097400 Causal Inference
Semester 2021w
Lior Danino ID: [redacted]
Daniel Goman ID: [redacted]

Final Submission

Introduction

Acts of crime and extreme violence have been an issue since ancient times, and even today the law forces are trying to address this issue in several ways. One of the ways this is approached is by having law enforcement forces respond to the event in real time – those being the police forces.

With the advanced technology we have nowadays, there are cameras that can records things at high quality, for example: movie cameras, phone cameras, car cameras and so on.
One innovation that followed the abovementioned ideas was the policemen body camera.
The body camera is a wearable audio, video, or photographic recording system used to record events in which law enforcement officers are involved.

The essence of policemen body cameras is to allow the policemen provide solid evidence and details of incidents without having the integrity and accuracy of their report double-checked.
The incorporation of body camera usage among the police forces also works in another way – it gives more assurance to the public that justice is delivered properly, and that the rights of the suspects are not being unnecessarily violated. On top of that, in states where the usage of body cameras is mandatory, it helps preventing any case of abuse-of-power by the law enforcement forces.

In our work, we examine a dataset that contains records of people killed by policemen, which is discussed in the Data section.

Our work focuses on two features of the incidents that exist in the dataset:

1. body_camera – a feature which says whether the policeman who killed the suspect wore a body camera while the event occurred or not.
   Available values: {FALSE, TRUE}, FALSE meaning that the policeman wasn't wearing a body camera, and TRUE meaning that he did wear one.
2. manner_of_death – a feature which says whether the suspect was shot to death, or tasered and shot to death.
   Available values: {shot, shot and Tasered}

We attempt to grasp the effect of wearing a body camera on the manner in which a suspect is killed.

Formally speaking, our research question is: "Does wearing a body camera affect the level of policemen violence on alleged threats?" where we regard to "policemen violence" as straight away shooting the suspects without attempting to taser them first.

Our thought process behind this is as follows: shooting suspects without attempting to apprehend them by using non-lethal approaches such as a tasering them first is more violent, while attempting to taser the suspects first is not as violent.

Thus, our research question is: Does wearing a body camera makes policemen attempt to apprehend suspects by tasering them first more often than when not wearing a body camera?

The treatment is **wearing a body camera**.

The outcome is the **manner of death**.

Our hypothesis is that policemen who wear a body camera are more likely to taser and shoot a suspect instead of immediately shooting them.

Data

The dataset we explore in this work is a dataset consisting of records of incidents where a suspect was killed by policemen in the US in the years 2015-2020.

The dataset has the following features:

Feature	Meaning	Available values
manner_of_death	The way the suspect was killed	• shot • shot and Tasered
armed	The weapon the suspect was armed with during the incident	89 different weapons
age	The age of the suspect	Various ages from 6 to 91
gender	The gender of the suspect	1. Male 2. Female
race	The race of the suspect	Asian, White, Hispanic, Black, Other, Native
city	The city in which the	2288 cities in the US

	incident occurred	
state	The state in which the incident occurred	51 states of the US
signs_of_mental_illness	Whether the suspect showed signs of mental illness	1. True 2. False
threat_level	The level of threat the suspect possessed	1. attack 2. other 3. undetermined
flee	Whether the suspect was trying to get away	1. not fleeing 2. car 3. foot 4. other
body_camera	Whether the policeman that killed the suspect wore a bodycam	1. True 2. False
Arms_category	The type of weapon the suspect had	12 types of weapons

There are a few more features which we did not use so they're not mentioned here (features such as id, name, etc).

We found no missing data and no erroneous data.

In our code we encode the features into numerical values, where each available value is mapped to some integer (starting from 1).

For the features body_camera and manner_of_death we specifically define:

- body_camera: $False \to 0, True \to 1$
- manner_of_death: $'shot' \to 0, 'shot\ and\ Tasered' \to 1$

Assumptions

As we mentioned before, we are looking to evaluate the Average Treatment Effect of wearing a body camera on the manner of death of the suspects. In this setting both

treatment and outcome features are binary, while some other features present in the dataset are categorial.

At this point we would like to discuss the assumptions that upload with respect to our dataset:

SUTVA – as far as we know, each record in the dataset is a record of an incident where a suspect was killed, which is independent of other records in the dataset, meaning the potential outcomes for any suspect is independent of the treatment other suspects received.
Thus, the SUTVA assumption upholds.

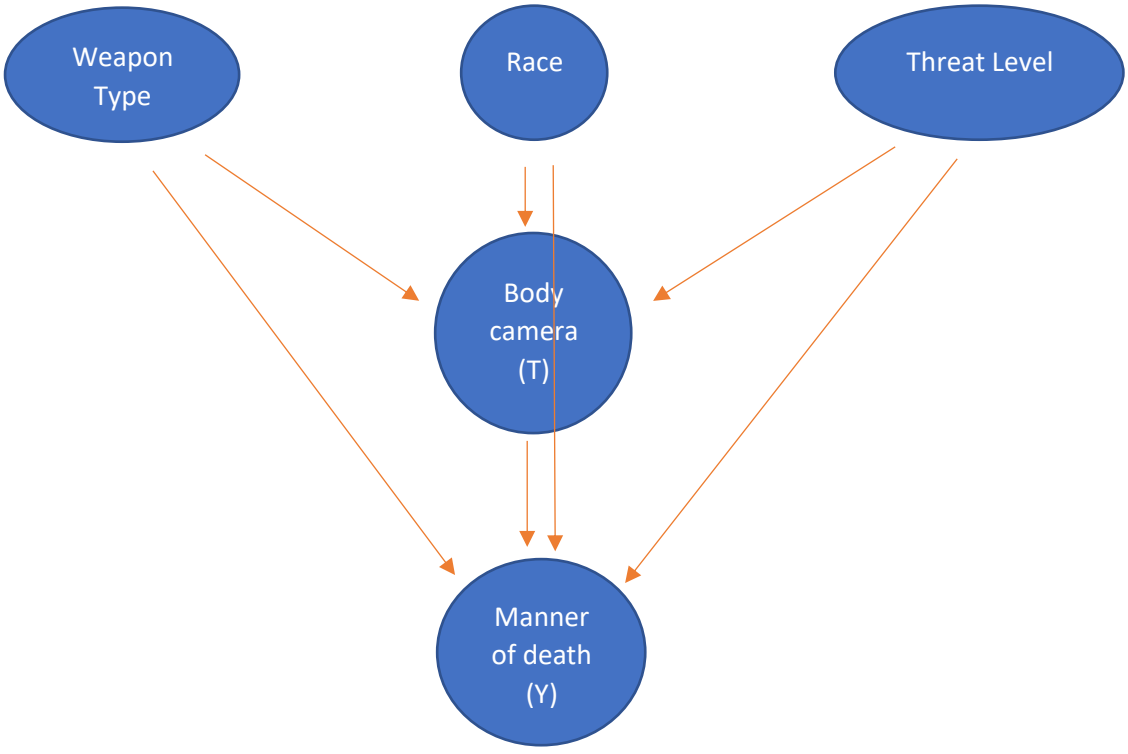
Consistency – we didn't observe any erroneous data, which leads us to assume the data is indeed intact and has valid information regarding the manner of death of each suspect.
Thus, the Consistency assumption upholds.

Ignorability – we cannot guarantee that the ignorability assumption upholds, as we do not observe all the information regarding the incidents. For example, there's no information regarding the location of the event, the amount of respondent policemen, the number of suspects, the gender of the policemen, etc, information that could prove to be a confounder in our settings. Moreover, even if it says in the dataset that the policeman wore the body camera, there's no guarantee that it was activate on scene. In total, the ignorability assumption does not hold.

Common support – operating under the assumption that all states in the US have body cameras at their disposal, the Common support assumption upholds, as the policemen who killed a suspect had the option to wear a body camera. This gives us that

$$P(T = t \mid X = x) > 0 \; \forall t, x$$

Confounders – We suspect that we have some confounders in our data, those confounders being: $\{'arms\_category', 'race', 'threat\_level'\}$, which can be represented in the following manner:



In our work we will attempt to estimate the causal effect of wearing a body camera on the manner of death, while avoiding the confounding that may be present in the data.

Methods

Our goal is to estimate the Average Treatment Effect of a policeman wearing body camera on the manner of death of the suspect who was killed by the that policeman.

For the purpose of estimating the ATE, we employ several methods which attempt to estimate the true causal effect of the former on the latter.

Some of the methods we used require the usage of ML models of our choice. We decided to use Random Forest Classifier, as our data is mostly categorial, and so we thought RFC is an adequate model to employ in such settings.

The methods we used are as follows:

Naive ATE:

First of all, just for reference, we calculated the ATE in the naive manner: simply calculating the difference in outcome on the treated and control.
Mathematically speaking:

Denote:

$$Treated \coloneqq \{i \; s.t. \; t_i = 1\}$$
$$Control \coloneqq \{i \; s.t. \; t_i = 0\}$$

We then calculated the naive ATE as follows:

$$\widehat{ATE} = \frac{1}{|Treated|} \sum_{i \in Treated} y_i - \frac{1}{|Control|} \sum_{i \in Control} y_i$$

IPW:

A method to estimate ATE using propensity score.
IPW tries to remove confounding by re-weighting samples using the propensity scores, which creates “new” samples in which treatment is independent of the confounders.

The algorithm:
1. Given a sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$, estimate the propensity score using a

selected machine-learning method.
2. Estimate ATE using the following formula:

$$\widehat{ATE} = \frac{1}{n}\sum_{i\ s.t.\ t_i=1}\frac{y_i}{\hat{p}(t_i=1|x_i)} - \frac{1}{n}\sum_{i\ s.t.\ t_i=0}\frac{y_i}{\hat{p}(t_i=0|x_i)}$$

In our project, we estimated $\hat{p}(T=t|x)$ using Random Forest Classifier.

1-NN Matching:

A method to estimate ATE based on matching pairs of samples.

The algorithm:
1. Given a sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$, calculate the distances between all pairs of samples using a selected distance metric.
2. Using 1-NN, match each sample i its nearest neighbor $j(i)$ from opposite treatment group.
3. Calculate $\widehat{ITE}(i)$ by the following formula:

$$\widehat{ITE}(i) = \begin{cases} y_i - y_{j(i)} & , t_i = 1 \\ y_{j(i)} - y_i & , t_i = 0 \end{cases}$$

4. Estimate ATE using the following formula:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\widehat{ITE}(i)$$

In our project, the features used to calculate the distance were all nominal categorical features. Therefore, the distance metric used was hamming distance: for two different strings, the hamming distance between them is the number of character positions in which they differ.
In our case, instead of strings, we have vectors where each entrance represents a feature and contains its value.
This metric works well for nominal categorical data since it ignores the specific values and any order between them.

S-Learner (Single Learner):

For this method we employ the following mindset: what if the outcome on the treated and the control could be represented through some model which, given a row of data, would determine the outcome on that specific row.

In other words, this approach attempts to fit a model to the data, with the target feature being the outcome feature which we're interested in.

This allows us to estimate outcomes for scenarios we don't have in the real world, for example, we don't know y_0 for suspect who got $t = 1$, we only know y_1.

After fitting a model M, we proceed to going over all of the data, and predict the outcome using the model, twice for each record, once as treatment and once as

control.
Meaning, for record x_i we set $t_i = 1$ and predict the outcome using the model M, which we denote as $M(x_i, 1)$, and then we set $t_i = 0$ and predict $M(x_i, 0)$.
This is done for all x_i in the dataset. Then we calculate the ATE in the following way:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} M(x_i, 1) - M(x_i, 0)$$

In our project we use a Random Forest Classifier to predict the outcome.

T-Learner (Two Learner):

For this method we employ a similar mindset to the one employed for the S-learner, where we assume the outcome could be represented as some function of the data, and we predict the complementary scenario which is not present in the data.

The difference from the S-Learner is that the T-Learner creates two different models, one on the data of the treated, to predict the outcome for the treated, and one on the data of the control, to predict the outcome for the control.

We define the two models M_1 and M_0 which are trained on $\{x_i \; s.t. \; t_i = 1\}$ and $\{x_i \; s.t. \; t_i = 0\}$ respectively.

We then proceed to estimate the ATE as follows:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} M_1(x_i) - M_0(x_i)$$

In our project we use a Random Forest Classifier to predict the outcome for both M_1 and M_0.

Backdoor adjustment:

For this method we use do-calculus. To allow this we employ backdoor adjustment, which says that given the confounders, the expected outcome under the do operator boils down to the expected outcome given the confounders and T.

Formally speaking:

Denote the group of confounders as c. We then have that:

$$\widehat{ATE} = E[y \mid do(T = 1)] - E[y \mid do(T = 0)]$$
$$= E_c\big[E[y \mid c, T = 1]\big] - E_c\big[E[y \mid c, T = 0]\big]$$

In our data we calculate this assuming the confounders we have in the data are $\{'arms\_category', 'race', 'threat\_level'\}$, and so the utility over c goes over all existing combinations of the available values for these features.

The probabilities required to calculate the abovementioned utilities are estimated from the data.

Results

Table 1 – ATE results of selected estimation methods:

Estimation method	ATE
Naive ATE	0.0132
IPW	0.0011
S-Learner	−0.0014
T-Learner	−0.001
Matching	0.0239
Backdoor adjustment	0.0012

In this table we display the results of the ATE from running the various methods we went over in detail earlier.

Table 2 – Testing various subsets of suspected confounders:

Features used for backdoor adjustment	ATE
'arms_category'	0.0046
'race'	0.0134
'threat_level'	0.0116
'race', 'threat_level'	0.0125
'race', 'arms_category'	−0.0003
'threat_level', 'arms_category'	0.0062
'threat_level', 'race', 'arms_category'	0.0012

In this table we display the results of running backdoor adjustment with different subsets of features out of the features we suspect to be confounders. (the result of backdoor adjustment which appears in Table 1 is for running it with all 3 suspected confounders)

Table 3 – Model sensitivity:

Estimation method	Random Forest Classifier	Decision Tree Classifier
IPW	0.0011	0.0012
S-Learner	−0.0014	−0.0004
T-Learner	−0.001	−0.001

In this table we experiment using Decision Tree Classifier in comparison to Random Forest Classifier in the estimation methods which use ML models.

Possible Weaknesses

As we have mentioned earlier in the Assumptions - Ignorability section, the ignorability assumption does not hold, which may harm the reliability of our results. Ignorability doesn't hold as we cannot guarantee that there are no hidden confounders. For instance, having a policeman wear a body camera does not necessarily mean it was active on scene, which may sound ridiculous, but we have found information of such cases actually happening.

Moreover, some possible hidden confounders could be the location of the event, the amount of respondent policemen, the number of suspects, the gender of the policemen and so on. All those and more can potentially affect the decision of wearing a body camera, as well as the decision of the approach to apprehend the suspects on scene.

One more potential downfall of our approach lies in the assumption that all states in the US have body cameras at their disposal, which perhaps didn't hold true in our dataset.

The SUTVA assumption may not hold in cases of large incidents where more than one suspect was shot to death as some records in the dataset may be correlated.

Discussion

In this work we employ several methods in an attempt to estimate the Average Treatment Effect of having a body camera worn by a policeman on the manner of death of a suspect who was killed by that policeman.

Observing the results in Table 1, the naive ATE we got in comparison to rest of the results excluding Matching, is larger.
According to the way we encoded our treatment and outcome variables, this implies that wearing a body camera increases the chances of the respective suspect to be shot and tasered, instead of immediately being shot.

When observing the rest of the results excluding Matching, we actually see that the results mostly imply that there's no causal effect between the former and the latter. From this we may conclude that there is confounding to some extent.

Observing the results in Table 2, wherever we condition on 'arms_category' the ATE becomes lower, which may imply that 'arms_category' is a confounder in the dataset, while the other two – 'race' and 'threat_level' are not as much if at all.

From observing the results in Table 3, which are quite similar for both models, it seems that the methods which use ML models are not sensitive to model selection.

Overall, from our tests it seems that although the naive method implies that there's a causal effect between a policeman wearing a body camera and the manner of death

of a suspect which was apprehended by that policemen, there actually is no causal effect.

The one exception we have is the results of the Matching method we used, but as the rest of the methods agree on the same result, we conclude that there's no causal effect between the former and the latter.

Appendix

Link to git rep: https://github.com/DanielGoman/Causal-Inference-Final-Project