstrap-t technique for improved inference with Difference-in-differences when the number of clusters is small, which was proposed in chapter 2.
4.1 Introduction

In recent years, several jurisdictions in North America have become concerned with the future of the local labour force; in particular, there is concern over a shortage of skilled labour. For example, several Canadian provinces have expressed such a concern brought about in part by the retirement of baby-boomers, and in part by low historical educational attainment levels. These provinces have started to offer tax credits, known as Graduate Retention Credits or Rebates, to recent post-secondary education graduates residing within the province. These programs are quite costly, and their impacts have yet to be analyzed. The four provinces which have implemented these graduate retention programs (herein referred to as GRP) are Saskatchewan, Manitoba, New Brunswick, and Nova Scotia. In the GRP provinces, excepting Nova Scotia, the rates of post-secondary graduation are below the national average.\(^1\)

Immigration reform could possibly alleviate this problem, but the provinces’ lack direct influence on the number and characteristics of immigrants to a particular province. Labour mobility in Canada means that provinces are free to compete with one another for talent within Canada. Historically, labour mobility has worked against these provinces, as residents have left for higher wages in other provinces. In a study using longitudinal administrative data, Bernard and St-Jean. (2008) found that emigrants from all of the GRP provinces experienced above-average wage increases compared to emigrants from all other provinces except Quebec and the other Atlantic provinces. With individuals exiting the labour force due to retirement, some jurisdictions are worried that a contracting labour force will hamper future economic

\(^1\)Estimates from the Labour Force Survey show the share of individuals aged 40-49 with a post-secondary education is significantly below the average in Manitoba, Saskatchewan, and New Brunswick, while the share in Nova Scotia is highest amongst all provinces.
growth or possibly lead to economic contraction (The Saskatchewan Labour Market Commission, 2009). The Canadian Department of Finance released a report which suggested that the growth rate of the labour force in Canada is expected to fall from 1% to 0.5% from 2016-2030.²

These GRP programs have similar aims to Merit Scholarships being offered in several American states.³ The Georgia Hope Scholarship program is one of the better known Merit Scholarship programs.⁴ It offers scholarships to students from Georgia who go to college in Georgia and has been studied by Dynarski (2000) and many others. It is similar to the retention programs in that it ties educational funding to a geographic location, but it does not restrict where recipients can reside post-graduation. Both the merit scholarships and GRP attempt to raise the average level of educational attainment within a jurisdiction. The Merit Scholarships offer incentives to study in a specified location, whereas the GRP offer incentives to reside in a specified location after graduation.

Aside from direct concerns about having a sufficiently skilled labour force in the future, these provinces may have other reasons to desire a better-educated populace. There is speculation that many new jobs will require higher levels of education. According to the United States Bureau of Labor Statistics, job growth in occupations requiring some post-secondary education is expected to outpace job growth in occupations requiring a high school education or less over the coming decade. (Bureau of Labour Statistics, U.S. Department of Labor, 2012) Additionally, the post-secondary education (PSE) spillovers literature presents evidence that higher rates of PSE have

²See Department of Finance Canada (2012) for details.
³Fifteen states now offer some form of merit scholarship, the effectiveness of these programs has been studied by Fitzpatrick and Jones (2012).
⁴Some of the states offering merit scholarships extend eligibility to National Merit Scholars from other states as well.
benefits for the community at large. For example, in a study looking at the returns to education that accrue from minimum schooling laws, Acemoglu and Angrist (2001) find evidence that there are small positive external returns to an extra year of education. Moretti (2004) uses the presence of a land grant university as an instrumental variable to examine the impact of increasing the share of college graduates in a city. He shows that a percentage point increase in presence of college graduates is associated with increased wages for others: a 1.9% wage increase for high-school dropouts and a 1.6% increase for high school graduates. In a similarly designed study, Shapiro (2006) examines American data at the metropolitan level and finds that a 10% increase in a city’s concentration of college graduates was followed by a 0.8% increase in employment growth. Using American data, Aghion et al. (2009) find that investment in four-year college educations has a positive effect on economic growth.

The pattern of internal migration has been studied in detail, with particular attention paid to the migration patterns of college-educated individuals or couples. In examining the locational choice of college-educated individuals, Costa and Kahn (2000) find that increasing shares of college-educated couples are locating in large metropolitan areas, which they attribute to a co-location problem. In addressing the same question, Compton and Pollak (2007) argue that the over-representation of college-educated couples in large cities is a result of couples matching in large cities, rather than from couples migrating to large cities. Either way, both studies shed light on the fact that there is a tendency for the college educated to move to large cities and to match with other college-educated individuals in a dating or marriage market. In examining life cycle migration, Chen and Rosenthal (2008) show that cities with improving business conditions acquire more workers. A recent study by

---

5 This study looked at additional years of high school as opposed to post-secondary education.
Albouy, Leibovici and Warman (2013) has estimated that the business conditions in major cities in the GRP provinces are amongst the lowest across Canadian cities.\textsuperscript{6} These findings provide suggestive evidence as to why the GRP provinces have started offering retention subsidies. Whether the amounts on offer are sufficient to change recent graduates’ migration decisions is an empirical question that this chapter hopes to answer.

Regardless of whether they are successful in altering migration decisions, the GRPs could otherwise be deemed successful if they encourage more individuals to enroll at, and subsequently graduate from, a post-secondary institute. After all, the average level of education in a province can be improved along the intensive and extensive margins. The educational choice literature is well studied, and Sartarelli (2011) provides a recent survey. The choice to obtain a post-secondary education is considered briefly in section 4.3, which provides a simple model to guide considerations about the theoretical impacts of graduate retention programs. Ignoring general equilibrium effects about the relative supply of post-secondary graduates, we should expect that a program which reduces the net present costs of obtaining a post-secondary education will, on average, induce more individuals to obtain more education.

Offering subsidies to recent graduates may seem a rather indirect route to increase the average educational level in a province; however, the provincial and federal governments invest heavily in both post-secondary institutions and students. Essaji and Neill (2010) provide a summary of the characteristics and costs of the various student funding programs currently in place in Canada, and thoroughly summarize the graduate retention programs. In addition to the programs mentioned in that paper, \textsuperscript{6}Here business conditions are proxied by productivity and measured by above-average local wages after controlling for observable characteristics such as education, experience, industry, and occupation.
several provinces have recently adjusted their approach to funding post-secondary education. For instance, in Ontario there is a new tuition credit, under which students are eligible for a 30% rebate of their Ontario college or university tuition.\textsuperscript{7} Additionally, some post-secondary funding decisions have gathered a great deal of attention from stakeholders: the 2012 proposed tuition increases in Quebec resulted in student strikes lasting for over six months.

While most education funding for students has been offered with few geographical constraints (aside from the differing tuition for foreign versus domestic students), there have been a few exceptions. Some jurisdictions, including the Atlantic provinces, have experimented with targeted retention/attraction programs to attract people in certain occupations, such as doctors and nurses (Reamy, 1994). Recently, the state government in Kansas has started to offer incentives such as student debt repayments and income tax waivers to attract individuals to rural Kansas. Unlike the graduate retention programs, to be eligible individuals must prove they have resided outside of the state for at least the previous five years. Since this program started in the 2012 tax year, its effects have yet to be studied.\textsuperscript{8} To my knowledge, the programs under study in this chapter are the first large scale, non-targeted programs aimed at attracting and retaining recent graduates from all disciplines.

The remainder of this chapter is organized as follows. Section 4.2 gives a detailed overview of the various graduate retention programs under study. Section 4.3 provides a simple model to consider the potential impact of the programs. Section 4.4 describes the estimation strategy and motivates the methodology for inference. Section 4.5

\textsuperscript{7}The grant tops out at $1680 for university tuition and $770 for college tuition, provided that their parents earn $160,000 or less per year, see \url{https://osap.gov.on.ca/OSAPPortal/en/PostsecondaryEducation/Tuition/index.htm} for more details.

\textsuperscript{8}For more information of the Kansas program see \url{http://http://www.kansascommerce.com/index.aspx?nid=320}
describes the four data sources. Section 4.6 discusses the results, which provide little support for the effectiveness of these programs, and section 4.7 concludes.

4.2 Background on the Graduate Retention Programs

The four provinces in Canada that have implemented graduate retention programs are Saskatchewan, Manitoba, New Brunswick, and Nova Scotia.\textsuperscript{9} Table 4.1 provides a quick overview of the various program attributes and how they differ across provinces.\textsuperscript{10} Broadly speaking, the programs are quite similar, and the amounts they offer to PSE graduates are of the same order of magnitude. The programs all act as tax credits, though the properties of the credits vary. With one exception none of the credits is refundable, though most of the credits do roll over.\textsuperscript{11} Alberta and Saskatchewan do not require a separate application for the graduate retention credits, and the credits can be claimed on income tax returns. Nova Scotia and New Brunswick require a separate application to claim the credits.

Three of the four programs determine the maximum individual retention credit based on the amount of tuition paid, while Nova Scotia offers a fixed amount to each

\textsuperscript{9}See Essaji and Neill (2010) for a summary of the various GRP programs.

\textsuperscript{10}Quebec also operates a smaller wage subsidy program for people in remote, resource-rich regions who work in the resource industry, with eligibility contingent on holding a degree related to your current occupation. Given the specificity of this program, it is ignored in the analysis. For more details about the Quebec program see \texttt{http://www.revenuquebec.ca/en/citoyen/credits/credits/credits_reduisant/nouv_diplome/}

\textsuperscript{11}Specifically, until 2012 Saskatchewan offered a refundable credit, which meant that if individuals did not earn sufficient income to claim the maximum annual amount, they were able to receive the difference in the form of a refund. In 2012, Saskatchewan moved to end the refundability of the tax credit, opting instead to make it possible to roll over the credit. All but Nova Scotia offer a roll over provision, which means that eventually an individual will be able to claim the maximum allowable credit, provided they owe taxes each year.
recent graduate. The proportion of tuition refunded in each province also varies, with up to 100% of tuition being refundable in Saskatchewan and 50% being refundable in New Brunswick. The maximum amounts of the credits are the same in both of these provinces; however, given the differing tuition refund percentages, a student would have had to pay $40,000 in tuition to receive the maximum credit in New Brunswick, but would have only had to pay $20,000 to receive the maximum credit in Saskatchewan. Finally, the total cost of the programs is largely similar in each of the provinces, ranging from $24-$35 million per year. These figures are estimated to increase over the coming years, as many of these programs are new and there are not yet six or seven cohorts of graduates claiming these credits.

The graduate retention programs are distinct for several reasons when compared with the other means of government funding for post-secondary education. One unique feature is that their benefits are provided solely after graduation. The budget constraints students face while in school are a current concern in educational funding policy since it has been shown that increased financial aid improves college enrollment (Sartarelli, 2011). The GRP programs do nothing to ease this constraint, but instead they expand a student’s post-graduation budget. Offsetting this aspect is the fact that most student loans require repayment following graduation and the payouts from these programs coincide with the repayment schedule for student loans. Moreover, in most cases individuals need to have a sizable income to receive the full amount of the credit.\textsuperscript{12} Another distinct aspect of these programs is the linking of the subsidy to a locational decision. This is an unusual characteristic of education funding, but it does align the subsidy with the provinces’ goal to have a well-educated labour force.

\textsuperscript{12}The roll over provisions offered by many provinces means that the entire credit will eventually be received, but over a longer time horizon than for someone with a higher income.
rather than a large number of college and university students. In part, these programs may be a beneficial alternative to increasing funding to universities or to students, as historically many graduates have moved provinces after they graduate.\textsuperscript{13} Similarly, in their examination of American data Aghion et al. (2009) argue that investments in institutions in states far away from the technology frontier\textsuperscript{14} tend to benefit states closer to the frontier, as graduates of said institutions tend to migrate toward the frontier.

4.3 A Simple Model for Educational and Migration Decisions

Having explained how these programs operate institutionally, it is worth considering the theoretical influence they may have on an individual’s decisions. Let us consider the situation of a student who has just finished high school. The individual wants to maximize her utility, which we assume she is able to do so by choosing only two things: whether to pursue additional education and which city to reside in. Cities influence the quality of life (QoL) that the individual enjoys as well as the earnings they receive and the taxes they pay.\textsuperscript{15} Conceptually, we can imagine this individual solving the following problem:

\textsuperscript{13}The Saskatchewan government mentions that there “has been significant leakage of post-secondary graduates outside the province” (The Saskatchewan Labour Market Commission, 2009).

\textsuperscript{14}The authors define a ‘close-to-the-frontier’ state as one where industries are more likely to depend on innovation for economic growth.

\textsuperscript{15}Taxes in Canada are determined at the provincial and federal levels. In choosing a city, individuals will also determine which provincial tax schedule they are subject to.
\[
\max_{\text{educ} \in \{hs,c,u\}, \text{city}_1,2 \in X} \ U_1(C_1, QoL_1) + \beta U_2(C_2, QoL_2),
\]
such that
\[
QoL_i = f(\text{city}_i),
\]
\[
C_1 = Y_1(\text{educ}, \text{city}_1) + SL(\text{educ}) - \text{taxes}(\text{city}_1) - \text{rent}(\text{city}_1) - \text{tuition}(\text{educ}),
\]
\[
C_2 = Y_2(\text{educ}, \text{city}_2) - (1 + r) \cdot SL(\text{educ}) - \text{taxes}(\text{city}_2, \text{educ}) - \text{rent}(\text{city}_2).
\]

The problem is to maximize her utility across two periods, one in which she possibly continues her education, and the subsequent period. Utility in the second period is discounted by an exogenously given discount factor, \(\beta\). To make things easier, assume that she only has the following educational choices: she can choose no further education, and remain a high school graduate (hs), she can attend a community college (c) or she can attend university (u). Additionally, she can choose the city in which she will either work or attend school from the set of cities \(X\).

In the first period, the educational decision will affect consumption in several ways. Obtaining additional education requires a good deal of time, and as such will negatively impact earnings in the first period. Additional education has a direct tuition cost, which must be paid in the first period. Student loans (SL) are available to help pay tuition and provide additional consumption in the first period if additional education is chosen, but they must be repaid with interest charged at rate, \(r\), in the second period. The individual is free to relocate to a new city in the second period. Education is assumed to increase earnings in the second period. The graduate retention programs introduce an education argument to the tax function in the second period.
period, as in certain cities the individual will face lower taxes, given earnings, than a similar person with a lower level of education on account of receiving a graduate retention tax credit. Rent is included in the model to reflect the fact that there are some non-traded goods that people must consume, the price of which is determined by the city of residence.

Reflecting on the impact of the graduate retention programs through the lens of this model reveals several dimensions in which the program might have an impact. The most direct impact is that consumption in the second period is higher in a city where an individual is eligible for a GRP, with all else being equal. Accordingly, if wages, taxes, and quality of life are in fact equal between two cities, then if one offered a GRP and the other did not, we should expect the city with the GRP to be the chosen residence. Moreover, the presence of a GRP program will increase overall consumption in the two periods for those who go to school, so we should expect that a GRP program will increase educational attainment. This follows from the fact that if a location decides to offer a GRP, overall two-period consumption is then higher for graduates than it was prior to the introduction of the GRP. This increase in consumption makes obtaining more education relatively more attractive. A third possible impact may occur because the decrease in taxes from a GRP credit coincides with student loan repayments. If there is heterogeneity in the amount of student loans outstanding, then it is probable that those with larger amounts of debt would be more likely to locate in a GRP city.

Although not modeled explicitly, the graduate retention programs have one additional channel of influence. If people have already decided to enroll in school in expectation of receiving a GRP upon graduation, the consequences of dropping out
of school are more severe than they would have been otherwise. Dropping out of post-secondary education is not a rare occurrence: in Ontario the graduation rate for colleges was 65% in 2011, (Colleges Ontario, 2012) while the overall graduation rate at Ontario universities is closer to 80%.\footnote{Author’s approximation using data taken from an online tool at the Council of Ontario University’s website \url{http://cudo.cou.on.ca/} } In the absence of a GRP credit, the potential penalty for dropping out is missing out on the higher wage one might earn after graduating. The presence of the GRP increases the penalty of dropping out by disqualifying an individual from receiving the graduate retention tax credit.

The model above is framed in terms of an individual’s decision, which matches well with most of the data used in this chapter. However, if we consider that there are many individuals making these decisions and perhaps being influenced by the presence of graduate retention programs, then we should expect to see the presence of these influences in the aggregate data. If we accept as plausible the three potential impacts – the increased educational attainment, the increased school enrollment, and the decreased dropout probability – then we should expect these impacts to be observable amongst cohorts of a given age.

\section{4.4 Estimation Strategy}

To analyze the impact of the GRP, a difference-in-differences (DiD) approach will be used. In the analysis, the provinces offering graduate retention programs listed in table 4.1 will be regarded as the ‘treated’ group, while the other provinces will be regarded as the ‘comparison’ group. For the DiD estimates to work, there needs to be common support between the treatment and comparison groups, in addition
to common trends prior to the introduction of the programs in both groups, see DiNardo and Lee (2011) for additional details. The primary analysis will look at the impact of the programs in the Atlantic Provinces, namely: Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland and Labrador. The Atlantic provinces offer a useful framework for conventional DiD analysis, as the Atlantic region is comprised of two provinces with retention programs, Nova Scotia and New Brunswick, and two provinces without, Prince Edward Island and Newfoundland and Labrador. There is the additional benefit of interprovincial migration being nearly symmetric within the Atlantic Region.\textsuperscript{17} Accordingly, the assumption of common support is more realistic when looking at policies within the Atlantic provinces. If the programs have a substantial impact on migration, the estimates will perhaps be an upper bound of the impact. For example, in analyzing the proportion of university graduates within a province, a graduate who moved from PEI to Nova Scotia would simultaneously increase the stock of graduates in Nova Scotia and reduce the stock in PEI. The analysis for the Atlantic provinces will be estimated by the following equation:

\[
Y_{istm} = c + \beta_{GRP} * I[ProvGRP_{istm} * YearGRP_{istm}] \\
+ PROV_{s} + YEAR_{t} + I[PROV_{s}] * YEAR_{t} + MONTH_{m} + AGES_{istm} + \epsilon_{istm}.
\]

(4.1)

The data varies along four dimensions, as the variable $Y_{istm}$ contains an observation for individual $i$, in province $s$, in year $t$, in month $m$. Not all of the datasets have monthly data; in the sets that do not, the monthly identifiers are dropped. In the

\textsuperscript{17}See Statistics Canada, CANSIM table 051-0019.
equation there is a set of province dummy variables, a set of year dummy variables, a set of province specific time trends, and a set of month dummy variables. Aside from age dummies, there are no control variables because the unconditional effect is of primary importance. The coefficient of interest is $\beta_{GRP}$ which will capture the impact on those individuals living in GRP provinces in GRP years. This variable will be equal to one for those in Nova Scotia in 2006-2013, those in New Brunswick in 2005-2013, and set to zero otherwise. This equation estimates a common treatment effect for the two provinces offering graduate retention credits.

Following the model, there are two primary margins on which the graduate retention programs could have an impact. The most direct impact would be if they were to alter individuals’ migration decisions. That is, if the programs made those currently in GRP provinces think twice about moving to a non-GRP province, or nudged those in non-GRP provinces to relocate to a GRP province. To examine the most direct impact, migration, it would be ideal to look at migration decisions directly. Unfortunately, given the data used in this chapter, there is very limited information available on interprovincial migration. The Survey of Labour and Income Dynamics records whether individuals have moved provinces within the last calendar year, and this variable is analyzed directly. It should be noted that it is difficult to establish the expected sign of this coefficient. The retention programs could make individuals both less likely to emigrate from a GRP province, which would result in a negative sign, and make individuals more likely to immigrate to a GRP province, which would result in a positive sign. For these reasons, in addition to looking at migration directly, the aggregate proportion of individuals within a cohort with a given level of education
will be analyzed. These aggregate levels or aggregate proportions have many analytical advantages. The first of these is they correspond well with the model since there are many people making educational decisions within each of the provinces. Additionally, if the programs are influential in either attraction or retention, we should see an increase in the concentration of PSE graduates within the GRP provinces. A number of different educational attainment variables are analyzed in an attempt to quantify the programs’ impacts, specifics of which are discussed in greater detail in section 4.5.

The second major impact they might have is on the educational attainment decisions made by those currently in the GRP provinces. If high-school-aged individuals in the GRP provinces are sufficiently forward-looking, then they might view the retention credits as large subsidies to education, and thus on average obtain more education than they would have otherwise. This might appear in the data by increasing the education level of those who were initially tempted by GRP several years ago, and by increasing the school enrollment rates for the younger cohorts. The data sets contain information of whether individuals are currently enrolled in school, or if they have ever been enrolled in college or university. We can thus estimate the concentration of current and former students in a given province.

While equation (4.1) describes the estimation strategy to be used in this analysis, particular attention needs to be given to the method of inference. Two of the four data sets used in this chapter use individual data in which there are a great number of individuals sampled in each of the provinces. Many of the variables used in equation (4.1) are invariant within cluster, and this causes a violation of the i.i.d. errors assumption required for OLS. The OLS standard errors are inappropriate for
inference as the error terms are no longer independently distributed. The within cluster correlation problem causes standard errors to be too small, and inference in this environment is particularly challenging as there are only four clusters. This problem is the focus of chapter 2, and the small cluster problem was addressed by proposing the following 6-point bootstrap weight distribution:

\[ v_g = -\sqrt{\frac{3}{2}}, -\sqrt{\frac{2}{2}}, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{2}{2}}, \sqrt{\frac{3}{2}} \quad \text{w.p.} \quad \frac{1}{6}. \]

This chapter is the first empirical application of the 6-point wild cluster bootstrap-t technique. Throughout the analysis, 399 bootstrap replications are performed. Additional work by MacKinnon and Webb (2013) has suggested that variation in cluster sizes can also cause a problem for inference. The combination of a small number of clusters and variable cluster sizes will cause problems for the cluster robust variance estimator. A small Monte Carlo simulation calculates the rejection frequency using the 6-point wild bootstrap using simulated data constructed to match the cluster structure of the data used in this chapter. The setup for the simulations run here are very similar to those in chapter 2. The major distinction is that the simulations are done with clusters of unequal size. The four clusters have observations proportional to the populations of the four Atlantic provinces, where Nova Scotia is 39.90% of the sample, New Brunswick is 31.88%, Newfoundland and Labrador is 22.13%, and Prince Edward Island is 6.09%. This Monte Carlo experiment was run with 1000 replications and a sample size of 2500 observations. The result of this simulation is an estimated rejection frequency of 4.7%, at the 5% level. This result suggests that the wild cluster bootstrap works well with only four, unequal sized clusters.
4.5 Data

Data from four different data sets are used. Two of the data sets should be regarded as primary, while the others should be thought of as providing robustness checks. In general, there are data collected from individuals, institutions, and a census of the institutions. The various data sets all have different advantages and limitations, which are discussed in Appendix B. Survey weights are used throughout the analysis to enhance the comparability of the sample to the Canadian population.

Table 4.2 shows the means of all the variables used in the analysis. These are presented for both the pre and post period, and the GRP and non-GRP provinces. The LFS variables are all binary variables that indicate whether the person has obtained a certain level of education, has enrolled in school or whether the person has dropped out. Thus the estimation strategy is very flexible. The variables are calculated for the entire age subsample that is being estimated, and as such can be regarded as provincial averages for the various indicators. While most of the variables are calculated independently for college or university education levels, data limitations require that some of the variables be calculated jointly for either college or university. Note that within the treatment and comparison groups the means are increasing over time, and the regression analysis will determine whether the differences in each group are in fact statistically different from each other. The gender ratio is quite similar in both the GRP provinces and the non-GRP provinces. However, the percentage of the sample that is single is lower in the GRP provinces.

The variables university ever attended or college ever attended are set to one if an individual was ever enrolled in university or college. For young individuals, these variables can be informative of whether an individual is a student, and for older
individuals they can be used to make inferences about whether individuals dropped
out of post secondary education. The variable STEM or Business university major is
used to examine whether the programs had any impact on the majors that individuals
chose while at university. The annual panel nature of the SLID allows for a level
of analysis not possible with the LFS. In particular, movements between provinces
can be observed. The variable moved province is set to one for individuals who have
moved provinces in the previous twelve months. A similar indicator for PSE student
in same province as high school records whether someone is currently in university
or college, and whether their PSE institution is in the same province as where they
attended high school. It is difficult to assign a prior to the sign of the DiD coefficient
on this variable. If more students within a GRP province go to school in province
then it would be positive, although if more students from other provinces went to
school in GRP provinces then it would be negative.

The annual nature of the data also allows us to analyze what individuals do in
their high school graduation year. Direct to College indicates whether an individual
who graduated high school in the current year enrolled in college in the same year.
Similarly, Direct to University indicates whether an individual who graduated high
school in the current year enrolled in university in the same year. In line with the
LFS, gender is used for subsample analysis. Within the SLID the gender ratio is
fairly balanced, and the percentage of individuals who are single is lower in the GRP
provinces. The parent with university degree variable is used as a subsample variable
in the regressions. This variable is used for two reasons, one being to capture the

---

18 The specific majors classified as being either science, technology, engineering math or business
also includes medical, dental, account and architectural majors.
19 To determine whether graduates have moved this variable is calculated for the subsample of
people who have a highest level of education of a college diploma or above.
inter-generational persistence in educational attainment, and the other as a proxy for family economic resources.

The public use PSIS reports levels of enrollment for each province and breaks the enrollment numbers down by province of origin as well as international student status. For the purposes of the analysis, the enrollment levels have been changed into annual growth rates in the levels. From the PSIS tables on University Enrollment, I calculate the following variables to examine how university enrollment varied by province: 1) From Province records how many individuals from a given province are enrolled in university in Canada, that is to say how many Albertans are enrolled in universities across Canada. 2) Domestic in Province records how many Canadian students go to school in a given province. 3) Within Own Province records how many people from a given province go to university in their home province, i.e. how many Albertans go to university within Alberta. 4) Total Enrollment in Province records how many students, foreign and domestic, from any province, go to school in any province. The regression estimates of $\beta_{GRP}$ from equation (4.1) will be presented as both pooled estimates and estimates on a province by province basis to capture heterogeneous treatment effects.

The Maclean’s variable measures the retention rate for each university in the sample. The retention rate reports the proportion of first year full time students who re-enroll at the same university the following school year. This variable is a good proxy for overall graduation rates as the majority of individuals who do not finish university leave before their second year of university.\textsuperscript{20} The estimates here are also presented as both pooled estimates and estimates on a province by province basis.

\textsuperscript{20}See the Globe and Mail focus on ‘Our Time to Lead: Education’ from October 6, 2012.
4.6 Results

Table 4.3 estimates for the full sample estimates of the impact of the GRPs in the Atlantic Provinces, and compares procedures for inference by presenting four alternative p-values. The coefficients themselves are quite surprising, as they are in general quite small. The estimates here are per 1000 individuals, so the range of these coefficients is from -8.13% to 4.65%. Moreover, the coefficients are often estimated to be in the opposite direction of what would be expected if the programs were having an impact. In particular, there is an estimated decrease in the prevalence of university, college and PSE grads using the LFS data set. The university student and college student variables within the LFS suggest that there was a decrease in the prevalence of college students, with an almost opposite increase in the prevalence of university students. A similar pattern exists in the SLID results, with the proportion of university graduates increasing, and the proportion of college graduates decreasing. Meanwhile, the likelihood of a student going directly from high school to university decreased, while the likelihood of going to directly to college increased. Interestingly, the proportion of within province PSE students decreased, as did the proportion of university grads with a STEM or business degree. These results, taken together, suggest that it is unclear whether these programs have had the desired effects.

As mentioned in the methodology section, inference using a difference-in-differences methodology requires special considerations. The problem of small clusters, discussed above, is particularly pronounced in this analysis using the Atlantic provinces, as there are only four clusters. It is natural to treat the province as the primary sampling unit here as the retention programs that are being analyzed operate at the provincial level. The table shows that the implication of the analysis varies greatly depending
on which p-values are used. The four p-values that are considered are OLS, CRVE $N(0,1)$, CRVE $T(G-1)$, and the wild bootstrap. The OLS p-values are calculated using robust standard errors, and assume that the t-statistics follow a normal $(0,1)$ distribution. Both CRVE p-values use cluster robust standard errors, and assume that the t-statistics follow either a normal $(0,1)$ distribution, or a t-distribution with $G-1$ degrees of freedom. The wild bootstrap p-value follows the procedure discussed in chapter 2. These p-values were all considered in the Monte Carlo simulations in chapter 2. The Monte Carlo results suggest that the OLS p-values and both CRVE p-values are on average too small, while the bootstrap p-values are on average closer to the correct size. Monte Carlo simulations suggest that the observed CRVE rejection frequencies are roughly double the desired 5% level. If the results from those simulations are applicable to the estimates being made here, then we should expect to see the p-values increasing as we move from left to right in the table. Indeed, this pattern generally holds in both the LFS and the SLID panels.

Within the LFS, if one were to only look at the OLS p-values, then one would erroneously conclude that many of the estimates are highly significant. The CRVE p-values are an improvement, especially when a t-distribution is assumed. However, relying on CRVE p-values would still result in many coefficients erroneously being regarded as significant. This is also the case within the SLID, where the t-distribution increases several p-values from 0.000 to the 0.02-0.04 range. The wild bootstrap p-values are generally larger than the CRVE p-values, and hardly any of the estimates are significant at even the 10% level.\footnote{We need not be concerned with cases when the bootstrap p-value is lower than the CRVE p-value in cases when both are highly statistically insignificant.} Given that the other tests are over-sized and that the bootstrap tests are close the appropriate size, we can regard the bootstrap
p-values as being closer to the ‘true’ p-values. As this table shows, ignoring the small cluster problem would lead to inaccurate inference, and as a result only the bootstrap p-values will be displayed in the following tables.

Examining the results of this table with the bootstrap p-values as a means of inference paints a rather dim picture on the effectiveness of the graduate retention programs. There is only one coefficient estimated to be significant at the 5% level. This could well be due to chance, as 18 separate regressions are being run, so there is the matter of multiple testing to bear in mind when interpreting all of the results being presented. This sole significant coefficient estimates that the concentration of post secondary students attending school in the same province as their high school has decreased. A potential explanation for this is that the concentration of out of province students has increased in the GRP provinces, though it is unclear whether these students moved on their own or with their parents. If we relax the stringency of the analysis and examine coefficients at the 10% level, there are only two additional significant results. There is a decrease in the likelihood of a university graduate being a STEM or business major. Table 4.2 shows that there was a large increase in these majors in the non-GRP provinces between the pre and post periods, and only a slight decrease in the GRP provinces. This effect is interesting, but it is likely not being caused by the programs.

Finally, the concentration of PSE grads in the GRP provinces is estimated to decrease slightly using the LFS data. This decrease exists despite a lack of significance for either the concentration of university grads or college grads independently. Despite how small this decline is, it is clearly in the opposite direction of what would
be expected if these programs were working as intended. In addressing the three predictions from the model, that if effective the GRP should increase the concentration of PSE graduates, increase school enrolments, and decrease dropout rates, there are no statistically significant effects to support those predictions.

Having examined the full sample results, it is interesting to see if there are any effects within any specific subsamples. Table 4.4 reports the results of the pooled DiD estimates for the Atlantic provinces, specified in equation 4.1. The first column of this table is identical to the first column in table 4.3. The most notable result from this table is that almost every coefficient is statistically insignificant. Given that 18 variables are estimated across the full sample and four subsamples, 90 coefficients are displayed in this table. By chance alone we should expect that 4-5 coefficients should be significant at the 5% level. The programs seem to have not been effective in achieving their primary goals of increasing the prevalence of either college or university graduates. The coefficient on the combined PSE graduate concentration variable is estimated to be $-17.881$ for females and is significant at the 5% level, which is closely aligned with the full sample estimate of $-18.059$ which is significant at the 10% level. Otherwise, there are no significant impacts on the concentration of graduates.

In regards to the subsamples themselves, there are some interesting patterns. In general there are more significant effects for females than for males in both data sets. The coefficients for singles are all highly insignificant, while there are some significant coefficients for married individuals. The prevalence of married university graduates is estimated to decrease by 4.4% under the programs with a p-value of 6%. Additionally, married individuals were more likely to be university students and less likely to be college students in the years and provinces in which the programs were operational.
In comparing the results for those who have a parent with a bachelor’s degree and those that do not, we can see that the individuals with more educated parents are more likely to go directly to either college or university, though none of these effects are remotely significant. Additionally, the children of more highly educated parents are less likely to drop out of college or university, though this is also not significant. Females in the SLID were estimated to be 4.4% less likely to drop out of college under the program period, with a p-value of 3.2%. From this table it is difficult to say with certainty if the GRP programs had any real impact, though the programs certainly have not had the desired impact of increasing the presence of graduates and students in both college and university in the GRP provinces.

Table 4.5 reports the results obtained using both the PSIS data and the Maclean’s retention rate. The table reports both the pooled and individual estimates for both the Atlantic provinces and all provinces. Here the p-values being used for inference are calculated using heteroskedasticity robust standard errors.\textsuperscript{22} The top panel of the table reports the results for the Atlantic provinces. All of the estimates for the PSIS variables are highly insignificant in both the pooled and individual regressions. Recall that the PSIS variables are reported as year over year growth rates. Interestingly, the coefficients for \textit{From Province} and \textit{Within Province} are estimated to be positive. The bottom panel reports the estimates using all provinces and many of the coefficients are estimated to be statistically significant. The pooled estimates are quite large for three of the four PSIS variables and estimated to be significant at the 5% level. However, within the heterogeneous estimates, the only province with a statistically significant impact is Manitoba.\textsuperscript{23}

\textsuperscript{22}CRVE standard errors are not used as the PSIS data only contains province level data, and the Maclean’s data has very few schools in each province.

\textsuperscript{23}In the raw data there is a significant drop in all of the Manitoba variables in 2006, but the
The coefficients on the retention rate variable are also insignificant in the Atlantic provinces regressions, though the coefficients for all provinces tell an interesting story. The pooled estimate suggests that at the universities within GRP provinces, the retention rate increased by 2.3 percentage points during the post GRP period compared to the change observed in non-GRP provinces. The means presented in table 4.2 suggest that this effect comes from the retention rates falling in non-GRP provinces, while they held steady in GRP provinces. In looking at the individual province estimates we observe that three of them are positive and significant at the 10% level. In New Brunswick, the estimate is negative, but insignificant, while it is positive and insignificant in Nova Scotia. The estimates are sizeable and significant, with p-values of 0.01 for Saskatchewan and 0.11 for Manitoba. As mentioned above, this retention rate is a proxy for overall university dropout rates, and it might be the case that the graduate retention programs are having an impact on the frequency with which people return for a second year of university. Aside from the increase in student retention within the GRP provinces, the PSIS estimates suggest that there is little evidence that the programs are encouraging students from the GRP provinces to enroll in university. Interestingly, this lack of effectiveness is consistent between measures of the sources and destinations of university students.

As a robustness check, an additional set of regressions were estimated. This is equivalent to a falsification test in which it is assumed that the programs came into effect in 2002 in New Brunswick and 2003 in Nova Scotia. These estimates are very similar to those presented in equation (4.1). However, these estimates use only variables all return to their previous levels in 2007. When the analysis is re-done with a dummy variable in place for the 2007 growth rate in Manitoba, the p-values of the pooled estimates for from province, domestic, and within province, are on the order of 8%. The Manitoba coefficients on these same variables are smaller, yet highly significant. It is possible that the 2006 observation is a coding error, but given the source of the data the results without the dummy variable are presented here.
the data from 2000-2004. Here the ‘treatment’ variable is set to one for those in Nova Scotia in 2003 and 2004, and those in New Brunswick from 2002-2004.\textsuperscript{24} This period is the pre-program period where no provinces are treated. These estimates can be regarded as a hybrid falsification-common trend test. The regressions estimate whether the GRP provinces had a different experience in 2003-2004 compared to the non-GRP provinces.

The results of these estimates can be found in table 4.6. If there is common support, we should expect to see no estimates to be either economically or statistically significant. Recall that the estimates are examining the impact per 1000 individuals. The LFS results are nearly ideal since most of the coefficients are quite small, especially for the full sample. There are a few larger coefficients, particularly for the male subsample in terms of increased college and PSE graduate concentrations. However, not one of these coefficients is estimated to be statistically significant at the 5% level. In regards to the SLID results, there are a few estimates which are economically significant, specifically a 39.2% decline amongst males majoring in science or business, though this is only significant at the 10% level. The choices made by recent high school graduates are also interesting, as it seems like there is a steep decline in males going directly to college, and a steep increase in females going directly to college. At the same time, there is a decrease in individuals with parents with a B.A. going to college, and an increase in these individuals going to university; however, these estimates are highly insignificant. Almost none of the SLID results are estimated to be statistically significant at 5%. Given that 55 separate regressions are estimated here, we should expect between two and three coefficients to be significant by chance at the 5% level. There are only two coefficients that are estimated to be significant

\textsuperscript{24}The unequal treatment periods matches the unequal treatment periods in the original estimates.
at the 5% level and none at the 1% level. In general, these estimates lend support to
the estimation strategy being used, as the null hypothesis of common trends is largely
not rejected.

\section{Conclusion}

In this chapter, I examine a set of programs to encourage recent post-secondary
graduates to settle in a specified province. The programs operate in four provinces
and offer generous tax credits to those who reside within the province after graduating.
The programs were put in place in an effort to mitigate potential skill shortages as
baby boomers begin to retire, and to curb a perceived outflow of recent graduates
to the largest Canadian cities. Overall, there is little evidence to suggest that these
programs have had the desired effect. In many cases, the estimated impact of the
program is in the opposite direction that one would expect, where the relative share of
the targeted population with a post secondary education has decreased in comparison
with the provinces not offering such retention programs. These estimates suggest that
the GRP provinces are not fulfilling their desire to increase the level of educational
attainment within their provinces.

The analysis performed on limited migration data suggests that the retention pro-
grams have had no impact on interprovincial migration decisions. There is some pre-
liminary evidence that the programs have decreased the share of university students
that leave university after one year of studies. The lack of evidence for a substan-
tial impact of these programs is not surprising given the rather limited value of the
credits in comparison with the potential gains from switching provinces, as well as
the lack of marginality in the programs’ designs. The programs are rather generous,
but historical wage gaps continue, and there are potentially large gains to be had by emigrating from a GRP province. The lack of marginality is important in two dimensions, since people are eligible for these credits regardless of whether they would have gone to post secondary in the absence of the program, and regardless of whether the program induced a change in their migration decision. This is in stark contrast to a similar program recently introduced in rural Kansas which requires people to demonstrate out of state residency to be eligible for a recent graduate tax credit.

If the primary purpose of these programs is to retain recent graduates within the province, it is difficult to conceive of a mechanism which efficiently sorts those who would have actually moved to another province were it not for the programs, from those who would merely claim to be part of that group to qualify for the tax credit. Given the annual cost of these programs, it is evident that their existence is fairly well known, though from the analysis conducted here it seems as though the majority of the benefits accrue to inframarginal individuals. Future research should attempt to analyze the share of benefits which go to marginal individuals, and perhaps assess some of the normative implications of the graduate retention programs.
<table>
<thead>
<tr>
<th></th>
<th>SK</th>
<th>MB</th>
<th>NS</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximum amount</strong></td>
<td>20k</td>
<td>25k</td>
<td>15k</td>
<td>20k</td>
</tr>
<tr>
<td><strong>Rebate per year</strong></td>
<td>10%, 20%</td>
<td>4k, 10%</td>
<td>2.5k</td>
<td>4k</td>
</tr>
<tr>
<td><strong>NPV ($000) @ 5%</strong></td>
<td>16.9</td>
<td>14.1</td>
<td>13.3</td>
<td>12.6</td>
</tr>
<tr>
<td><strong>Refundable credit</strong></td>
<td>Y*</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td><strong>Roll over credit</strong></td>
<td>N*</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Eligibility duration</strong></td>
<td>7</td>
<td>10</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td><strong>Application req.</strong></td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Tuition based</strong></td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td><strong>Tuit. % refunded</strong></td>
<td>100%</td>
<td>60%</td>
<td></td>
<td>50%</td>
</tr>
<tr>
<td><strong>Program costs</strong></td>
<td>35m</td>
<td>34m</td>
<td>25m</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** * reflects the change in 2012 that Saskatchewan announced, where the credit was no longer refundable, but would instead roll over from one year to the next. The NPV calculation assumes a 5% discount rate, sufficient earnings to get the maximum credit in all years, and $22,663 in tuition paid in earning a 4 year B.A. degree from Queen’s University.
Table 4.2: Variable Means from All Data Sets

<table>
<thead>
<tr>
<th></th>
<th>Provinces without GRP</th>
<th>Provinces with GRP</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>LFS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University Graduate</td>
<td>15.9%</td>
<td>20.2%</td>
<td>20.0%</td>
</tr>
<tr>
<td>College Graduate</td>
<td>40.7%</td>
<td>39.0%</td>
<td>37.6%</td>
</tr>
<tr>
<td>University or College Graduate</td>
<td>57.4%</td>
<td>59.5%</td>
<td>58.2%</td>
</tr>
<tr>
<td>University or College Dropout</td>
<td>10.2%</td>
<td>10.0%</td>
<td>9.9%</td>
</tr>
<tr>
<td>University Student</td>
<td>14.7%</td>
<td>15.7%</td>
<td>13.5%</td>
</tr>
<tr>
<td>College Student</td>
<td>6.9%</td>
<td>7.8%</td>
<td>5.1%</td>
</tr>
<tr>
<td>University or College Student</td>
<td>21.6%</td>
<td>23.5%</td>
<td>18.5%</td>
</tr>
<tr>
<td>Male</td>
<td>50.1%</td>
<td>50.1%</td>
<td>49.7%</td>
</tr>
<tr>
<td>Single</td>
<td>71.9%</td>
<td>72.6%</td>
<td>68.8%</td>
</tr>
<tr>
<td>SLID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University Ever Attended</td>
<td>35.1%</td>
<td>38.5%</td>
<td>40.0%</td>
</tr>
<tr>
<td>College Ever Attended</td>
<td>40.3%</td>
<td>41.6%</td>
<td>37.8%</td>
</tr>
<tr>
<td>University Graduate</td>
<td>8.8%</td>
<td>11.0%</td>
<td>10.9%</td>
</tr>
<tr>
<td>College Graduate</td>
<td>27.8%</td>
<td>27.2%</td>
<td>26.5%</td>
</tr>
<tr>
<td>University Dropout</td>
<td>8.4%</td>
<td>8.1%</td>
<td>11.8%</td>
</tr>
<tr>
<td>College Dropout</td>
<td>7.9%</td>
<td>7.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>High School to University</td>
<td>30.1%</td>
<td>35.9%</td>
<td>33.2%</td>
</tr>
<tr>
<td>High School to College</td>
<td>16.8%</td>
<td>23.7%</td>
<td>15.7%</td>
</tr>
<tr>
<td>Moved Province</td>
<td>3.0%</td>
<td>2.3%</td>
<td>2.7%</td>
</tr>
<tr>
<td>PSE Stdt in Prov. of High School</td>
<td>25.7%</td>
<td>25.2%</td>
<td>20.8%</td>
</tr>
<tr>
<td>STEM or Business Univ. Major</td>
<td>37.9%</td>
<td>51.9%</td>
<td>55.2%</td>
</tr>
<tr>
<td>Male</td>
<td>49.3%</td>
<td>50.5%</td>
<td>50.5%</td>
</tr>
<tr>
<td>Single</td>
<td>78.3%</td>
<td>81.7%</td>
<td>75.5%</td>
</tr>
<tr>
<td>Parent with University Degree</td>
<td>15.3%</td>
<td>22.7%</td>
<td>20.6%</td>
</tr>
<tr>
<td>PSIS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>From Province</td>
<td>6.02</td>
<td>2.27</td>
<td>-3.54</td>
</tr>
<tr>
<td>Domestic in Province</td>
<td>7.36</td>
<td>2.97</td>
<td>-4.25</td>
</tr>
<tr>
<td>Within Own Province</td>
<td>7.25</td>
<td>2.48</td>
<td>-4.63</td>
</tr>
<tr>
<td>Total Enrollment in Province</td>
<td>3.47</td>
<td>3.01</td>
<td>0.88</td>
</tr>
<tr>
<td>Maclean’s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention Rate</td>
<td>88.56</td>
<td>85.80</td>
<td>80.40</td>
</tr>
</tbody>
</table>

Sample: SLID years 2000-2010, LFS years 2000-2013, ages 17-29, unless otherwise noted. PSIS, growth rates for all provinces. Maclean’s all surveyed universities in all provinces.
## Table 4.3: Comparison of P-values for Full Sample Estimates

<table>
<thead>
<tr>
<th>LFS</th>
<th>$\beta_{GRP}$</th>
<th>OLS</th>
<th>CRVE N(0,1)</th>
<th>CRVE t(G-1)</th>
<th>Wild 6pt</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Graduate</td>
<td>-15.924</td>
<td>0.009</td>
<td>0.101</td>
<td>0.200</td>
<td>0.541</td>
</tr>
<tr>
<td>College Graduate</td>
<td>-0.197</td>
<td>0.978</td>
<td>0.979</td>
<td>0.980</td>
<td>0.957</td>
</tr>
<tr>
<td>PSE Graduate</td>
<td>-18.059</td>
<td>0.015</td>
<td>0.109</td>
<td>0.207</td>
<td>0.085</td>
</tr>
<tr>
<td>University Student</td>
<td>12.820</td>
<td>0.004</td>
<td>0.220</td>
<td>0.307</td>
<td>0.306</td>
</tr>
<tr>
<td>College Student</td>
<td>-12.909</td>
<td>0.001</td>
<td>0.093</td>
<td>0.192</td>
<td>0.236</td>
</tr>
<tr>
<td>PSE Student</td>
<td>11.227</td>
<td>0.000</td>
<td>0.077</td>
<td>0.175</td>
<td>0.531</td>
</tr>
<tr>
<td>PSE Dropout</td>
<td>-1.682</td>
<td>0.698</td>
<td>0.897</td>
<td>0.905</td>
<td>0.982</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLID</th>
<th>$\beta_{GRP}$</th>
<th>OLS</th>
<th>CRVE N(0,1)</th>
<th>CRVE t(G-1)</th>
<th>Wild 6pt</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Graduate</td>
<td>36.900</td>
<td>0.000</td>
<td>0.000</td>
<td>0.036</td>
<td>0.116</td>
</tr>
<tr>
<td>College Graduate</td>
<td>-27.534</td>
<td>0.399</td>
<td>0.307</td>
<td>0.382</td>
<td>0.592</td>
</tr>
<tr>
<td>University Dropout</td>
<td>-3.295</td>
<td>0.888</td>
<td>0.764</td>
<td>0.783</td>
<td>0.872</td>
</tr>
<tr>
<td>College Dropout</td>
<td>-6.735</td>
<td>0.714</td>
<td>0.657</td>
<td>0.687</td>
<td>0.872</td>
</tr>
<tr>
<td>University Ever Attended</td>
<td>-1.647</td>
<td>0.949</td>
<td>0.917</td>
<td>0.924</td>
<td>0.892</td>
</tr>
<tr>
<td>College Ever Attended</td>
<td>-0.299</td>
<td>0.990</td>
<td>0.950</td>
<td>0.954</td>
<td>0.984</td>
</tr>
<tr>
<td>PSE Stdt in Prov. of High School</td>
<td>-40.806</td>
<td>0.041</td>
<td>0.000</td>
<td>0.022</td>
<td>0.048</td>
</tr>
<tr>
<td>High School to University</td>
<td>-20.699</td>
<td>0.828</td>
<td>0.600</td>
<td>0.637</td>
<td>0.612</td>
</tr>
<tr>
<td>High School to College</td>
<td>46.495</td>
<td>0.562</td>
<td>0.304</td>
<td>0.379</td>
<td>0.436</td>
</tr>
<tr>
<td>Moved Province</td>
<td>8.570</td>
<td>0.633</td>
<td>0.472</td>
<td>0.524</td>
<td>0.784</td>
</tr>
<tr>
<td>STEM or Business Univ. Major</td>
<td>-81.260</td>
<td>0.192</td>
<td>0.000</td>
<td>0.032</td>
<td>0.064</td>
</tr>
</tbody>
</table>

**Notes:** OLS p-value calculated using robust standard errors. CRVE p-value calculated using both a N(0,1) distribution and a T(G-1) distribution, while the Wild p-value is calculated using a Wild cluster bootstrap-t technique and a six point distribution.
Table 4.4: LFS and SLID Estimates for the Atlantic Provinces

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>male</th>
<th>female</th>
<th>single</th>
<th>married</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>wild βGRP</td>
<td>p-value</td>
<td>wild βGRP</td>
<td>p-value</td>
<td>wild βGRP</td>
</tr>
<tr>
<td>University Graduate</td>
<td>-15.924 0.541</td>
<td>-27.400 0.140</td>
<td>-4.961 0.727</td>
<td>11.692 0.747</td>
<td>-44.276 0.060</td>
</tr>
<tr>
<td>College Graduate</td>
<td>-0.197 0.957</td>
<td>11.239 0.060</td>
<td>-11.613 0.311</td>
<td>-17.704 0.226</td>
<td>19.616 0.155</td>
</tr>
<tr>
<td>PSE Graduate</td>
<td>-18.059 0.085</td>
<td>-18.556 0.436</td>
<td>-4.961 0.727</td>
<td>11.692 0.747</td>
<td>-44.276 0.060</td>
</tr>
<tr>
<td>University Student</td>
<td>12.820 0.306</td>
<td>7.833 0.456</td>
<td>17.601 0.546</td>
<td>-3.533 0.947</td>
<td>26.573 0.045</td>
</tr>
<tr>
<td>College Student</td>
<td>-12.909 0.236</td>
<td>-14.286 0.266</td>
<td>-11.387 0.251</td>
<td>-14.012 0.556</td>
<td>-16.179 0.045</td>
</tr>
<tr>
<td>PSE Student</td>
<td>11.227 0.531</td>
<td>8.201 0.551</td>
<td>14.232 0.040</td>
<td>12.061 0.576</td>
<td>6.240 0.185</td>
</tr>
<tr>
<td>PSE Dropout</td>
<td>-1.682 0.982</td>
<td>-6.085 0.997</td>
<td>2.844 0.792</td>
<td>-1.951 0.952</td>
<td>-9.938 0.201</td>
</tr>
</tbody>
</table>

|                | wild βGRP   | p-value      | wild βGRP     | p-value       | wild βGRP     | p-value       | wild βGRP     | p-value     |
| University Graduate | 36.900 0.116 | 32.133 0.132 | 41.093 0.156  | 39.761 0.344  | 27.843 0.168  |
| College Graduate  | -27.534 0.592 | -30.247 0.784 | -20.317 0.104 | 7.370 0.876  | -46.740 0.064 |
| University Dropout| -3.295 0.872 | 13.537 0.324 | -19.431 0.624 | 20.708 0.128  | 6.722 0.821  |
| College Dropout   | -6.735 0.872 | 29.594 0.200 | -43.711 0.032 | 35.961 0.232  | 1.599 0.920  |
| University Ever Attended | -1.647 0.984 | 6.300 0.896  | -15.842 0.284 | 16.856 0.108  | 4.243 0.876  |
| College Ever Attended | 0.299 0.984 | 14.638 0.596 | -15.231 0.340 | -13.543 0.864 | -2.978 0.756 |
| PSE Stdt in Prov. of High School | -40.806 0.048 | -54.228 0.036 | -31.955 0.056 | -30.247 0.096 | -33.185 0.060 |
| High School to University | -20.699 0.612 | 14.122 0.960 | -88.056 0.524 | 66.355 0.668 | 11.249 0.976 |
| High School to College | 46.495 0.436 | 21.970 0.752 | 41.417 0.576 | 106.708 0.332 | -1.954 0.916 |
| Moved Province    | 8.570 0.784 | -4.595 0.812 | 16.905 0.196  | 24.147 0.532  | 4.651 0.820  |
| STEM or Business University Major | -81.260 0.064 | -27.170 0.396 | -99.157 0.156 | -84.357 0.364 | -80.715 0.028 |

Notes: Wild bootstrap conducted with 399 bootstrap replications. All regressions have province and year fixed effects, age fixed effects, and province specific time trends.
<table>
<thead>
<tr>
<th></th>
<th>From Prov $\beta_{GRP}$</th>
<th>Domestic in Prov $\beta_{GRP}$</th>
<th>Within Prov $\beta_{GRP}$</th>
<th>Total in Prov $\beta_{GRP}$</th>
<th>Retention Rate OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic</td>
<td><strong>Pooled</strong></td>
<td>1.415</td>
<td>0.474</td>
<td>0.718</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>NS</td>
<td>0.999</td>
<td>0.729</td>
<td>-0.171</td>
<td>0.949</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>1.776</td>
<td>0.430</td>
<td>1.489</td>
<td>0.619</td>
</tr>
<tr>
<td>Canada</td>
<td><strong>Pooled</strong></td>
<td>16.916</td>
<td>0.049</td>
<td>21.228</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>NS</td>
<td>3.266</td>
<td>0.594</td>
<td>3.056</td>
<td>0.653</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>4.079</td>
<td>0.469</td>
<td>4.679</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td>SK</td>
<td>11.378</td>
<td>0.087</td>
<td>14.377</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>MB</td>
<td><strong>61.788</strong></td>
<td>0.006</td>
<td><strong>79.573</strong></td>
<td>0.008</td>
</tr>
</tbody>
</table>

**Notes:** The first four variables come from the PSIS, while the retention rate variable comes from the Maclean's data.
<table>
<thead>
<tr>
<th>LFS</th>
<th>all</th>
<th>male</th>
<th>female</th>
<th>single</th>
<th>married</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Graduate</td>
<td>0.560</td>
<td>0.967</td>
<td>-17.150</td>
<td>0.807</td>
<td>18.673</td>
</tr>
<tr>
<td>College Graduate</td>
<td>13.156</td>
<td>0.817</td>
<td>61.505</td>
<td>0.571</td>
<td>-33.929</td>
</tr>
<tr>
<td>PSE Graduate</td>
<td>11.767</td>
<td>0.887</td>
<td>42.694</td>
<td>0.566</td>
<td>-17.036</td>
</tr>
<tr>
<td>University Student</td>
<td>-13.837</td>
<td>0.221</td>
<td>-19.008</td>
<td>0.110</td>
<td>-9.110</td>
</tr>
<tr>
<td>College Student</td>
<td>18.308</td>
<td>0.546</td>
<td>10.894</td>
<td>0.206</td>
<td>25.715</td>
</tr>
<tr>
<td>PSE Student</td>
<td>-11.500</td>
<td>0.120</td>
<td>-10.436</td>
<td>0.065</td>
<td>-12.454</td>
</tr>
<tr>
<td>PSE Dropout</td>
<td>6.808</td>
<td>0.827</td>
<td>0.458</td>
<td>0.882</td>
<td>13.261</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLID</th>
<th>all</th>
<th>male</th>
<th>female</th>
<th>parent B.A</th>
<th>parent B.A.</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Graduate</td>
<td>-11.998</td>
<td>0.112</td>
<td>-13.338</td>
<td>0.240</td>
<td>37.376</td>
</tr>
<tr>
<td>College Graduate</td>
<td>-8.001</td>
<td>0.112</td>
<td>-16.632</td>
<td>0.692</td>
<td>-0.435</td>
</tr>
<tr>
<td>University Dropout</td>
<td>23.882</td>
<td>0.136</td>
<td>17.345</td>
<td>0.868</td>
<td>31.114</td>
</tr>
<tr>
<td>College Dropout</td>
<td>20.127</td>
<td>0.148</td>
<td>-15.976</td>
<td>0.448</td>
<td>56.851</td>
</tr>
<tr>
<td>University Ever Attended</td>
<td>11.248</td>
<td>0.756</td>
<td>42.975</td>
<td>0.556</td>
<td>-15.629</td>
</tr>
<tr>
<td>College Ever Attended</td>
<td>7.285</td>
<td>0.932</td>
<td>16.281</td>
<td>0.056</td>
<td>30.437</td>
</tr>
<tr>
<td>PSE Stdt in Prov. of High School</td>
<td>8.023</td>
<td>0.932</td>
<td>26.065</td>
<td>0.584</td>
<td>-10.120</td>
</tr>
<tr>
<td>High School to University</td>
<td>49.792</td>
<td>0.052</td>
<td>80.466</td>
<td>0.692</td>
<td>150.126</td>
</tr>
<tr>
<td>High School to College</td>
<td>-80.652</td>
<td>0.068</td>
<td>-204.592</td>
<td>0.400</td>
<td>12.716</td>
</tr>
<tr>
<td>Moved Province</td>
<td>-5.181</td>
<td>0.572</td>
<td>-19.276</td>
<td>0.412</td>
<td>8.838</td>
</tr>
<tr>
<td>STEM or Business Uni. Maj</td>
<td>-48.900</td>
<td>0.460</td>
<td>-392.311</td>
<td>0.096</td>
<td>162.554</td>
</tr>
</tbody>
</table>

Notes: Wild bootstrap conducted with 399 bootstrap replications. All regressions have province and year fixed effects, age fixed effects, and province specific time trends.
Bibliography


Bernard, Andre; Finnie, Ross, and Benoit St-Jean. (2008) ‘Interprovincial mobility and earnings.’ *Perspectives on Labour and Income*


Department of Finance Canada (2012) ‘Economic and fiscal implications of canada’s aging population’


Francesconi, Marco, and Wilbert van der Klaauw (2007) ‘The socioeconomic consequences of “in-work” benefit reform for british lone mothers.’ *Journal of Human Resources*


104


Appendix A

Proof of $2^{G-1}$ Unique Absolute Value t-statistics

Recall that a bootstrap sample is generated by:

$$y_i^* = X\tilde{\beta} + \tilde{u}_i^*,$$  \hspace{1cm} (A.1)

where $\tilde{u}_i^*$ is the Hadamard product $\tilde{u} \circ v_i$, and $v_i$ is the vector of draws of the bootstrap weights. The Rademacher weights are $-1$ and $+1$, so every possible $v_i$ is equal to $-1 \circ v_j$ for some $i \neq j$. These two bootstrap weight draws will generate the following bootstrap samples: $y_i^* = X\tilde{\beta} + \tilde{u} \circ v_i$ and $y_j^* = X\tilde{\beta} + \tilde{u} \circ v_j$. Since $v_j = -1 \circ v_i$ we can rewrite $y_j^*$ as $y_j^* = X\tilde{\beta} - \tilde{u} \circ v_i$.

We then test the null hypothesis $\beta_i^*$ and $\beta_j^*$ are $= \beta_o$ and calculate a t-statistic of the form:

$$\frac{(X'X)^{-1}X'y_i - \beta_o}{\left(\frac{u_i' u_i}{X'X(n-k)}\right)^{1/2}}.$$  \hspace{1cm} (A.2)
The denominator in equation (A.2) is constant for either $i$ or $j$, as $X$ and $n - k$ are invariant and $u_i' u_i = u_j' u_j$ because $u_j = -1 \circ u_i$.

Let us consider the numerator in equation (A.2), where we have an expression in terms of $\beta_i^*$ and $\beta_o$. If we start with the expression:

$$(X'X)^{-1}X'y_i - \beta_o,$$

using the identity that $y_i = X\tilde{\beta} + \tilde{u} \circ v_i$ we get the following,

$$(X'X)^{-1}X'(X\tilde{\beta} + \tilde{u} \circ v_i) - \beta_o.$$

With a little algebra we get:

$$(X'X)^{-1}X'(X\tilde{\beta} + \tilde{u} \circ v_i) - \beta_o = (X'X)^{-1}X'X\tilde{\beta} + (X'X)^{-1}X'(\tilde{u} \circ v_i) - \beta_o = \tilde{\beta} - \beta_o + (X'X)^{-1}X'(\tilde{u} \circ v_i).$$

Because the bootstrap samples impose the null hypothesis, $\tilde{\beta} = \beta_o$. The numerator then simplifies to:

$$(X'X)^{-1}X'(\tilde{u} \circ v_i).$$

Because $v_i = -1 \circ v_j$, the numerator for the t-statistic of $\beta_j^*$ will be the negative of the numerator for the t-statistic of $\beta_i^*$. Because the denominators are also the same, the t-statistics are equal in absolute value. If we reverse the sign on the weight vector $v_i$ we reverse the sign of the t-statistic, but preserve the magnitude. Thus the $2^G$ unique bootstrap samples will only result in $2^{G-1}$ unique t-statistics in absolute value.
Appendix B

Data Set Descriptions – Chapter 4

LFS

The Labour Force Survey (LFS) surveys roughly 100,000 people who reside in roughly 54,000 households. The survey collects data on a reference week, which is chosen so as to include the 15th day in each month. The LFS data used in this chapter covers the period from January 2000 to March 2013, and I use the public use data which is limited in its demographic and locational variables.\footnote{For instance, the public use data only indicates which province someone resides in and whether they live in Toronto, Montreal, or Vancouver.} Individuals are surveyed once per month and are in the data set for a maximum of six months.

SLID

The Survey of Labour and Income Dynamics (SLID) contains roughly 60,000 people in each wave, and people are in the survey for six years, with new waves starting every three years. The confidential microfile data from 2000-2010 is used. The annual panel nature of the SLID allows us to follow individuals across years. It is possible to
conduct a similar analysis using certain variables from both the SLID and the LFS, the recency of the LFS makes it the preferred data set. For the overlapping variables analysis is conducting using both data sets.

PSIS

The Post Secondary Information System (PSIS) surveys all colleges and universities in Canada and collects information on enrollment and graduation numbers as well as various demographic details about the school’s current students. The public use files from the PSIS contain only two-way tabulations which allow for limited analysis. The survey is conducted on an annual basis, and the surveys from 2000-2008 are used in this chapter. Unlike the LFS and SLID data, estimates from all of Canada will be calculated in addition to estimates using data from the Atlantic provinces. This is done to see if people from GRP provinces in the Atlantic are more likely to go to school in other parts of Canada after the programs are introduced.

Maclean’s

As a robustness check I have collected the historical retention rate from the Maclean’s Magazine University Rankings.\(^2\) The university ranking data reports self reported statistics from nearly 50 universities in Canada. The annual nature of the rankings allows for the construction of a panel. For this chapter I have a panel covering the years 2002 through 2011. Similarly to the PSIS data, estimates are provided separately for both the Atlantic provinces and all of Canada.

\(^2\)I am very appreciative of Mary Dwyer from Maclean’s for providing me with past editions of the Maclean’s rankings.