Imperial College of Science, Technology and Medicine
Department of Mathematics

Estimating the Causal Effect of 20mph Zones in London on Traffic Accidents

Author: Luxi Liu
Supervisor: Prof. Emma McCoy

February 27, 2021

CID: 01068720

Plagiarism Statement

I declare this is my own work except where otherwise stated.

February 27, 2021

Luxi Liu

CID: 01068720

Abstract

By 2020, the Mayor of London plans on implementing a blanket 20mph zone across London in the hopes of reducing accidents. In order to assess whether this policy is effective, we aim to measure the causal effect of current 20mph zones on accidents, using non-20mph roads as our control group. Our secondary aim is to assess the impact of cycle superhighways. We achieve this by using propensity score methods, including matching, stratification and inverse probability of treatment weighting. We find that with the available data 20mph zones/cycle superhighways cause more accidents in comparison to non-20mph zones/non-cycle superhighways.

Acknowledgements

I would like to thank my supervisor, Professor Emma McCoy, for her guidance, support and time. Her encouragement helped me complete my project. I'd also like to thank my friends and family for supporting me throughout my academic career.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

20mph speed zones have been encouraged by the UK government since 2013 [1], when the Department for Transport revised guidelines in a bid to urge authorities into implementing more 20mph speed limits, especially on residential streets and roads with heavy foot and cycle traffic. The supposed safety benefits were quickly noticed by local councils, and since in most cases there was no expectation of additional police enforcement required, there has been a growing interest in the busiest cities in the UK. London is no exception, with Mayor Saddiq Khan planning on rolling out a speed limit of 20mph on all TfL roads within the Congestion Zone by 2020 [2], in a bid to reduce road accidents. It is more relevant now than ever to assess the effectiveness of road speed interventions such as lowering the speed limit.

In a 2018 study commissioned by the government [1], they found using journey speed analysis, that following implementation in cities such as Brighton, Liverpool and Winchester, 47% of drivers in residential areas and 65% of drivers in city center areas complied with the new 20mph limit although this represents only a 1mph reduction in average speed. Additionally, 69% of residents agree that the 20mph limits are beneficial for cyclists and pedestrians. Other research shows that higher speeds can lead to an increase in collisions and accident severity, although currently the government consensus is that there has been insufficient evidence to

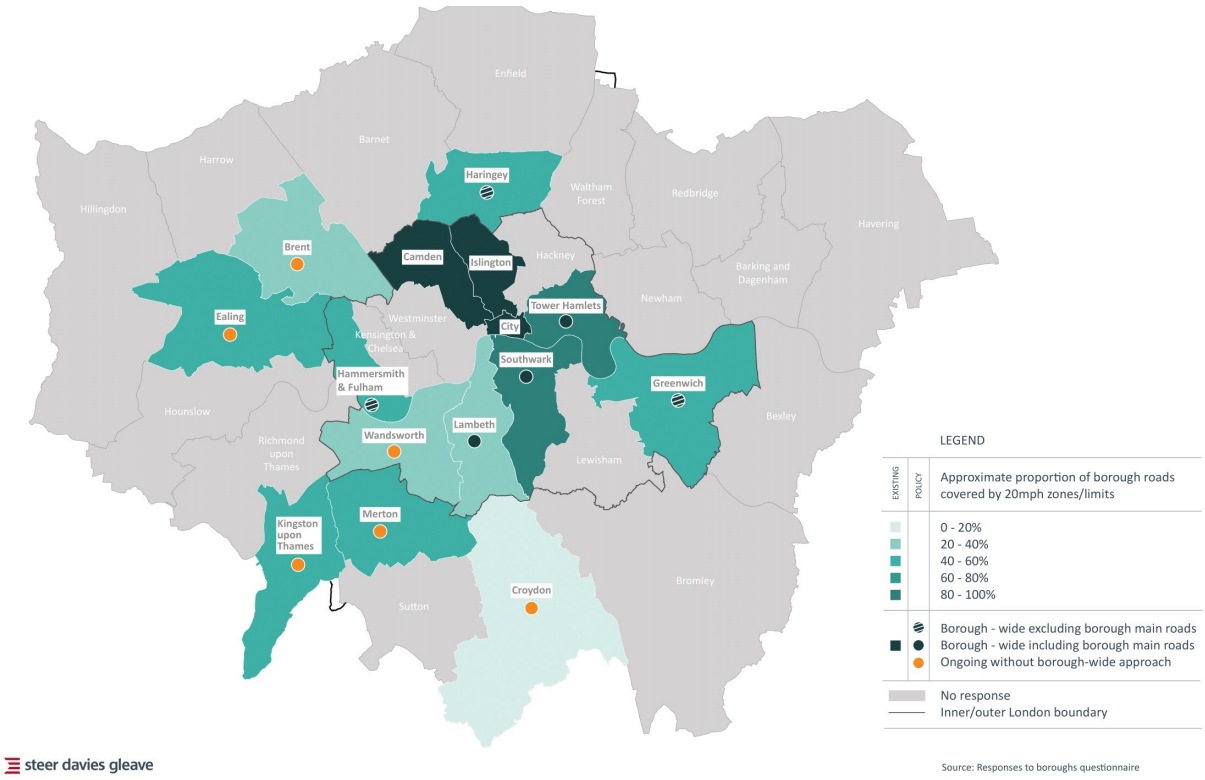

Figure 1.1: Current and Future 20mph zones as of 2016 [16]

conclude that there has been a direct impact from the implementation of the 20mph speed limit on accident rates in these cities.

A good method of assessing this treatment on London is to analyze data from boroughs which have already had changes made to their speed limit. In this project, we assess the difference in number of accidents on roads with 20mph speed limit and on non-20mph roads.

1.2 Motivation and Objectives

The primary goal of this project is to investigate whether 20mph zones have less collisions than non-20mph zones in London. Our secondary goal is to investigate the effects of 20mph zones in conjunction with cycle superhighways as an example of another method of traffic calming.

1.3 Project Structure

- Chapter 1: An introduction into the project, outlining my motivations and goals
- Chapter 2: A literature review of current studies on implementing 20mph zones as a safety measure, and current research on propensity score as a method to eliminate selection bias.
- Chapter 3: A definition of propensity score and an overview of the methodology utilised, such as nearest neighbour matching algorithm and estimators used in this project.
- Chapter 4: A self-generated simulation study to practice propensity score methods and analysis of results in preparation for the real traffic data.
- Chapter 5: Preliminary data analysis to give an idea which variables to use and a direction to approach from. A general assessment of whether the overlap condition is fulfilled.
- Chapter 6: By using the propensity score framework for binary treatments, assessing the effect of 20mph zones by first using the nearest neighbour matching method, then the stratification method.
- Chapter 7: By using the propensity score framework for multi-level treatments, assessing the effect of 20mph zones in conjunction with cycle superhighways, using the weighted regression method
- Chapter 8: A conclusion summarising the results we have found and possible future research directions.

Chapter 2

Literature review

2.1 Previous Studies on 20mph Zones

There has been plenty of literature supporting the relationship between speed and the injury risk for different speed levels. Figure 2.1 shows the that the expected change in accidents when average speed changes 1mph as estimated in a 2006 study [3].

Reference Speed	50 km/h	70 km/h	80 km/h	90 km/h	100 km/h	120 km/h
Injury accidents	4.0%	2.9%	2.5%	2.2%	2.0%	1.7%
Serious injury accidents	6.1%	4.3%	3.8%	3.4%	3.0%	2.5%
Fatal accidents	8.2%	5.9%	5.1%	4.5%	4.1%	3.3%

Table 2.1: Expected change in accidents when average speed changes 1mph

Therefore it seems natural, given that in London there were 2,805 killed or severely injured road casualties in 2011 [4], that speed limits may need to be reduced. The idea of implementing a 20mph zone across London has been considered since 2013, when the Mayor of London at the time released an executive summary [4] detailing his vision for the future of TfL. The British government has supported this by releasing reports analysing the effectiveness of 20mph zones as an accident-reducing measure, the most recent being [1], a technical report analysing the biggest factors relating to the effectiveness of 20mph zones across the UK; speed compliance, impact on number of collisions and impact on severity of collisions.

This study found that based on in-car GPS data, before the speed limit had been lowered, 65

% of drivers in residential areas and 79 % in city centre areas were already travelling at speeds less than 24mph, and after the enforcement this increased to 70 % and 86 % respectively, representing a mediocre change in real speed, as most drivers were self-enforcing a lower speed already. The study explained this phenomenon by stating that most roads that had the new limit implemented were already roads that had certain characteristics that required slower driving - i.e. dense traffic, limited visibility - and so the biggest impact of implementing a new 20mph would be on faster roads. But reduction in speeds on these faster roads are very low, around 1mph. However, due to lack of data, this study only managed to concluded that 20mph zones had no significant reductions to number or severity of collisions. In this project, I hope to assess this more conclusively.

In *Driving speed and the risk of road crashes: A review* [3], it has also been suggested that the implementation of this reduced speed limit may be encouraging alternative methods of transport, such as walking and cycling, which then in turn may have an effect on safety, but since other traffic calming measures such as cycle lanes or bike hire schemes were also introduced, it is hard to isolate the impact of 20mph schemes alone. Therefore, in this project, I hope to assess the individual impacts of cycle lanes and 20mph zones by viewing them as multiple treatments.

2.2 Propensity Score

The idea of propensity scores was first introduced in 1983 by Paul Rosenbaum and Donald Rubin [1] as a way to estimate treatment effect when random controlled trials are not feasible [5]. They showed that instead of considering covariates directly to remove confounding, it was sufficient to consider just the propensity score. Propensity scores allow for treatment effect to be estimated by directly comparing outcomes between treated and untreated units [6]. In the past, researchers have used regression adjustment to reduce confounding when using observational data, but this method does not account for lack of overlap, leaving it extremely dependent on model design [7].

One of the first propensity score methods was matching [8], fully introduced by Rosenbaum

and Rubin in 1985, which utilises the propensity score by matching treated and control units with similar scores and then directly comparing outcomes. Stratification (stratifying subjects into mutually exclusive subsets based on propensity score) was considered as early as 1968 [9], but it was done on a single continuous confounding variable. This result was extended to a propensity score in 1984 by Rosenbaum and Rubin. Inverse Probability of Treatment Weighting (IPTW), proposed by Rosenbaum in 1987 [10], is the final method considered in this paper, and it involves weighting subjects based on propensity score. An example of a recent study successfully utilising some of these propensity score methods is [11] a paper studying the effectiveness of Alcoholics Anonymous, which employed matching and stratification methods.
There have been several studies comparing the efficacy of each method. [12] [13] found that the matching method is more effective over stratification and marginally more than IPTW by analysing the Type I errors, coverage of confidence intervals and variance estimation.
An aspect of propensity score methods that has changed in recent years is propensity score estimation. The most popular approach is still logistic regression, but with machine learning packages becoming more readily available [12], researchers have considered using alternative methods. The main disadvantage of logistic regression is the assumption of linearity, even if interaction terms are included - and deciding which terms to include and how many can be difficult in itself. Machine learning methods such as a generalized boosted model [14] therefore seems like a natural alternative, as it automatically performs this variable selection by splitting trees on different covariates. A disadvantage of this method is that it is often difficult to interpret the results.
Rosenbaum and Rubin's work mostly only covered binary treatment, so some additional work was required to generalize the known methods into multiple treatments [15]. By redefining the framework, and defining a generalized propensity score, the previously discussed methods (matching, stratification, IPTW) were found to be valid in reducing selection bias.

Chapter 3

Methodology

3.1 Motivation for Propensity Score

Randomized control trials are considered the ideal setup when estimating the effect of a treatment on an outcome [1], because the results will be unaffected by confounding factors. However, in reality, randomized control trials may be expensive, difficult or indeed impossible to run (as data may be retrospective). It is necessary to develop tools to be able to assess observational studies that are not randomized control trials, for which treatment selection can be influenced by subject characteristics, leading to systematic differences. The goal is to eliminate the bias caused by treatment selection. Selection bias is caused by non-overlapping supports of X in the treatment and control group, and unbalance between the observable and non-observable confounders. Traditionally, researchers have used regression adjustment to deal with this confounding, but recently there has been interest in utilizing propensity scores instead.

3.2 Potential Outcomes Framework and Average Treatment Effects for Binary Treatments

In this framework, there are possible outcomes but only one outcome is actually observed for each subject, based on whether they did or did not receive treatment. For simplicity, we currently consider only 1 possible treatment, and two outcomes. [16]
Define pair of potential outcomes for an individual i where 0 denotes no treatment and 1 denotes treatment: $Y_i(0)$ and $Y_i(0)$. Define Z_i as an indicator variable describing whether the individual received treatment, where $Z_i = 0$ denotes no treatment and $Z_i = 1$ denotes treatment)

Then, we define the Average Treatment Effect:

$$ATE = E(Y_i(1) - Y_i(0))$$

And define Average Treatment Effect for the Treated:

$$ATT = E(Y_i(1) - Y_i(0)) \mid Z = 1)$$

It is up to the researchers to discern which measure is of more interest, as the first requires a control group so may be more expensive in some circumstances.
For RCTs, an unbiased estimate for ATE is simply $E(Y(1)) - E(Y(0))$. This is consequence of the randomization, and makes estimating the treatment effect very simple. However, generally in observational studies, we have $E(Y(1) \mid Z = 1) \neq E(Y(1))$, so we cannot find an unbiased estimate for ATE directly.
This bias can be illustrated in a simple example: [16]

	$Y \mid Z = 0$	$Y \mid Z = 1$	Z
Adam	0*	1	0
Ben	1*	0	0
Celine	0	0*	1
Debra	1	1*	1
Ella	1	1*	1

In real life, we can only observe the outcome for only one value of Z; that is, we cannot both give and not give a treatment to a patient. We can only observe:

$$Y_i^{obs} = Y_i(Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0))$$

This leads to: [6]

$$E(Y^{obs} \mid Z = 1) - E(Y^{obs} \mid Z = 0) = E(Y(1) \mid Z = 1) - E(Y(0) \mid Z = 0)$$

$$= E(Y(1) \mid Z = 1) - E(Y(0) \mid Z = 1) + E(Y = 0 \mid Z = 1) - E(Y = 1 \mid Z = 0)$$

$$= ATE + Bias$$

The observed outcome for the example above has been starred. From this, we can calculate the conditional probability $Pr(Y = 1 \mid Z = z)$. From this example, we see that the probability of surviving given the treatment is $\frac{2}{3}$, while the probability of surviving having not been given the treatment is $\frac{1}{2}$ - so it appears that the treatment has a positive impact on the probability of survival. However, we also know that the risk of death if everyone had been given the treatment is $\frac{2}{5}$, which is the same as if everyone had not been given the treatment; so the average treatment effect is actually null. The question is, how can we eliminate the selection bias caused by treatment selection so we can recognise that this is an ineffective treatment?

3.3 Propensity Score

We define propensity score as $e_i = Pr(Z_i = 1 \mid \boldsymbol{X_i})$, which can be expressed as the probability individual i is selected for treatment given their observed covariates $\boldsymbol{X_i}$. Rosenbaum and Rubin [17] have been able to show that conditioning on the propensity score will lead to unbiased estimates for average treatment effects, if treatment assignment is strongly ignorable, which is defined thusly:

1. $(Y(1),\ Y(0)) \perp\!\!\!\perp Z \mid X$ (i.e. weak unconfoundedness - treatment assignment is independent of potential outcomes conditional on observed covariates)

2. $0 < P(Z = 1|X) < 1$ (or overlap)

But clearly, 1. is a large assumption because it assumes that all confounders that affect treatment assignment have been measured. However, this assumption can be measured with the introduction of a second control group to examine whether the adjustment actually eliminates bias.
Then, given these conditions, it can be proven that:

$$Z \perp\!\!\!\perp (Y(1), Y(0) \mid e$$

Calculating propensity score itself is a simple logistic regression problem, although recently machine learning methods have also been used.

3.4 Generalized Boosted Regression Model (GBM) to calculate propensity scores

Although most R packages tend to use logistic regression methods to find a propensity score, in this project we also explore using a machine learning method, and use a Generalized Boosted Regression model [5]. The exact details of this method are out of the scope of this project, but simply put, a GBM iteratively combines simple regression trees in order to predict a dichotomous outcome (i.e. a binary treatment), where new tree is chosen to minimise some loss function, with the iterations stopping when some stopping function is fulfilled. The stopping function we will utilise is the Kolmogorov-Smirnov statistic, which compares the distribution of covariates across the groups, and so is a measure of balance. In order to create a GBM for multiple treatments, we create dummy variables for each treatment method, and then fit separate GBMs to each treatment method. Intuitively, this method works because although the outcome is not dichotomous, the weights for each treatment group requires knowing the probability a patient is assigned to said group, and not the probability of another of the others. As discussed in the literature review, GBMs have been found to be the most promising machine-learning approach to generate the propensity score [18] and so it is the method we shall explore.

3.5 Traditional Regression-Based Approach

Firstly, to understand why regression adjustment approaches are in most cases less desirable than propensity score methods, we will examine the approach, which implicitly assumes unconfoundedness [7]:

$$Y_i^{obs} = \beta_0 + \beta_1 Z_i + \boldsymbol{\beta}\boldsymbol{X} + \epsilon$$

where $\beta_1, \boldsymbol{\beta}$ are the coefficient estimates for the corresponding variables. β_0 is the estimate of the intercept. The goal of this model is to estimate β_1, the estimator of the average treatment

effect.

However, there are many problems with this approach. Firstly, unconfoundedness is untestable [7]. Secondly, although this approach assumes unconfoundedness, it does not assume overlap, and therefore this method is not enough to achieve strong ignorability. To understand why this is a problem, consider the following example [7]:

Let a variable X_1 takes values $1, 2, 3$ among the treated but only $1, 3$ among the control. This implies that those treated with value 2 for X_1 do not have a comparable counterpart in the controls. Regression analysis would ignore this.

In general, if the covariates are unbalanced between treated and control, this means the results are highly dependent on model design. In order to circumvent this, we should use non-parametric models - like propensity score methods.

3.6 Propensity Score Matching

Simply put, this method matches treated and untreated subjects with similar propensity scores. There are several methods of doing this, the most common being Nearest Neighbour matching. [1] Once the data has been matched, theoretically selection bias has been eliminated between the matched subjects, and so treatment effect can be estimated by comparing outcomes between the treated and untreated subjects, and similarly so can variance and statistical significance (although this calculation is affected by whether the treated and untreated outcomes should be treated as independent, a topic up for debate).

3.7 Nearest Neighbour Matching

NN matches treated and control units by taking each treated unit and finding its "nearest neighbour". [1]There are several choices that must be made before undertaking matching. Firstly, we must decide whether we opt for one-to-one matching, or one-to-many matching, as treated units could be matched to more than one neighbour if so desired. Next, we must

decide whether the match is made with or without replacement from the untreated patients. Finally, the researcher must decide whether a greedy or optimal matching algorithm is suitable. Greedy matching selects the treated subject at random, then finds the closest available untreated subject, and then repeats this until it has used all the treated subjects. Optimal matching minimizes the total difference in propensity score between all pairs. [7]

When deciding what type of matching to do, we must consider our data, as well the trade-off between bias and variance. One-to-one matching minimizes bias at the cost of additional variance. However, if the size of the treated and control groups are greatly unbalanced, you will end up discarding a large amount of control data. Matching with replacement minimizes bias at the cost of variance.

In this project, the default method of matching used will be greedy one-to-one Nearest Neighbour matching with replacement. However, the following illustration of the implementation will be for one-to-M with replacement matching, for completion [7]:

1. Let $e_i(x_i) = p_i$

2. $I_{m(i)}$ denotes the index of the control unit that is m-th closest to unit i in terms of distance based on the norm $||.||$, where $\mathbb{I}$ is the indicator function

$$(\sum_{j:Z_j \neq Z_i} \mathbb{I}(||p_j - p_i|| \leq ||p_l - p_i)||)) = m$$

3. Let $C(i)_M = \{I_1(i), ...I_M(i)\}$, where $C(i)_M$ denotes the set of indices for the first M matches for i. Then:

$$\hat{Y}_i(0) = \frac{1}{M} \sum_{j \in C(i)_M} Y_j^{obs}$$

3.8 Stratification

This method involves stratifying the data by propensity score [6], and divides subjects into subgroups and then each subgroup is treated as a randomized controlled trial.

More precisely, to generate m strata, we define c_i such that :

$$0 = c_0 < c_1 < c_2 < ... < c_m = 1$$

$$I_k = (c_{k-1}, c_k]$$

Under the regularity assumption that the treatments effect is a continuous function of the propensity scores, the subjects with comparable propensity scores show similar treatments effect, and so we can assume:

$$E(Y_i(j) \mid e_i \in I_k) \approx E(Y_i(j) \mid Z_i = j, e_i \in I_k)$$

where $j = 0, 1$ in the single treatment model.
In other words, in each strata, we have $E(Y(1)|Z = 1) = E(Y(1))$, so the average treatment effect can be very simply calculated.

3.9 Calculating ATT

Under strong ignorability, the first condition assumed for the propensity score, the following holds [7]:

$$E(Y(i) \mid X) = E(Y(i) \mid Z = i, X) = E(Y^{obs} \mid Z = i, X)$$

Therefore, after applying the NN matching method, we can estimate ATT thusly:

$$ATT^{NN} = \frac{1}{N^T} \sum_{Z_i=1} [Y_i^{obs} - \sum_{j \in C(i)_M} w_{ij} Y_j^{obs}]$$

$$= \frac{1}{N^T} \sum_{Z_i=1} Y_i^{obs} - \frac{1}{N^T} \sum_{j \in C(i)_M} w_j Y_i^{obs}$$

Where:

- N^T is the number of observations in the treated group
- w_{ij} is equal to $\frac{1}{N_i^C}$, the number of controls matched to treated observation i if j is in the control units of i, and 0 otherwise
- $w_j = \sum_i w_{ij}$

The variance of this estimator can also be estimated: [7]

$$Var(ATT^{NN}) = \frac{1}{(N^T)^2}\{\sum_{Z_i=1} Var(Y_i^{obs}) + \sum_{j \in C(i)_M} (w_j)^2 Y_i^{obs}\}$$

$$= \frac{1}{(N^T)^2}\{N^T Var(Y_i^{Z_i=1}) + \sum_{j \in C(i)_M} (w_j)^2 Y_i^{Z_i=0}\}$$

$$= \frac{1}{N^T} N^T Var(Y_i^{Z_i=1}) + \frac{1}{(N^T)^2} \sum_{j \in C(i)_M} (w_j)^2 Y_i^{Z_i=0}$$

For the stratification method, we can estimate the ATT using a similar method: [19]

$$ATT_k^{Strat} = \frac{\sum_{i \in I(k)} Y_i^{obs}}{N_k^T} - \frac{\sum_{j \in I(k)} Y_j^{obs}}{N_k^C}$$

where $I(k)$ is the same indexing as previously defined, and N_k^T and N_k^T are the numbers of controls and treated in the k-th block.

$$ATT^{Strat} = \sum_k ATT_k^{Strat} \frac{\sum_{i \in I(k)} Z_i}{\sum_i Z_i}$$

The variance is as follows (assuming independence of outcomes across units):

$$Var(ATT^{Strat}) = \frac{1}{N^T}\{Var(Y_i^{Z_i=1}) + \sum_k \frac{N_k^T}{N^T}\frac{N_k^T}{N_k^C}Var(Y_j^{Z_i=0})\}$$

3.9.1 Multiple Treatments

It has been shown that the propensity score method can be extended to meaningfully consider multiple treatment cases [5], and it has been justified that the use of propensity scores can eliminate bias in these cases. [20]

In this case, define $T = \{t_1, ..., t_n\}$ set of n possible treatment values, and $Z(t)$ as the binary treatment level indicator [21]. Then the generalized propensity score for individual i defined thusly:

$$e_{i,j} = Pr(T = t_j \mid X_i) = E(Z_i = t_j \mid X_i)$$

where $j = \{1, ..., n\}$

It can be proven that given similar assumptions for strong ignorability, as stated for binary treatments, the generalized propensity score can be used as a method for estimating multiple treatment effects, where $T_i(t) := \mathbb{I}(T_i = t)$ ($\mathbb{I}$ is the indicator function):

1. $(Y(1),\ Y(0)) \perp\!\!\!\perp \mathbb{I}(T = t) \mid X$ (i.e. weak unconfoundedness - treatment assignment is independent of potential outcomes conditional on observed covariates)

2. $0 < P(T = t|X) < 1$ (or overlap, true for all X, t)

Once a generalized propensity score has been computed, we can apply the same methods as for binary treatment (propensity score matching and stratification) to calculate the ATE or ATT. However, instead in this project, we find it useful to next employ Inverse Probability of Treatment Weighting (IPTW Method) using the generated propensity scores, as the matching/stratification method usually requires larger sample sizes for each treatment group. This method adjusts the baseline covariates by their weights based on the propensity score, creating a synthetic sample that is independent of treatment assignment [6]. We calculate the weights

based on which estimand we want, ATE or ATT. To find the pairwise ATE, define the following [5]:

$$E(\hat{Y(t)}) = \frac{\sum_{i=1}^{n} T_i(t) Y_i w_i(t)}{\sum_{i=1}^{n} T_i(t) w_i(t)}$$

where $w_i(t) = \frac{1}{Pr(T(t)=1|X)}$

Therefore, the pairwise ATE is:

$$ATE_{t,t*} = E(\hat{Y}(t)) - E(\hat{Y(t*)})$$

To find pairwise ATT, define:

$$E(\hat{Y(t)} \mid T = t') = \frac{\sum_{i=1}^{n} T_i(t') Y_i w_i(t,t')}{\sum_{i=1}^{n} T_i(t') w_i(t,t')}$$

where $w_i(t,t') = \frac{Pr(T(t)=1|X)}{Pr(T(t')=1|X)}$

Therefore, the pairwise ATT is:

$$ATT_{t,t',t'',t'''} = E(\hat{Y(t)} \mid T = t') - E(\hat{Y(t'')} \mid T = t'''')$$

However, the only important case we need to consider are when $t' = t'''$. The other cases are theoretical but have limited applications [5]. What this case measures will be explained in the next paragraph.

Choice of estimand is dependent on the scenario. For multiple treatments, it is much easier to compare effects on the population as a whole using ATE, but this assumes that each treatment can be offered to every member of the population [5]. However, if you are looking at whether

the effectiveness of one treatment t if it were to replace an alternative treatment t' that is typically offered to a group, then it is a lot more effective to analyse $ATT_{t',t',t',t}$, because this is the relative effectiveness of a treatment on a population that typically receives t'.

Chapter 4

Simulation Study

We find it useful to apply the propensity score methods we have explored on a generated example to ensure they are effective for our needs and so we can be sure of the type of results we expect. All coded work was done in R [22]. Appendix A contains the code used in this chapter.

4.1 Binary Treatment Model

I randomly generate patient data, $\boldsymbol{X_1}, , \boldsymbol{X_n}$, independently distributed, $n = 10000$, with age, happiness and health variables data i.i.d. such that for individual i:

$$X_{i,1} = Age \sim Exp(0.1)$$

$$X_{i,2} = Happiness \sim Norm(0, 1)$$

$$X_{i,3} = Health \sim LogNorm(0, 1)$$

I then generate an indicator variable $X_{i,4}$, where $X_{i,4} = 0$ if they receive no treatment, and $X_{i,4} = 1$ if they receive treatment. This variable is dependent on the age, happiness and health of each patient, mimicking the kind of confounding that is present during treatment selection.

I reason that $X_{4,i} \sim Bernoulli(p_i)$, with:

$$p_i = \frac{1}{1 + exp^{-\boldsymbol{X_i\alpha}}}$$

since The link function is the logit function:

$$\boldsymbol{X\beta} = log(\frac{q}{1-q})$$

I choose coefficients $\boldsymbol{\alpha} = (-0.5, -1, -1)$ to ensure sufficient overlap. I sample from this distribution to get the outcome $X_{4,i}$ for each i.

Finally, I generate some binary outcome data. I do this by again randomly sampling the Bernoulli distribution with parameter q_i (with $Y_i = 0$ meaning "survive" and $Y_i = 1$ meaning "death").

So:

$$q_i = \frac{1}{1 + exp^{-X_i\beta}}$$

First, I generate a random outcome that depends on age, happiness and health but not whether the patient was treated, so I take $\boldsymbol{\beta} = (1, -3, -3, 0)$ - mimicking a treatment that has no effect. Then I generate another outcome that is dependent on all age, happiness, health and treatment variables, so I take $\boldsymbol{\beta} = (1, -3, -3, -5)$ - mimicking a treatment that is successful.

Figure 4.1 shows the overlap between the treated and control. I chose such coefficients for treatment selection in order ensure reasonable overlap, although clearly it is not perfect.

4.2 Results and Analysis

Using this data, I generate a propensity score for each patient, and then use propensity score matching methods in order to eliminate the bias that arises from treatment selection. Table 4.1 shows the sizes of the control and treated groups, and the sizes of the matched and unmatched groups. Since we are employing one-to-one matching without replacement, most of the control

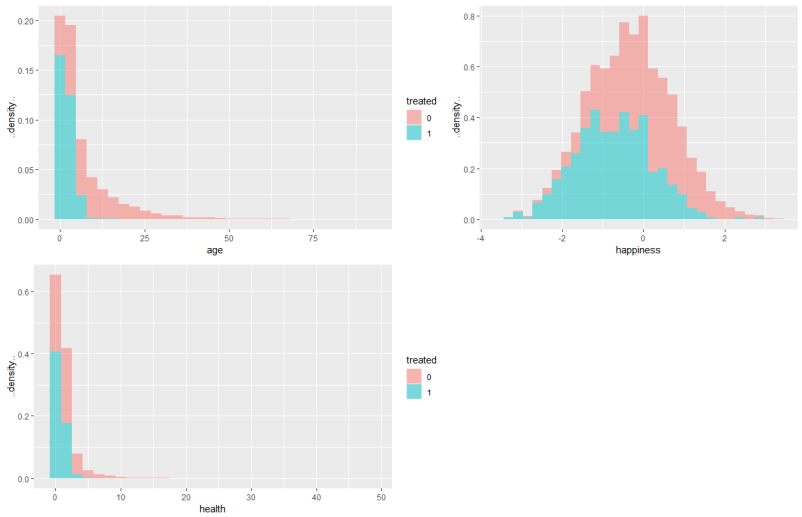

Figure 4.1: Plot of variable distributions for treated and control

group was discarded.

	Control	Treated
All	9413	587
Matched	587	587
Unmatched	8826	0
Discarded	0	0

Table 4.1: Sizes of Control/Treated Matched/Unmatched/Discarded groups

We can also assess balance before and after matching. Because this matching is dependent only on treatment selection and not the outcome, the balance charts will be the same for both random and non-random outcomes. Table 4.2 shows that before matching the treated and control means for each variable are very different, indicating imbalance.

	Treated Mean	Control Mean	SD Control	Mean Diff.
distance	0.3177	0.0426	0.0939	0.2751
age	1.9761	10.4830	10.2533	-8.5069
happiness	-0.7413	0.0504	0.9860	-0.7917
health	0.7523	1.7342	2.2840	-0.9819

Table 4.2: Simulation study covariate balance before matching

Table 4.3 shows balance after matching. There has been a high % reduction in mean difference, and the means are now a lot closer. This indicates the matching method has been successful.

Figure 4.2 is the distribution of the propensity scores before and after matching. I generate this graph, as I do all the graphs in this project, using ggplot [23]. We can see that for higher

	Treated Mean	Control Mean	SD Control	Mean Diff.	% Reduction
distance	0.317666829	0.286316588	0.183908736	0.031350242	88.6047249
age	1.976101137	2.008146133	1.834604936	-0.032044996	99.62330425
happiness	-0.741272446	-0.638272634	0.892658518	-0.102999812	86.98987899
health	0.752289181	0.762286449	0.575502788	-0.009997268	98.98187867

Table 4.3: Simulation study covariate balance after matching. Last column is the % reduction in mean difference.

values of the propensity score for treated units, there is insufficient overlap with control units. This could affect our results.

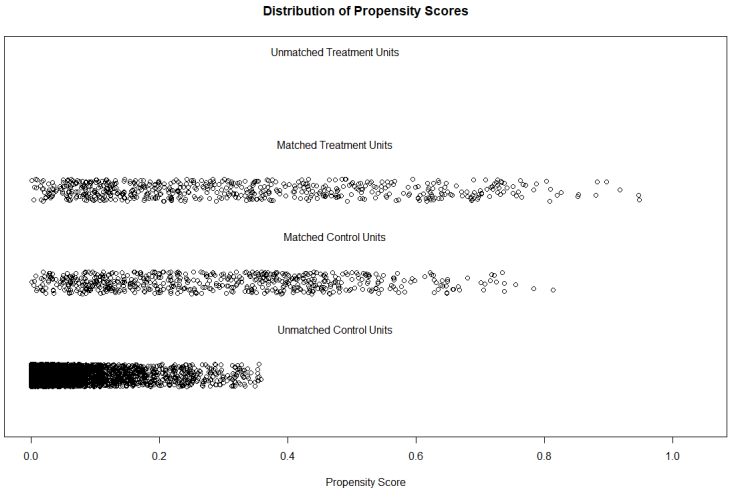

Figure 4.2: Simulation study distribution of scores before and after matching

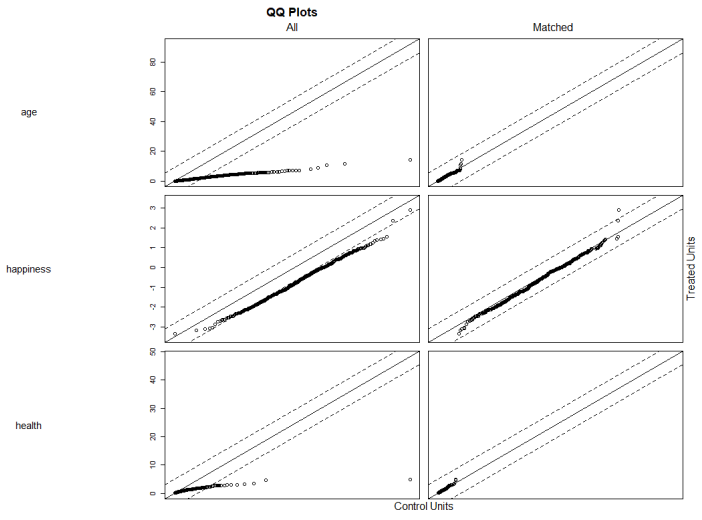

Figure 4.3: QQ plot of the covariates before and after matching

Figure 4.3 shows a QQ plot of the covariates based on propensity score based on the control and treated units, and compare before and after matching. After matching, most of the poitns

lie perfectly on the $y = x$ line of the QQ plot, indicating very close matching.

Finally, using the generated outcomes, we can measure the ATT by generating treated and control outcomes using Zelig. We expect to see little to no ATT for the outcome data that is not dependent on the treatment, but a noticeably negative ATT for the outcome data dependent on the treatment (since 1 indicates death, and 0 indicates survival, we expect the treatment to lessen the probability of death).

	2.5%	50%	97.5%
Random Effect	-0.01361709	0.00976096	0.03339213
Successful Effect	-0.4733956	-0.4519654	-0.4296088

Table 4.4: ATT for the binary treatment simulation study

Table 4.4 shows the quantiles for ATT, and that the results are as expected. At a 5 % significance level, you cannot reject that the random treatment has an ATT of 0, while for the successful treatment you can reject that the ATT is 0, as the ATT is greatly lower than zero, indicating it is effective.

4.3 Multiple Treatment Model

We can decide to expand our simulation study further by considering multiple treatments $T = \{t_1, t_2, t_3\}$. In order to do this, we use the same $X_{i,1}, X_{i,2}, X_{i,3}$ as previously defined, but use the multinomial distribution to find the $X_{i,4}$, the factor variable for which treatment individual i receives. This multinomial distribution has parameters $n = 1$, and $p_{i,j}$ defined as the probability of treatment i for an individual j for $i = 1, 2$: [24]

$$p_{i,j} = \frac{exp^{\boldsymbol{\alpha_i}.\boldsymbol{X_j}}}{1 + \sum_{k=1}^{2} exp^{\boldsymbol{\alpha_k}.\boldsymbol{X_j}}}$$

And the probability of treatment 3 for an individual j is:

$$p_{3,j} = \frac{1}{1 + \sum_{k=1}^{2} exp^{\boldsymbol{\alpha_k}.\boldsymbol{X_j}}}$$

A constraint is $\sum_{i=1}^{3} p_{i,j} = 1$ for all $j = 1, .., n$.

With these probabilities, I select each individual for a treatment group according their age, happiness and health. I do this by randomly sampling from the multinomial distribution, taking $\boldsymbol{\alpha_2} = (-5, 0, -5)$, $\boldsymbol{\alpha_3} = (0, 2, -3)$. I put them in treatment group 1 otherwise. Once assigned, I create dummy indicator variables $X_{i,4}, X_{i,5}, X_{i,6}$, with $X_{i,4} = 1$ when individual i falls into treatment group 1 and is 0 otherwise, $X_{i,5} = 1$ when the individual falls into treatment group 2 adn 0 otherwise, and so on.

Similar to binary treatment study, I then generate an outcome based on which treatment group the individual falls into, with probability of death for individual i:

$$p_i = \frac{1}{1 + exp^{-\boldsymbol{X_i\beta}}}$$

where $\boldsymbol{\beta} = (1, -3, -3, 0, -2.5, -5)$

In order to demonstrate the importance of overlap, the coefficients chosen for treatment 2 are the same as the coefficients in the binary treatment (i.e. have very poor overlap relative to the treatment 1) while the coefficients for treatment 3 have almost full overlap with treatment 1 (which is supposed to have no treatment effect).

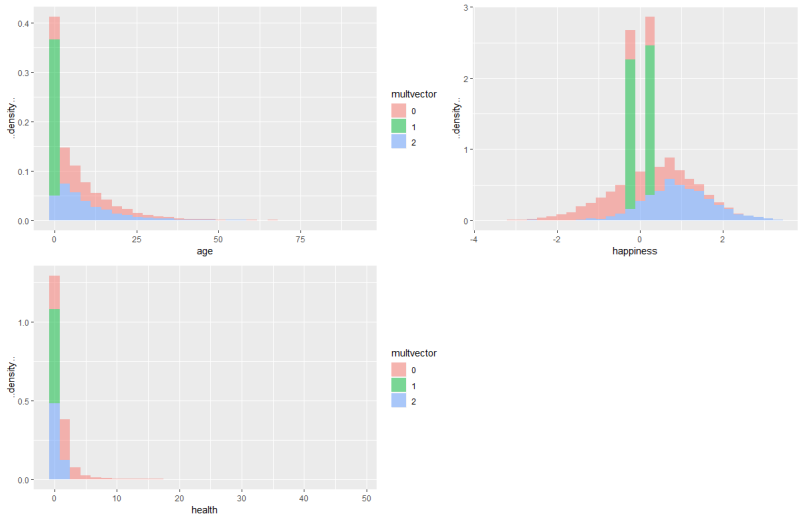

Figure 4.4: Plot of variable distributions for treatments $1, 2, 3$

So clearly treatment 2 has a positive effect on probability of surviving, but treatment 3 is more beneficial, while treatment 1 should have no effect at all. We can analyse whether the results reflect this, and consider the pairwise ATEs also.

I use the GBM to calculate the propensity score, as the matchit [25] package does not deal with multi-treatments. After generating the propensity score, it is important to check that the balance measure chosen (in this case, the mean Kolmogorov-Smirnov statistic) is no longer decreasing after the maximum number of iterations set, as you want to make sure that you have found the minimum [26]. I selected 3000 iterations.

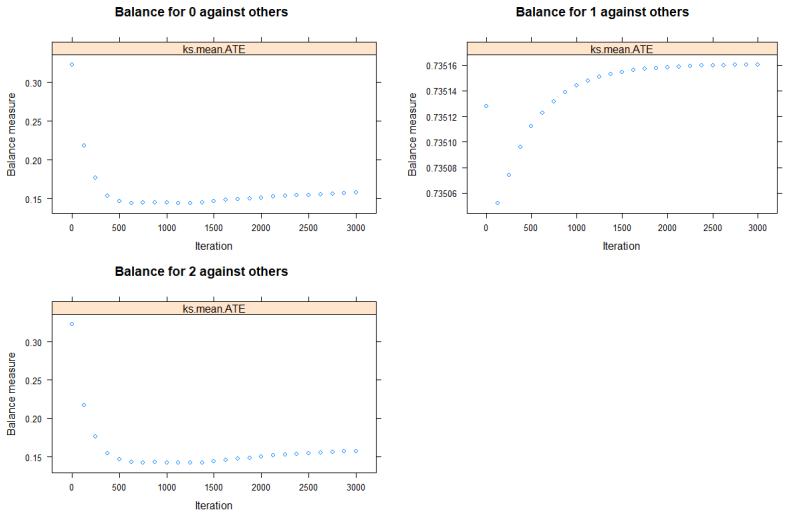

Figure 4.5: Simulation study plot of balance measures for each treatment

Finally, we can use the propensity score weighting method to find the ATE. We can fit a glm using the weights. The following at the coefficients.

	Coeff. Est.	SD	t-value	p-value
Treatment 1 (Intercept)	0.68996	0.02625	26.288	¡2e-16
Treatment 2	-14.14594	0.52078	-27.163	¡2e-16
Treatment 3	-0.95898	0.10502	-9.132	¡2e-16

Table 4.5: Coefficient estimates for binomial regression using IPTW-adjusted covariates

We used the binomial distribution as the outcome variable is binomial, and got a dispersion parameter of 0.9982802, very close to the theoretical dispersion parameter for binomial distributions (1) indicating a good fit [27].

When interpreting these coefficients, remember that we must interpret treatment 2 and treatment 3 relative to treatment 1 (intercept). So, we find, using the link function, that the ATE

is:

	Treatment 1	Treatment 2	Treatment 3
ATE	0.665958	-0.6659566	-0.2328103

Table 4.6: Simulation study ATE for multiple treatments

This is clearly not as we expected. This is because the overlap for treatment 3 against the others is very poor.

4.4 Conclusion

During this simulation study, I tried many different combination of numbers, and found that the matching method was much more effective when the treated group size was a lot smaller than the control group - which is perhaps unsurprising, as the implementation of nearest neighbour method we are using pairs a treated patient with its closest neighbour in the control group, regardless of distance, so in general if there were more points in the control group, there is clearly more choice and this would lead to closer matches. Perhaps I could have experimented with one-to-N matching, as it did mean that the vast majority of our control data went unmatched. When we apply what we have learnt from our simulation study to our traffic data, we may have to use a different version of the algorithm if our treated and control groups are of a similar size (for instance, matching with replacement).

Also, as we have discovered some variable selection, overlap can lead to a big bias in results, so we must carefully assess the overlap in our traffic data. When selecting the coefficients, I found if overlap wasn't fulfilled, I would get a lot of results that were not as expected, which is unsurprising. [28] found that using a propensity score weighting method, low overlap from [0,0.1) could introduce a relative mean bias of 9.9 %. There are methods that can be used to overcome lack of overlap, for instance the trimming technique [29], but it is not covered in this project.

Finally, for the simulation study I decided to only consider logistic regression for the binary treatment as a method for generating propensity score. This is because machine learning

techniques really shine when there are interactive terms between the variables [12], but since we know that the terms are all linear, I didn't find including GBM necessary.

Chapter 5

Accident Data

The data used was originally sourced from a previous study into the efficacy of London cycle superhighways [30], and so includes statistics relevant to cycling, including the existence of a cycle superhighways, cycling accidents and number of Santander stops. The data sourced had information the accidents on 450 roads in two different time periods, from the period of 2007 to 2014, and measured the change in accidents across some of the roads that introduced superhighways. I decided to adapt this for my context, and take only the data from the second period, to avoid bias from including the same road twice. The included variables, their interpretations, ranges and means are shown in Table 5.1.

5.1 Data Analysis

Firstly, we should undertake some preliminary data analysis to better understand our data, in particular overlap. In figure 5.1, I wanted to look at was whether the number of accidents correlated strongly with the length of the road, as it did seem like the number of accidents would be better measured per metre of road, as there was a great range in road length.

Surprisingly, there seems to be little correlation between road length and number of accidents. However, as we can already see, the number of accidents in 20mph zones actually appears to be

Variable Name	Interpretation	Range	Mean
X20mphzone	Existence of 20mph speed limit	{0,1}	0.511
Treatment	Existence of cycle superhighway	{0,1}	0.167
Length	Length of road (m)	[124.0, 1858.0]	1048.9
AADFTotal	Annual Average Daily Flow total	[5271, 117439]	26005
AADFPedal	AADF Cyclists	[0, 9051]	1045.0
AADFBus	AADF Buses	[8.25, 5396.75]	954.90
TotalCycleAcc	Total cyclist accidents	[0, 97]	16.33
SlightCycleAcc	Slight cyclist accidents	[0, 90]	14.38
AllAcc	All accidents	[2, 183]	47.86
AllAccSlight	All slight accidents	[0, 162]	42.91
AllAccKSi	All accidents with deaths/serious injuries	[0, 21]	4.947
Busstops	Number of bus-stops	[0, 22]	7.562
Speed	Average speed (km/h)	[15.23, 68.63]	28.06
Domes	% of domestic buildings	[11.89 , 28.351]	12.310
Non-Dom	% of non-domestic buildings	[0.510, 37.390]	10.876
RoadArea	% of road area	[3.345, 34.974]	19.587
IMD	Index of multiple deprivation	[5.829, 61.922]	31.189
Pop	Population /mˆ2	[0.000452, 0.019358]	0.008353
Employ	Employees /mˆ2	[0.0001976, 0.0109052]	0.0042642

Table 5.1: Accident data variable names, interpretation, ranges and means

Figure 5.1: Road length (m) against accidents

much higher than number of accidents on roads that do not have a 20mph speed limit. Indeed, this seems to pattern repeat itself as we compare other variables against whether it is a 20mph zone.

Figure 5.2: Number of accidents in the road in the year, coloured by 20mph zone

Similarly, Figure 5.2 shows that 20 mph zones seem to have a higher average number of accidents. This does seem perplexing - perhaps because 20mph zones are busier, there is more traffic and therefore there are more accidents? It is informative therefore to break down the accidents by type. Perhaps although there are more accidents in 20mph because they are busier, there will be less KSI accidents.

Figure 5.3 does not seem to support our hypothesis. Accidents in general seem to be much higher in 20mph zones. Therefore, we might begin to question whether people are actually following these 20mph speed limits, as these limits are usually self-policed, and it has been reported in past government-funded studies [1] that the average traffic speed before the introduction of a 20mph limit was often only just higher than 20mph, and some doubt was cast on whether there was any change in average speed at all. Unfortunately, we do not have data for whether there have been changes in average speed before and after the enforcement of a 20mph speed limit, so we can only compare current average speed between 20mph and non-20mph zones, as shown in figure 5.4. So it seems that broadly the average speed in 20mph zones is actually lower that that of non-20mph zones, although the mode is still over 20mph. It also shows that there is a

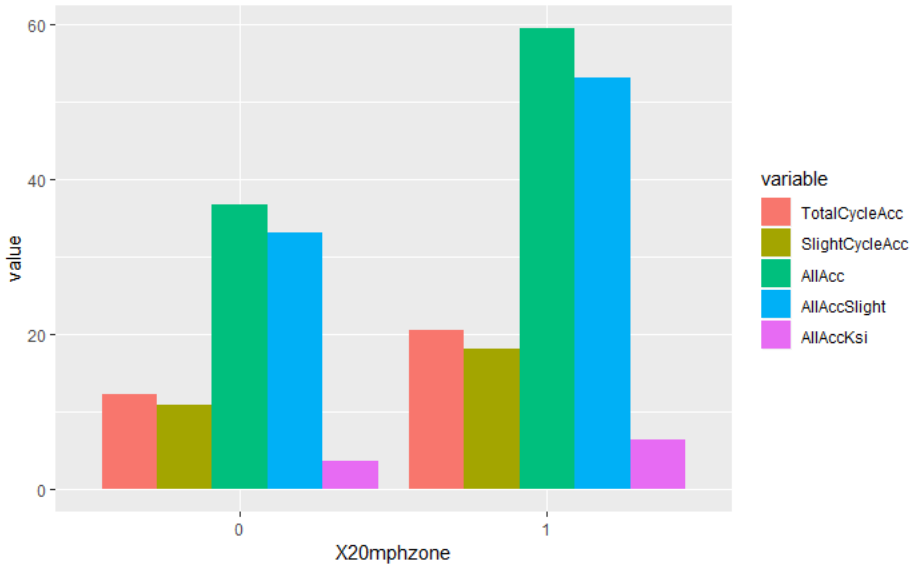

Figure 5.3: Average number of accidents sorted by type of accident and whether the speed limit is 20mph

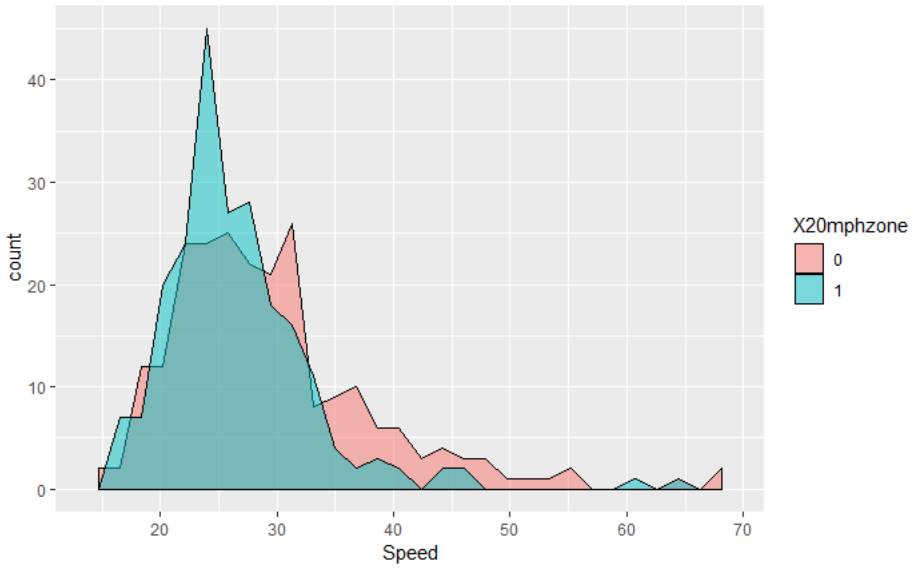

Figure 5.4: Average speed of traffic travelling down the road, coloured by 20mph zone

slight lack of overlap between control and treated for high speeds.

Chapter 6

Propensity Score Methods with Binary Treatment

Before applying propensity score matching methods, I decided to remove AllAccSlight, AllAccKSI, TotalCycleAcc, SlightCycleAcc. This is because our output variable is AllAcc, and clearly these other variables are a subset of AllAcc, so I would expect strong linearity/correlation that could affect our results. Appendix B contains the code used in this chapter.

6.1 Assessing Balance

I generate some information assessing the balance between the treated and control groups. Table 6.1 shows the balance before matching, using the package matchit [25], a package that can computes balance as well as undertake nearest-neighbour matching. It is clear that there is a big difference between the treated and control covariates, so it is likely propensity score methods would help balance in this case.

	Treated Mean	Control Mean	SD Control	Mean Diff.
distance	0.6086	0.3744	0.19803	0.23416
treatment0	0.8000	0.8652	0.3422	-0.0652
treatment1	0.2000	0.13478	0.3422	0.0652
length	1069.2045	1029.53478	184.3097	39.6698
AADFTotal	24872.6608	27088.3034	21088.8243	-2215.6426
AADFPedal	1256.4141	842.7351	1017.01730	413.6790
AADFBus	1113.3198	803.3640	489.1144	309.9558
Busstops	8.1591	6.9913	3.2602	1.1678
Speed	26.5568	29.5070	8.9077	-2.9502
Domes	0.1279	0.1185	0.05636	0.0093
Non_Dom	0.1235	0.0947	0.0686=	0.0289
RoadArea	0.2088	0.1835	0.0561	0.0253
IMD	36.0140	26.5730	11.2139	9.4410
Pop	0.0095	0.0072	0.0040	0.0023
Employ	0.0047	0.0039	0.0022	0.0008

Table 6.1: Balance between Unmatched Treated and Control Confounders

6.2 Nearest Neighbour Matching

Firstly, we apply propensity score matching using the nearest neighbour matching discussed in the methodology. As seen in figure 6.2, because our control and treated groups sizes are very balanced, I decided that we should apply nearest neighbour matching with replacement - otherwise, we could have a control and treated unit matched that have very different propensity scores, but because they are one of the last units to be matched. Although this does mean we discard some of the data, which may lead to less robust results.

	Control	Treated
All	230	220
Matched	90	220
Unmatched	140	0
Discarded	0	0

Table 6.2: Treated and Control Group Sizes, All, Matched, Unmatched and Discarded

Table 6.3 shows the balance after the matching process has been applied, with the last column showing % reduction in mean difference. There has been a 99 % reduction in distance (difference in propensity scores), so the matching method has appeared to be very successful.

Figure 6.1 is a plot of the propensity scores of the treated and control groups against the

	Treated Mean	Control Mean	SD Control	Mean Diff	% Reduction
distance	0.6086	0.6073	0.2238	0.0013	99.4448
treatment0	0.8000	0.8500	0.3591	-0.0500	23.3333
treatment1	0.2000	0.1500	0.3591	0.050	23.3333
length	1069.2045	1066.5909	169.0950	2.6136	93.4115
AADFTotal	24872.6608	23482.7451	15865.1975	1389.9158	37.2681
AADFPedal	1256.4141	924.4192	928.1794	331.9948	19.7458
AADFBus	1113.3198	1151.8393	754.2789	-38.5195	87.5726
Busstops	8.1591	8.4182	2.9047	-0.2591	77.8135
Speed	26.5568	25.8272	7.0224	0.7295	75.2725
Domes	0.1279	0.1421	0.0582	-0.0142	-52.1043
Non_Dom	0.1235	0.1109	0.0685	0.0126	56.1917
RoadArea	0.2088	0.2184	0.0595	-0.0096	62.1216
IMD	36.01404	35.9592	12.1770	0.0549	99.4189
Pop	0.0095	0.0100	0.0044	-0.0005	79.8640
Employ	0.0047	0.0050	0.0024	-0.0003	64.3592

Table 6.3: Balance between matched treated and control confounders using NN-matching method with replacement. % Reduction is the change in mean difference

variables which showed the biggest % change in mean difference. As we can see, once matched, the loess-smoothed lines are very close, indicating that the matching has been successful.

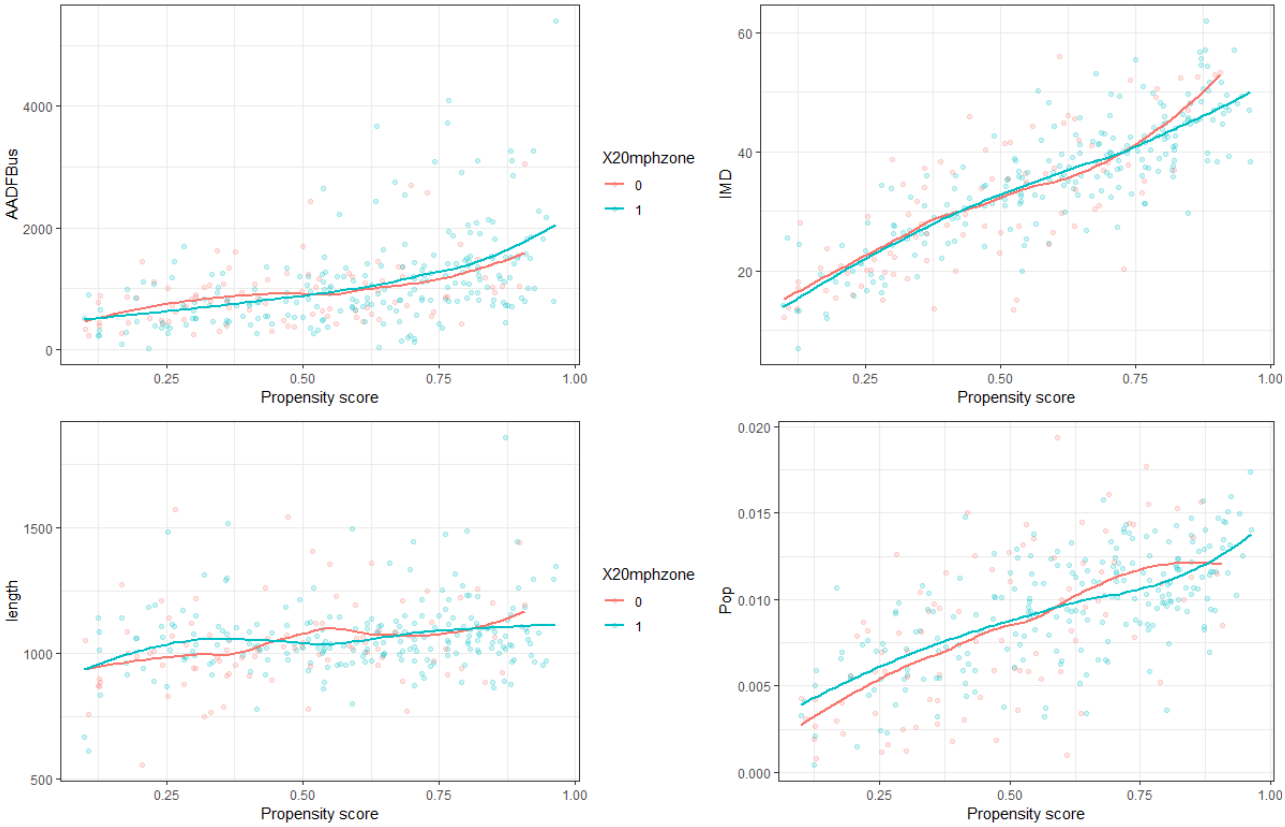

Figure 6.1: Distribution of propensity scores of treated and control populations with the biggest reduction in mean difference

Figure 6.2 shows the change in covariate balance in terms of absolute mean differences for matched and unmatched samples. The matching process seems very successful in improving the balance, except for Domes, where it has actually increased the distance.

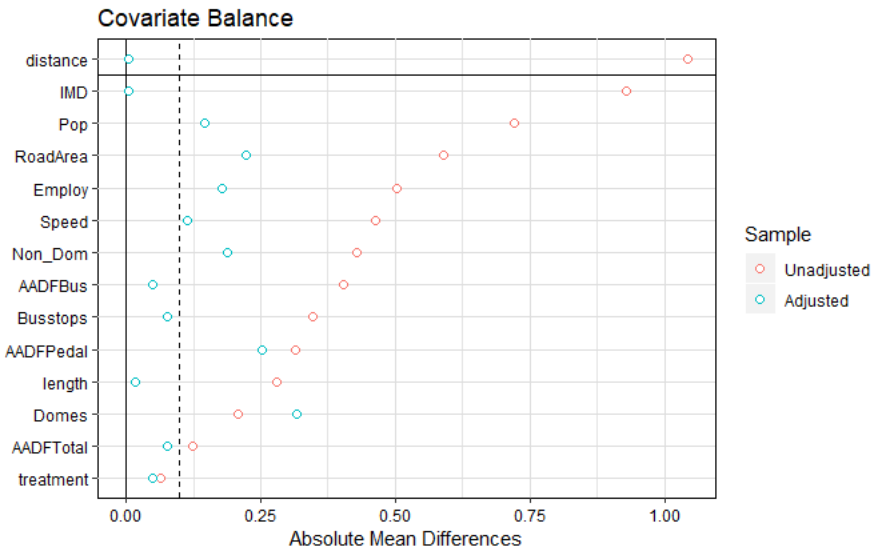

Figure 6.2: Absolute mean differences between treated and control variables before and after matching

Next, I found it informative to view the distribution of propensity scores of the matched and treated units, and see which which control units remained unmatched. Figure 6.3 shows this. It appears that lots of control units overlapping with the matched treatment units went unmatched. This suggest that there are some potentially informative units that could have not been used. Therefore maybe a one-to-M matching process would be more appropriate as it would allow inclusion of more of these unmatched control units.

6.2.1 Results

Next, using Zelig [31], we can simulate results for treated and control groups and thereby generate an estimated distribution for the ATT using the estimator described in the methodology for nearest-neighbour matching. The following is the mean, SD and the 2.5%, 50%, 97.5% quantiles [32].

I can also calculate the ATE:

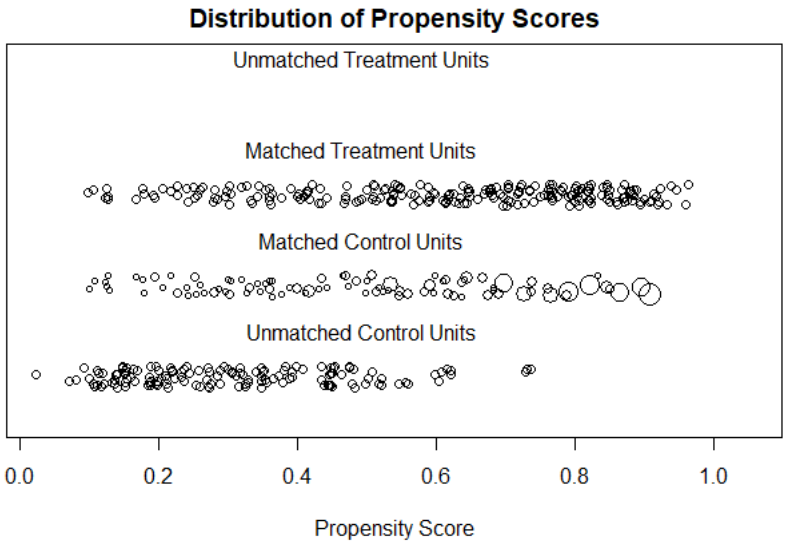

Figure 6.3: Distribution of propensity scores by matched/unmatched and control/treatment

Mean	SD	2.5%	50%	97.5%
7.9463	0.8734	6.2952	7.9492	9.6428

Table 6.4: Binary treatment matching method ATT

Mean	SD	2.5%	50%	97.5%
7.3019	0.9604	5.8031	7.1408	9.3358

Table 6.5: Binary treatment matching method ATE

The ATE does not seem too different from the ATT, so let us focus our analysis on the ATT. Figure 6.4 shows the distributions of the outcome given control group (x) or treated group (x1).

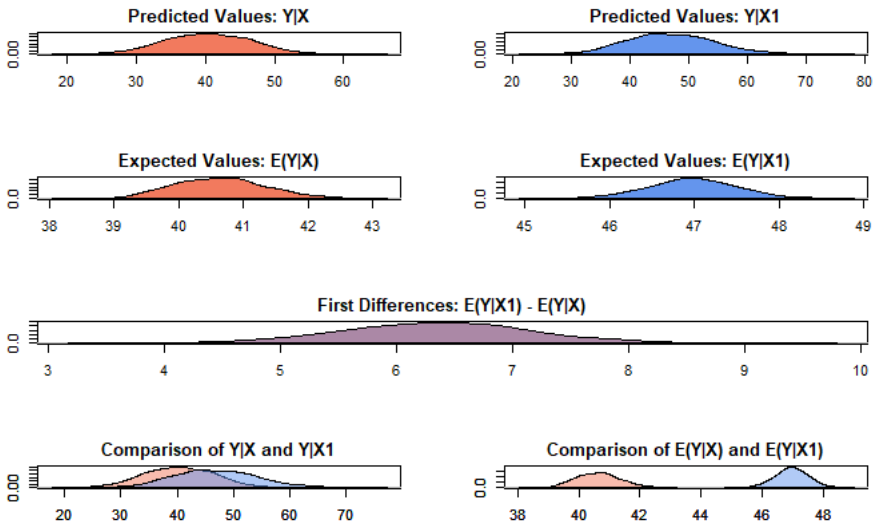

Figure 6.4: Zelig-generated graph of expected values for x (control) and x1 (treated) for binary treatment matching method

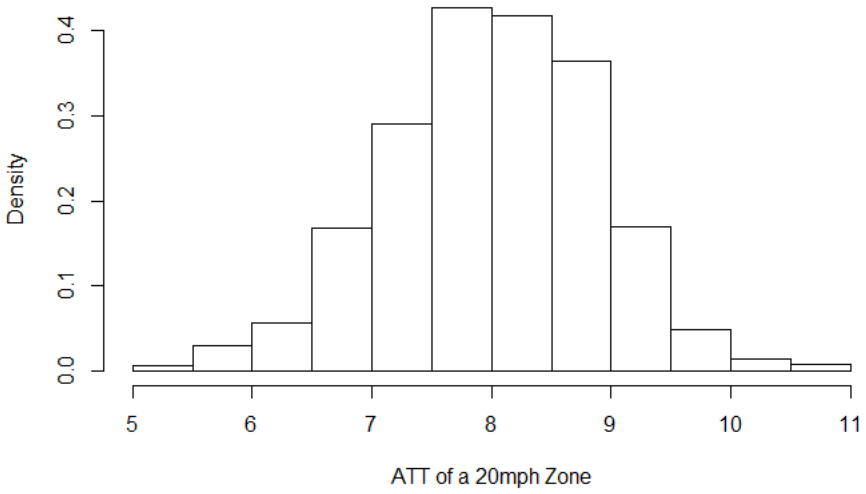

Figure 6.5: Distribution of ATT values generated by Zelig for binary treatement matching method

So despite the matching process resulting in a 99 % reduction in difference between "distance" (or propensity scores), we can see that the average ATT of a 20mph zone is 8.01, which can be interpreted as the average additional accidents per road given that the road has a 20mph speed limit. Figure 6.5 is a histogram of generated ATT. This result is very strange, as we were

predicting a negative ATT. This result may be because some of the conditions for using the propensity score as a selection bias reduction method were not fulfilled.

6.3 Stratification

We assess the stratification method by splitting the treatment and control into 6 subclasses by propensity score. Again, I use the matchit and Zelig packages.

	Subclass 1	Subclass 2	Subclass 3	Subclass 4	Subclass 5	Subclass 6
Control	119	62	26	12	6	5
Treated	37	36	37	36	37	37
Total	156	98	63	48	43	42

Table 6.6: Binary treatment stratification method group sizes for control and treated by subclass

Unfortunately, it is difficult using the matchit package to see the balance between variables. As such, I am including stratification as simply an alternative lesser method to matching - as discussed in the literature review, studies have found stratification as a worse method to remove bias, and I find the lack of interpretable outputs make the method much more difficult to analyze.

Although matchit does not allow for balance summaries across subclasses, we can utilise the cobalt [33] package to compare the the (standardized) difference in covariate means between the two groups after matching/stratifying between the matching method and the stratification method. The only difference we need concern ourselves with is distance, as it is average difference in propensity scores in each strata/each match.

	matching method	stratification method
distance	0.0058	0.0183

Table 6.7: Comparing difference in means between treated/control after applying the propensity score method

As we can see, the difference in means between treated and control after adjustment for the matching method is about 3x smaller than the difference for the stratification method. This

suggests that the stratification method is a less successful method in removing the bias caused by treatment selection, and the treated and control units "grouped" into strata have a larger differences in covariates than those matched by the matching method.

6.3.1 Results

We find that the ATT is:

Mean	SD	2.5%	50%	97.5%
2.714073	0.5670888	1.619078	2.709833	3.796958

Table 6.8: Binary treatment stratification method ATT

The ATE is:

Mean	SD	2.5%	50%	97.5%
2.8899	0.8611	1.2670	2.8565	4.6693

Table 6.9: Binary treatment stratification method ATE

The results are quite different from the results for the matching method, and are lower. From what we know of the matching method in comparison with the stratification method, I would suggest that these results are not as valid as the matching method. This may be because of a number of reasons. Maybe more strata are appropriate in order to reduce the differences between the control and treated units that are grouped.

6.4 GBM and IPTW Results

The final method we assess is the GBM to generate a propensity score, and then using the IPTW regression adjustment method to find the coefficient estimates for the adjusted Poisson regression model. I do this using the twang package [34] and then the survey package [35] for R.

Then, using the link function for the Poisson distribution ($e^{X\beta}$), we can calculate the ATE and ATT using their definitions.

		Coeff. Est	SD	t-value	p-value
ATT	Intercept (Control)	3.8240	0.0743	51.4440	<2e-16
	Treated	0.2623	0.0845	3.104	0.0020
ATE	Intercept (Control)	3.6922	0.0523	70.5740	2e-16
	Treated	0.2653	0.0697	3.8080	0.0002

Table 6.10: Binary treatment GBM/IPTW method ATE and ATT coefficients for the covariate-adjusted Poisson regression

ATT	13.7339
ATE	12.1931

Table 6.11: Binary treatment GBM/IPTW method ATE and ATT

Clearly, this is quite a lot larger than the previously estimated values. Using the cobalt package again, we find that the difference in means of the propensity score between treated and control after adjustment for the GBM/IPTW method for ATE/ATT were 0.8512/0.9311 respectfully - indicated the method was actually very unsuccessful at eliminating the bias. This is because covariate adjustment regression assumes that the relationship between the propensity score and the outcome has been correctly modelled [6].

Chapter 7

Propensity Score Methods with Multiple Treatments

We can analyse our data further by considering a second possible "treatment" - whether the road is also Cycle Superhighway. However, since this treatment and the 20mph treatment are not mutually exclusive, we can chose to split these into four treatments:

0. A cycle superhighway is not present and it is not a 20mph road - i.e. no treatment.
1. A cycle superhighway is present, but it is not a 20mph road
2. A cycle superhighway is not present, but it is a 20mph road
3. A cycle superhighway is present, and it is a 20mph road

However, it is clear that our previous methodology only applies if the probability that a road is a cycle superhighway and has a 20mph speed limit are independent.
We can these for independence using the Chi-squared test by splitting the data set into one of four treatments, and testing them against the expected outcomes given independence.

We get: X-squared = 2.9899, df = 1, p-value = 0.08378. So at 10 % significance level, we can assume independence.

Observed	20mph Zone		
Cycle Superhighway	No	Yes	Total
No	199	176	375
Yes	31	44	75
Total	230	220	450

Table 7.1: Observed group sizes for treatment 1 through 4

7.1 Generalized Boosted Regression Method

Appendix B contains the code used in this chapter. Firstly, we decide to start by using 3000 trees (or iterations) for the GBM. It is then important to assess whether this is the correct amount of trees for the Kolmogorov-Smirnov mean statistic stopping criterion [26]. We can do this by plotting the balance measures and seeing if they are still decreasing after 3000 iterations; if they are, it is likely we need to re-run the package with more iterations.

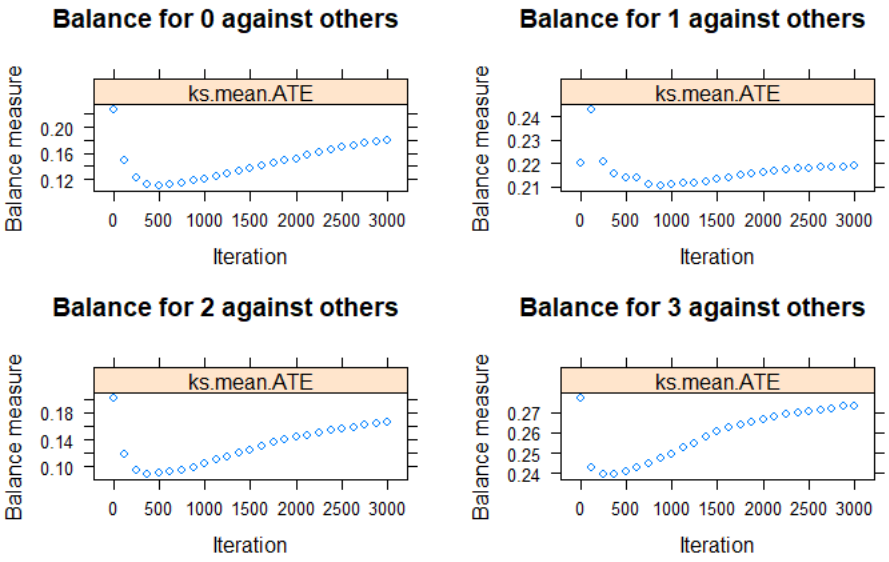

Figure 7.1: Plot of balance measures for each treatment

As we can see from figure 7.1, it appears that there have been enough iterations. Next, my biggest concern would be the overlap between the groups, since we have 4 treatments and we generate propensity scores for each of the treatments. We can assess this graphically using a box-plot displaying the quantiles, as in figure 7.2.

So it is likely that the overlap assumption is not fulfilled, especially for treatments 1 and 3, as you can see from the top-right and bottom right sub-figures in Figure 7.2. This may be because

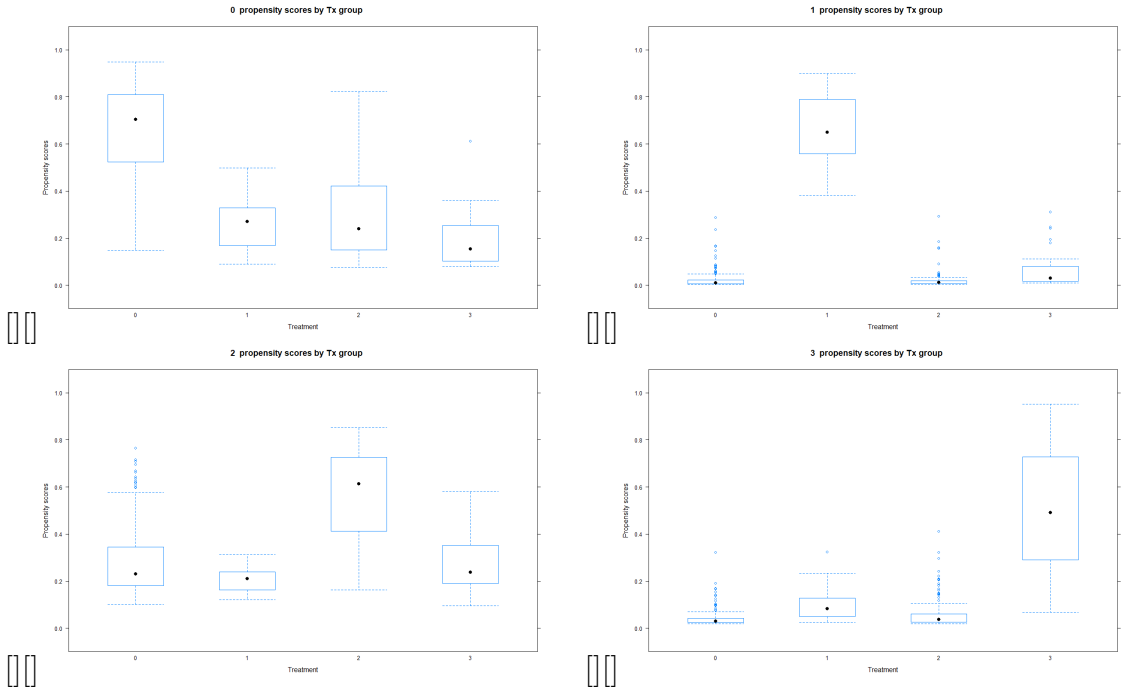

Figure 7.2: Plot of quantiles of propensity scores for each treatment

our treatment sizes are small, as show in Table 7.1. From our simulation study, we know that this lack of overlap can lead to erroneous results.

7.2 Results

Using the svyglm function from the survey package which uses IPTW regression, we are able to calculate coefficient estimates of the treatment effect, relative to treatment 0 (which we have defined to be no treatment). Firstly, let us look at the ATE.

	Coeff. Est.	SD	t-value	p-value
Treatment 0 (Intercept)	3.6867	0.0622	59.249	<2 e-16
Treatment 1	0.2865	0.1285	2.231	0.0262
Treatment 2	0.2240	0.0795	2.819	0.0050
Treatment 3	0.5825	0.1083	5.376	1.23e-07

Table 7.2: Multiple treatment GBM/IPTW method ATE coefficients for the covariate-adjusted Poisson regression

As we can see, all coefficients are significant at a 5 % level. From these coefficients, like the binary treatment, we can use the Poisson link function to calculate the ATE.

Treatment 1	13.2438
Treatment 2	10.0206
Treatment 3	31.5526

Table 7.3: Multiple treatment GBM/IPTW method ATE

As we can see, the treatment effect for each treatment is positive. Therefore, we conclude that with the confounders and data we have used, each treatment actually causes accidents.
We may also find it beneficial to analyse the pairwise ATT. When doing this, we must define first what the treatment of interest is - i.e. we are trying to draw inferences about the relative effectiveness of the other treatments relative to this treatment of interest. It seems natural to consider treatment 0 as the treatment of interest, because the project is about the effectiveness of applying 20mph zones/cycle superhighways to roads that are not 20mph zones/cycle superhighways.

	Coeff. Est.	SD	t-value	p-value
Treatment 0	3.5437	0.0444	79.88	<2 e-16
Treatment 1	0.3903	0.1128	3.46	0.0006
Treatment 2	0.1443	0.0776	1.86	0.0636
Treatment 3	0.6448	0.1172	5.50	6.4e-08

Table 7.4: Multiple treatment GBM/IPTW method ATE coefficients for the covariate-adjusted Poisson regression

The coefficients are significant at a 5 % level, apart from treatment 2, and do not seem to greatly differ from the ATE.

Treatment 1	16.5145
Treatment 2	5.3703
Treatment 3	31.3243

Table 7.5: Multiple treatment GBM/IPTW method ATT

So just like the binary treatment, we have received the opposite result to what we expected. The result for cycle superhighways also differs from the paper I sourced the data from; however, they were considering difference-in-difference estimation, and comparing the number of accidents from before the roads became cycle superhighways to after they became cycle superhighways. Our results suggest that we are missing some important confounders.

Chapter 8

Conclusion

8.1 Summary of Thesis Achievements

In this project, we wanted to assess whether 20mph zones in London have caused a reduction in the number of traffic accidents. This lead us to propensity score methods as they are common method for assessing treatment effect. We began our study by looking at the current research on the topic, reviewing the methods and estimators.

Next, we simulated a binary treatment scenario with a random outcome (a treatment with no effect) and a non-random outcome (a successful treatment) and applied nearest neighbourhood matching. We then analyzed the ATT. Then, we simulated a multiple treatment scenario (similarly with a non-successful, medium-successful and very-successful treatment) and used a weighted method to assess the ATE of each treatment. We found that the results were very dependent on the coefficients we chose for the distributions we sampled from. Even a treatment that should have shown no positive ATT showed a highly positive ATT if we was not careful about making sure the overlap condition was fulfilled.

Then, using data collected by researchers undertaking a similar project, and propensity score techniques, we investigated our primary interest and concluded that roads with 20mph speed limits seem to actually have higher accidents than on 20mph speed limit roads. The nearest neighbourhood matching technique seemed very successful, as it reduced the distance between

the matched roads by 99%. Using simulation methods, we also estimated an average ATT of 8.01. Then, using stratification, we estimated an average ATT of 2.72.

We also investigated our secondary interest, seeing whether the effects of 20mph zones were impacted by concurrent implementation of other traffic calming methods, namely cycle lanes. We did this by using multilevel propensity score matching, and also found that such methods are associated with roads with a higher number of accidents, with cycle superhighways having an ATE of 13.24, 20mph zones having an ATE of 10.02 and both combined having an ATE of 31.56.

There could be many reasons for this. Firstly, the conditions for conditioning on the propensity score may not have been fulfilled [17]. It is very likely that there are other confounding variables that we did not consider, the most likely being the number of accidents before the speed limit was reduced to 20mph, as it is likely that these roads had a high number of accidents beforehand anyway, which was why the speed limit implementation was necessary. Another condition that may not have been fulfilled was overlap, which we discussed the importance of in the simulation study and we illustrated was not fulfilled in Figure 7.2.

8.2 Future Work

Although our results have been negative, it is possible that we are just limited in the data we have. It is unfortunate that we do not have access to the amount of accidents before the road became a 20mph zone, so we could undertake Difference-in-Difference (DID) estimation [36]. This method focuses on measuring change in number of accidents rather than number of accidents. Indeed, it is the main sort of estimation undertaken by the study from which we source our data [30], which had much more successful results, suggesting it is better method of estimation when studying traffic accidents as it takes into account a lot of confounding variables that while may differ between 20mph and non 20mph roads, but would remain constant to each road despite the change in time. If given more time, I would have liked to have found better data. I considered in the beginning of my project, but found the only reliable source of information when it came to speed limits was Google Maps API, which was very expensive.

Another possible path that we could have explored was using trimming techniques in order to account for lack of overlap.

I am also aware that I did not assess multiple treatments as thoroughly as binary treatments. I did not use the regression method to generate propensity scores for multiple treatments, and I did not attempt the matching or stratification methods. This was because the sample sizes for the multiple treatment groups were far too small to get robust results from the matching and stratification methods. Additionally, there was a severe lack of overlap in propensity scores for treatment 1 and 3, so it was likely that the methods would not have worked. If given more time, I would have tried to use consistent methods across the project so I could compare results across binary and multiple treatments rather than considering the results in a self-contained way.

Appendix A

Simulation code

```
require(MatchIt)
require(PSAgraphics)
require(ggplot2)
require(gridExtra)
require(Zelig)
require(twang)
require(survey)

set.seed(1) #to make results reproducible
age = rexp(10000,0.1) #randomly sample distribution of age
happiness = rnorm(10000) #randomly sample distribution of happiness
health = rlnorm(10000) #randomly sample distribution of health
data=cbind(age,happiness,health)
data_name=c("age","happiness","health")

probabilitytreated=vector()
treated=vector()
coefficients = c(-0.5,-1,-1) #coefficients of probability treated based on covariates
for (i in 1:10000)
{probabilitytreated[i] = 1/(1+exp(-(sum(coefficients*data[i,])))) #generate probability of being treated
if (runif(1) < probabilitytreated[i]) #sample with appropriate probability
  {treated[i]=1}
else
{treated[i]=0}
}

data2=cbind(age,happiness,health,treated)
randomoutcome=vector()
coefficients2=c(1,-3,-3,0) #outcome dependent only on coefficients and not treatment
for (i in 1:10000)
{randomoutcome[i] = 1/(1+exp(-(sum(coefficients2*data2[i,])))) #generate probability of dying
if (runif(1) < randomoutcome[i]) #sample with appropriate probability
{randomoutcome[i]=1}
else
{randomoutcome[i]=0}
}

nonrandomoutcome=vector()
coefficients2=c(1,-3,-3,-5) #outcome dependent on coefficients and treatment
```

```
for (i in 1:10000)
{nonrandomoutcome[i] = 1/(1+exp(-(sum(coefficients2*data2[i,])))) #generate probability of dying
if (runif(1) < nonrandomoutcome[i]) #sample with appropriate probability
{nonrandomoutcome[i]=1}
else
{nonrandomoutcome[i]=0}
}

random_df = data.frame(cbind(data2,randomoutcome)) #making dataframe
nonrandom_df = data.frame(cbind(data2,nonrandomoutcome))
match1 = matchit(treated ~ age + happiness + health, method = "nearest", data=random_df) #propensity score matching
matched1 <- match.data(match1) #matched dataset with weights
summary(match1) #summary

g <- ggplot(data.frame(data2), aes(fill = factor(treated))) +
  geom_histogram(stat="bin", alpha=.5) +
  scale_fill_discrete("treated")
g1= g + aes(x = age, y = ..density..)
g2= g + aes(x = happiness, y = ..density..)
g3= g+ aes(x = health, y = ..density..)
grid.arrange(
  g1,
  g2+ theme(legend.position = "none"),
  g3,
  nrow = 2, widths = c(1, 0.8)
) #plotting distribution each cofounder to show overlap



m1=zelig(randomoutcome ~ treated + age + happiness + health, model="logit", data = matched1)
x=setx(m1, treated=0)
x1=setx(m1, treated=1)
s= sim(m1, x=x, x1=x1) #simulating treated and control outcomes for random outcome
summary(s)

m1.att <- m1%>% #calculating ATT for random outcome
  ATT(treatment = "treated", treat = 1) %>%
  get_qi(qi = "ATT", xvalue = "TE")
mean(m1.att)

m2=zelig(nonrandomoutcome ~ treated + age + happiness + health, model="logit", data = matched2)
x2=setx(m2, treated=0)
x12=setx(m2, treated=1)
s2= sim(m2, x=x2, x1=x12) #simulating treated and control outcomes for nonrandom outcome
summary(s2)

m2.att <- m2%>% #calculating ATT for nonrandom outcome
  ATT(treatment = "treated", treat = 1) %>%
  get_qi(qi = "ATT", xvalue = "TE")
mean(m2.att)

plot(match2, type="jitter") #plot of distribution of propensity score



grid.arrange(
  fn_bal(matched1, "age"),
  fn_bal(matched1, "happiness") + theme(legend.position = "none"),
  fn_bal(matched1, "health"),
  nrow = 3, widths = c(1, 0.8)
) #plotting propensity score against each confounder
```

```
100
101
102 set.seed(1)
103 multipleprobabilitytreated=matrix(, nrow = 10000, ncol = 2)
104 multipletreated=matrix(0,nrow=10000, ncol=3)
105 coefficients1 = c(-5,0,-5) #coefficients for being selected for treatment 1
106 coefficients2=c(0,2,-3) #coefficients selected for treatment 2
107 for (i in 1:10000)
108 {multipleprobabilitytreated[i,1] = exp(sum(coefficients1*data[i,]))/(1+exp(sum(coefficients1*data[i,]))+exp(sum(coefficients2*
      data[i,])))
109 multipleprobabilitytreated[i,2] = multipleprobabilitytreated[i,1]+exp(sum(coefficients2*data[i,]))/(1+exp(sum(coefficients1*
      data[i,]))+exp(sum(coefficients2*data[i,])))
110 randnum=runif(1)
111 if (randnum < multipleprobabilitytreated[i,1]) {multipletreated[i,2]=1} #randomly sample from multinomial distribution, then
      assign to dummy variable
112 else if (randnum < multipleprobabilitytreated[i,2]) {multipletreated[i,3]=1}
113 else {multipletreated[i,1]=1}}
114
115
116 data3=cbind(age,happiness,health,multipletreated)
117
118 multipleoutcome=vector()
119 coefficients2=c(1,-3,-3,0,-2.5,-5)
120 for (i in 1:10000)
121 {multipleoutcome[i] = 1/(1+exp(-(sum(coefficients2*data3[i,]))))
122 if (runif(1) < multipleoutcome[i]) #sample from bernoulli distribution for outcome
123 {multipleoutcome[i]=1}
124 else
125 {multipleoutcome[i]=0}
126 }
127
128 multvector=vector()  #convert from dummy variables to single variable with 3 factors
129 for (i in 1:10000)
130   {if (multipletreated[i,1]==1)
131   {multvector[i]=0}
132   else if (multipletreated[i,2]==1)
133   {multvector[i]=1}
134   else
135   {multvector[i]=2}}
136
137 data4=as.data.frame(cbind(age,happiness,health,multvector,multipleoutcome))
138 data4$multvector=as.factor(multvector)
139
140 g <- ggplot(data.frame(data4), aes(fill = factor(multvector))) + #plot covariate distributions
141   geom_histogram(stat="bin", alpha=.5) +
142   scale_fill_discrete("multvector")
143 g1= g + aes(x = age, y = ..density..)
144 g2= g + aes(x = happiness, y = ..density..)
145 g3= g+ aes(x = health, y = ..density..)
146 grid.arrange(
147   g1,
148   g2+ theme(legend.position = "none"),
149   g3,
150   nrow = 2, widths = c(1, 0.8)
151 )
152
153
154 mnps.AOD <- mnps(multvector ~ age + happiness + health, data = data4, estimand = "ATE", verbose = FALSE, stop.method = c("ks.
      mean"), n.trees = 3000) #finding IPTW weights
155  plot(mnps.AOD)
```

```
156 data4$w <- get.weights(mnps.AOD, stop.method = "ks.mean")
157 design.mnps <- svydesign(ids=~1, weights=~w, data=data4)
158 glm1 <- svyglm(multipleoutcome~ multvector, design = design.mnps, family="binomial")
159 summary(glm1) #finding covariate adjusted regression coefficietns
160
161 a=1/(1+exp(-(0.68996)))
162 b=1/(1+exp(-(0.68996-14.14594)))
163 c=1/(1+exp(-(0.68996-0.95898)))
164
165 a #calculating ATE using bernoulli link function
166 b-a
167 c-a
```

Appendix B

Data code

```
require(MatchIt)
require(plyr)
require(twang)
require(Zelig)
require(survey)
require(dplyr)
require(ggpubr)
require(reshape)
require(ggplot2)
require(cobalt)

#importing data and then selecting the points i want to use
data2<- read.csv("20data.csv")
combined <- read.csv("combined.csv")
data = data2[,-1]
data$treatment=as.factor(data$treatment)
data$X20mphzone=as.factor(data$X20mphzone)
dataused=select(data, X20mphzone, treatment, length, AADFTotal, AADFPedal, AADFBus, TotalCycleAcc, SlightCycleAcc, AllAcc,
     AllAccSlight, AllAccKsi, Busstops, Speed, Domes, Non_Dom, RoadArea, IMD, Pop, Employ)
datanow=dataused[451:900,]


#preliminary data analysis
summary(datanow)

#plotting speed, AllAcc and length, coloured by X20mphzone
g <- ggplot(datanow, aes(fill = factor(X20mphzone))) +
  geom_density(stat="bin", alpha=.5) +
  scale_fill_discrete("X20mphzone")
g + aes(x = Speed)
g + aes(x = AllAcc)
g + aes(x = length)

datanow2=datanow
combined =melt(datanow2[,c(1,7:11)], id = "X20mphzone")
combined2=aggregate(value ~ X20mphzone + variable, FUN = mean, combined)

#plotting accident type by 20mphzone
t<- ggplot(combined2, aes(X20mphzone,value)) + geom_bar(aes(fill = variable), position = "dodge", stat="identity")
```

```
t

#plotting accident by length
p <- ggplot(datanow, aes(x=length, y=AllAcc,colour=X20mphzone)) + geom_point()
p

#NN matching
match = matchit(X20mphzone ~ treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +
     RoadArea +IMD + Pop + Employ, method = "nearest", data = datanow, replace=TRUE)
summary(match) #checking balance
matched <- match.data(match)

#plots of propensity score distribution
plot(match1, type="jitter")
plot(match1, type="QQ")


fn_bal <- function(dta, variable) {
  dta$variable <- dta[, variable]
  dta$X20mphzone <- as.factor(dta$X20mphzone)
  support <- c(min(dta$variable), max(dta$variable))
  ggplot(dta, aes(x = distance, y = variable, color = X20mphzone)) +
    geom_point(alpha = 0.2, size = 1.3) +
    geom_smooth(method = "loess", se = F) +
    xlab("Propensity score") +
    ylab(variable) +
    theme_bw() +
    ylim(support)
}

#plotting variables against propensity score coloured by X20mphzone
library(gridExtra)
grid.arrange(
  fn_bal(matched, "AADFBus"),
  fn_bal(matched, "IMD") + theme(legend.position = "none"),
  fn_bal(matched, "length"),
  fn_bal(matched, "Pop") + theme(legend.position = "none"),
  nrow = 2, widths = c(1, 0.8)
)

#using matched data to simulate using zelig
z=zelig(AllAcc ~ X20mphzone + treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +
     RoadArea +IMD + Pop + Employ, model = "poisson", data=matched)
x=setx(z, X20mphzone=0)
x1=setx(z, X20mphzone=1)
s=sim(z, x=x, x1=x1)
summary(s)
plot(s)

#calculating ATT and quantiles
z.att <- z%>%
  ATT(treatment = "X20mphzone", treat = 1) %>%
  get_qi(qi = "ATT", xvalue = "TE")
mean(z.att)
sd(z.att)
quantile(z.att, c(0.025, 0.5, 0.975))

#ATT histogram
hist(z.att,
     main = NULL,
```

```
97      xlab ="ATT of a 20mph Zone",freq=FALSE)
98
99  #calculating ATE from ATT, and finding quantiles
100 z.att1 <- z%>%
101   ATT(treatment = "X20mphzone", treat = 1) %>%
102   get_qi(qi = "ATT", xvalue = "TE")
103 z.att2 <- z%>%
104   ATT(treatment = "X20mphzone", treat = 0) %>%
105   get_qi(qi = "ATT", xvalue = "TE")
106 ate.all <- c(z.att1, -z.att2)
107 mean(ate.all)
108 sd(ate.all)
109 quantile(ate.all, c(0.025, 0.5, 0.975))
110
111 #stratification method
112 matchstrata = matchit(X20mphzone ~ treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +
      RoadArea +IMD + Pop + Employ, method = "subclass", data = datanow)
113 matchedstrata <- match.data(matchstrata)
114
115 #balance tables for nn method and strat method
116 bal.tab(matchstrata)
117 bal.tab(match)
118
119 #zelig simulations
120 zstrata=zelig(AllAcc ~ X20mphzone + treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom
      +RoadArea +IMD + Pop + Employ, model = "poisson", data=matchedstrata, by="subclass")
121 xstrata=setx(z, X20mphzone=0)
122 x1strata=setx(z, X20mphzone=1)
123 sstrata=sim(zstrata,x=xstrata,x1=x1strata)
124 summary(sstrata)
125
126 #calculating ATT and quantiles
127 z.att <- zstrata %>%
128   ATT(treatment = "X20mphzone", treat = 1) %>%
129   get_qi(qi = "ATT", xvalue = "TE")
130 mean(z.att)
131 sd(z.att)
132 quantile(z.att, c(0.025, 0.5, 0.975))
133
134 #calculating ATE from ATT and quantiles
135 z.att1 <- zstrata%>%
136   ATT(treatment = "X20mphzone", treat = 1) %>%
137   get_qi(qi = "ATT", xvalue = "TE")
138 z.att2 <- zstrata%>%
139   ATT(treatment = "X20mphzone", treat = 0) %>%
140   get_qi(qi = "ATT", xvalue = "TE")
141 ate.all <- c(z.att1, -z.att2)
142 mean(ate.all)
143 sd(ate.all)
144 quantile(ate.all, c(0.025, 0.5, 0.975))
145
146 #GBM and IPTW method for binary ATE
147 data2=datanow
148 data2$X20mphzone=as.numeric(datanow$X20mphzone)
149 data3=data2
150 data3$X20mphzone=data2$X20mphzone-1
151 ps.AOD <- ps(X20mphzone ~ treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +RoadArea
      +IMD + Pop + Employ, data = data3, estimand = "ATE", verbose = FALSE, stop.method = c("ks.mean"), n.trees = 3000)
152
153 #using svglm to find coefficients for binary ATE
```

```
data3$w <- get.weights(ps.AOD, stop.method = "ks.mean")
design.ps <- svydesign(ids=~1, weights=~w, data=data3)
glm1 <- svyglm(AllAcc ~ as.factor(X20mphzone), design = design.ps, family = "poisson")
summary(glm1)

#calculating binary ATE
ATE=exp(3.69218+0.26530)-exp(3.69218)
ATE

#GBM and IPTW method for binary ATT
dataatt=data3
ps.AOD2 <- ps(X20mphzone ~ treatment + length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +RoadArea
      +IMD + Pop + Employ, data = dataatt, estimand = "ATT", verbose = FALSE, stop.method = c("ks.mean"), n.trees = 3000)

#using svglm to find coefficients for binary ATT
dataatt$w <- get.weights(ps.AOD2, stop.method = "ks.mean")
design.ps <- svydesign(ids=~1, weights=~w, data=dataatt)
glm2 <- svyglm(AllAcc ~ as.factor(X20mphzone), design = design.ps, family = "poisson")
summary(glm2)

#calculating binary ATT
ATT=exp(3.82403+0.26232)-exp(3.82403)
ATT

#finding diff.adj
bal.tab(ps.AOD)
bal.tab(ps.AOD2)

#chi-squared test for independence
chi=matrix(c(199,31,176,44),nrow=2,ncol=2)
colnames(chi) <- c("notcycle","cycle")
rownames(chi) <- c("not20","20")
chi<- as.table(chi)
chisq <- chisq.test(chi)
chisq

#GBM and IPTW method for multiple ATE
datamult=datanow
datamult$treat=as.factor(as.numeric(datanow$treatment)+2*as.numeric(datanow$X20mphzone)-3)
summary(datamult$treat)
mnps.AOD <- mnps(treat ~ length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +RoadArea +IMD + Pop +
     Employ, data = datamult, estimand = "ATE", verbose = FALSE, stop.method = c("ks.mean"), n.trees = 3000)
plot(mnps.AOD)


#using svglm to find coefficients for multiple ATE
datamult$w <- get.weights(mnps.AOD, stop.method = "ks.mean")
design.mnps <- svydesign(ids=~1, weights=~w, data=datamult)
glm1 <- svyglm(AllAcc ~ as.factor(treat), design = design.mnps, family="poisson")
summary(glm1)

#calculating multiple ATE
ATE1= exp(3.68671+0.28654)-exp(3.68671)
ATE2=exp(3.68671+0.22399)-exp(3.68671)
ATE3=exp(3.68671+0.58251)-exp(3.68671)
ATE1
ATE2
ATE3

#GBM and IPTW method for multiple ATT
```

```
datamult2=datanow
datamult2$treat=as.factor(as.numeric(datanow$treatment)+2*as.numeric(datanow$X20mphzone)-3)
mnps.AOD2 <- mnps(treat ~ length  + AADFTotal + AADFPedal +AADFBus + Busstops + Speed + Domes + Non_Dom +RoadArea +IMD + Pop +
      Employ, data = datamult2, estimand = "ATT", treatATT= 0, verbose = FALSE, stop.method = c("ks.mean"), n.trees = 3000)

#using svglm to find coefficients for multiple ATT
datamult2$w <- get.weights(mnps.AOD2, stop.method = "ks.mean")
design.mnps <- svydesign(ids=~1, weights=~w, data=datamult2)
glm2 <- svyglm(AllAcc ~ as.factor(treat), design = design.mnps, family="poisson")
summary(glm2)

#calculating multiple ATT
ATT1= exp(3.54365+0.39028)-exp(3.54365)
ATT2=exp(3.54365+0.14431)-exp(3.54365)
ATT3=exp(3.54365+0.64475 )-exp(3.54365)
ATT1
ATT2
ATT3
```

Bibliography

[1] Atkins, AECOM, and P. M. Maher, "20mph research study - process and impact evaluation: technical report," Nov 2018.

[2] BBC, "20mph speed limit to be set on central london roads," Jul 2018. [Online]. Available: https://www.bbc.co.uk/news/uk-england-london-44935036

[3] L. Aarts and I. van Schagen, "Driving speed and the risk of road crashes: A review," *Accident Analysis Prevention*, vol. 38, no. 2, pp. 215 – 224, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457505001247

[4] TfL, "Safe streets for london: The road safety action plan for london 2020," Jul 2013.

[5] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. Burgette, "A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models," *Stat Med*, 8 2013.

[6] P. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioural Research*, vol. 46, no. 3, p. 399424, 2011.

[7] L. Grilli and C. Rampichini, "Propensity scores for the estimation of average treatment effects in observational studies," Jun 2011.

[8] P. R. Rosenbaum and D. B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, vol. 39, no. 1, pp. 33–38, 1985.

[9] W. Cochran, "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*, vol. 24, no. 2, pp. 295–313, 1968.

[10] P. R. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 387–394, 1987. [Online]. Available: http://www.jstor.org/stable/2289440

[11] Y. Ye and L. Kaskutas, "Using propensity scores to adjust for selection bias when assessing the effectiveness of Alcoholics Anonymous in observational studies," *Drug and Alcohol Dependence*, vol. 104, no. 1-2, pp. 56–64, 2009.

[12] P. Austin, "Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses," *The International Journal of Biostatistics*, vol. 5, no. 1, 2009.

[13] P. Austin and M. Mamdani, "A comparison of propensity score methods: A case-study estimating the effectiveness of post-ami statin use," *Statistics in medicine*, vol. 25, pp. 2084–106, 06 2006.

[14] Y. Zhu, M. Schonbach, D. Coffman, and J. Williams, "Variable selection for propensity score estimation via balancing covariates," *Epidemiology*, vol. 26, no. 2, 2015.

[15] M. J. Lopez and R. Gutman, "Estimation of causal effects with multiple treatments: a review and new ideas," *Statistical Science*, vol. 32, no. 3, 2017.

[16] M. Hernn and J. Robins, *Causal Inference*. Boca Raton: Chapman Hall/CRC, forthcoming, 2019.

[17] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 04 1983. [Online]. Available: https://doi.org/10.1093/biomet/70.1.41

[18] D. Westreich, J. Lessler, and M. J. Funk, "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression," *Journal of Clinical Epidemiology*, vol. 63, no. 8, pp.

826 – 833, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895435610001022

[19] S. O. Becker and A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, vol. 2, no. 4, pp. 358–377, 2002.

[20] G. W. Imbens, "The role of the propensity score in estimating dose-response functions," *Biometrika*, vol. 87, no. 3, pp. 706–710, 2000. [Online]. Available: http://www.jstor.org/stable/2673642

[21] K. Imai and D. A. van Dyk, "Causal inference with general treatment regimes," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 854–866, 2004. [Online]. Available: https://doi.org/10.1198/016214504000001187

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[23] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: http://ggplot2.org

[24] W. Foundation, May 2019. [Online]. Available: https://en.wikipedia.org/wiki/Multinomial_logistic_regression

[25] D. E. Ho, K. Imai, G. King, and E. A. Stuart, "MatchIt: Nonparametric preprocessing for parametric causal inference," *Journal of Statistical Software*, vol. 42, no. 8, pp. 1–28, 2011. [Online]. Available: http://www.jstatsoft.org/v42/i08/

[26] L. Burgette, B. A. Griffin, and D. McCaffrey, "Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package," Jul 2017.

[27] R. Altman, "Lecture 25: The dispersion parameter," 2009. [Online]. Available: Lecture25:TheDispersionParameter

[28] Y. Zheng and L. Stapleton, "Effects of propensity score overlap on the estimates of treatment effects," 2018.

[29] M. Caliendo and S. Kopeinig, "Some practical guidance for the implementation of propensity score matching," *Journal of Economic Surveys*, vol. 22, no. 1, pp. 31–72, 2008.

[30] H. Li, D. J. Graham, and P. Liu, "Safety effects of the london cycle superhighways on cycle collisions," *Accident Analysis Prevention*, vol. 99, pp. 90 – 101, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457516304298

[31] K. Imai, G. King, and O. Lau, "Toward a common framework for statistical analysis and development," *Journal of Computational Graphics and Statistics*, vol. 17, no. 4, pp. 892–913, 2008. [Online]. Available: http://j.mp/msE15c

[32] G. King, "Examples," Dec 2010. [Online]. Available: https://r.iq.harvard.edu/docs/matchit/2.4-15/Examples2.html

[33] N. Greifer, *cobalt: Covariate Balance Tables and Plots*, 2019, r package version 3.7.0. [Online]. Available: https://CRAN.R-project.org/package=cobalt

[34] G. Ridgeway, D. McCaffrey, A. Morral, B. A. Griffin, and L. Burgette, *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*, 2017, r package version 1.5. [Online]. Available: https://CRAN.R-project.org/package=twang

[35] T. Lumley, "Analysis of complex survey samples," *Journal of Statistical Software*, vol. 9, no. 1, pp. 1–19, 2004, r package version 2.2.

[36] C. U. M. S. of Public Health, "Difference-in-difference estimation," Apr 2013. [Online]. Available: https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation