**Estimating the Average Treatment Effect of New Liquor Establishments on Crime Rates in Chicago**

Patrick Phillips

**Problem Summary**

Do food deserts lead to crime? Does liquor fuel violent tendencies? I try to answer these questions by identifying the influence of new liquor stores and new grocery stores on the local crime rate. Below is a map of the basic problem setup (only a fraction of the crime occurrences are plotted for visibility). The data consists of new liquor licenses or grocery store licenses, crime occurrences, and covariate information including socioeconomic factors, and population health statistics for each community area.

Graphical user interface, application

Description automatically generatedMap

Description automatically generatedChart

Description automatically generated

In the zoomed in visualization on the right of the UChicago campus area we notice a couple of potential difficulties: (1) the region around a new liquor establishment might intersect two or more community areas (which is the level of description that covariate information is available for) and (2) the region may also go outside of the city limits. To resolve the first issue, I take a weighted average of the covariate information between the two community areas. The second issue I leave unchecked, as the fraction of cases that have this type of overlap is less than 5% and accounting for population size, historical crime rate, and other covariates helps mask the issue as well.

Another potential problem is that the regions around new liquor establishments may overlap. To avoid this issue, I set the size of regions to 0.4km by 0.4km which I found was the largest area where there are no overlapping regions between licenses issued within +-3 months. I did experiment with circular vs square regions and bigger and smaller regions and saw no qualitative difference in results, even when regions began to overlap. In general, I observed that new liquor establishments did lead to increased amounts of crime. I was not able to find any significant effect of new grocery stores on the amount of surrounding crime.

**Causal Identification**

I estimate the Average Treatment Effect (ATE) of a new liquor establishment or grocery store on the crime rate in the surrounding region. Using the assumptions below, we can show the ATE is equivalent to the efficient double machine-learning Augmented Inverse Probability of Treatment Weighted (AIPTW) estimator.

A picture containing text, clock

Description automatically generated

**A: Treatment –** A new liquor establishment is added to an area (I consider the same setup where treatments are a new grocery store separately). More specifically the treated groups (A=1) are the 0.4km by 0.4km square around a new liquor establishment. The control groups (A=0) are 0.4km by 0.4km regions where no new liquor establishment was created in the past +-3 months.

**X: Confounders** – Socioeconomic factors for the area, historical crime rate in the area, population change of the area, population health statistics, time period (year). Except for historical crime rate and time period we only have data stratified at the community area level.

**Y: Outcome** – Subsequent change in number of crimes in the area over 3 months after treatment.

**Identification Assumptions**

1. (**Confounders Observed**) There are many confounders which one could imagine are left out (e.g. number of previous liquor stores/grocery stores could affect both the number of crimes and whether a new store is built in the area). I assume conditioning on the covariates I have included, in particular the historical violent crime in each area of consideration, is going to suffice to make the effects of these confounders small (i.e. the crime rate after the new grocery store is built is approximately conditionally independent of the previous grocery stores in the area given the historical crime rate). The later sensitivity analysis supports the claim that the unobserved confounders would likely have little impact.
2. (**Overlap**) The condition 0<P(a=1|x)<1 is largely satisfied with the constructed data. Control units are created using rejection sampling of random points in the same community are. Thus for each treated unit we are guaranteed to have a control unit with all confounders except crime rate the same. We can see that the propensity scores are generally around 0.5 due to this construction.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated  Figure : Liquor License Analysis. Random Forest g(x). | Chart, histogram  Description automatically generated  Figure : Grocery Store Analysis. Random Forest g(x). |

**The Data**

The [crime data set](https://catalog.data.gov/dataset/crimes-2001-to-present) has information from 2001 until one week ago (12/1/2022). The crimes are broken down by type below, with the most reported crime being theft. I do separate analysis for the effects on overall crime and the effects in particular on the number of violent crimes. There are roughly 10 million crimes overall of which 3.5 million are violent crimes (I include theft as violent crime to keep the number of crimes reasonably far from zero in each region).

**Chart

Description automatically generated**

Looking at this dataset, it becomes clear how important it is to control for global trends as we can see in the figure below. The number of crimes is both decreasing and cyclical, with peaks in crime during summer months. I make the outcome change in the number of crimes and I also make the historical number of crimes a covariate to take into account these global trends.

**Graphical user interface, chart, line chart

Description automatically generated**

The [business license dataset](https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr) is the key data used to find new liquor establishments and new grocery stores. There are roughly 1800 usable liquor establishments and 200 new grocery stores (joining with [this separate dataset](https://data.cityofchicago.org/Community-Economic-Development/Grocery-Stores-2013/53t8-wyrc) was required to identify grocery stores). For each treated unit I construct a control unit in the same region using a rejection sampling procedure which ensures no nearby liquor stores were created within the relevant date range.

The covariates include the historical number of crimes in the region, time period, and data from two datasets (1) [Public Health Statistics](https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu) and (2) [Socioeconomic Factors](https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2) both of which have information on each of the 77 community areas in Chicago. There are 30 covariates in total.

**Analysis**

Assuming the identification argument holds, the ATE can be interpreted as the number of additional crimes caused by introducing a new liquor establishment (or new grocery store) over the three months following in the surrounding region. I tabulate values of the ATE and its 0.95 confidence interval below for the different analysis scenarios. We see that new liquor establishments increase both the amount of crime and the amount of violent crime in the surrounding area. Meanwhile new grocery stores do not have a significant effect on nearby crime.

Table 1: Average Treatment Effect of New Liquor Establishments on Crime

|  |  |  |  |
| --- | --- | --- | --- |
| Notes on Data | g Model | Q Model | AIPTW Estimate with 0.95 CI |
| 0.4km Circles, Violent Crimes | Linear Regression | Logistic Regression | 12.762 ± 4.665 |
| 0.4km Circles, Violent Crimes | Random Forest | Random Forest | 6.692 ± 1.436 |
| 0.4km Circles, All Crimes | Linear Regression | Logistic Regression | 19.435 ± 15.514 |
| 0.4km Circles, All Crimes | Random Forest | Random Forest | 7.358 ± 1.837 |
| 0.4km Circles, All Crimes | MLP Regressor | Random Forest | 11.535 ± 1.954 |
| 1km Rectangles, All Crimes | Linear Regression | Logistic Regression | 101.166 ± 29.225 |
| 1km Rectangles, All Crimes | Random Forest | Random Forest | 33.572 ± 19.627 |

Table 2: Average Treatment Effect of New Grocery Stores on Crime

|  |  |  |  |
| --- | --- | --- | --- |
| Notes on Data | g Model | Q Model | AIPTW Estimate with 0.95 CI |
| 0.4km Circles, All Crimes | Linear Regression | Logistic Regression | 4.368 ± 6.290 |
| 0.4km Circles, All Crimes | Random Forest | Random Forest | 4.723 ± 5.886 |
| 0.4km Circles, Violent Crimes | Linear Regression | Logistic Regression | 3.173 ± 4.205 |
| 0.4km Circles, Violent Crimes | Random Forest | Random Forest | 2.609 ± 3.930 |

The models with the best test loss were generally Gradient Boosting or Random Forest regressors and classifiers. On a few occasions logistic regression did not improve cross entropy loss on held out data for g(x). All models improved the MSE loss for Q(a,x) and all models other than logistic regression improved the test cross entropy for g(x). The Gradient Boosting Classifier for the g(x) model often had extreme values of zero or one, failing to satisfy the overlap condition.

**Sensitivity Analysis and Feature Importance**

By far the most important feature in predicting both treatment and outcome was the historical crime rate of the region. I did think the historical number of crimes would be a significant covariate, especially considering that it is one of two covariates that is specific the exact region as opposed to the community area as a whole. I thought making the outcome change in the number of crimes would have moderated its significance, but it still appears to be far and away the leading predictor of both treatment and outcome. The other most important features for g(x) were (1) Number of Previous Crimes, (2) Year, (3) Teen Birth Rate (4) Fertility Rate and for Q(a,x) they were (1) Number of Previous Crimes, (2) Year (3) Preterm Births (4) Treatment.

In the Austen Plots below it is clear how significant previous crime rate is in predicting outcome and treatment. In some ways these sensitivity analyses suggest that since the majority of confounders have little impact, maybe any unobserved confounders would also have negligible effect. However, considering the other confounders do not relate to the specific region in consideration like previous crime rate does, maybe other unobserved confounders which did relate to the specific region would have non-negligible impacts.

|  |  |
| --- | --- |
| Random Forests Liquor Store Violent Crime Sensitivity Analysis | Random Forests 1km Rectangles Liquor Store Crime Sensitivity Analysis |

|  |  |
| --- | --- |
| Chart, line chart, scatter chart  Description automatically generated Linear/Logistic Regression Grocery Store Sensitivity Analysis | Linear/Logistic Regression Liquor Store Sensitivity Analysis |

**Conclusion**

While new grocery stores do not appear to affect crimes rates, there does appear to be a significant effect of new liquor establishments on crime and violent crime. However, there certainly is skepticism that can be raised about the first identification assumption requiring no unobserved confounding. The covariate data is by no means comprehensive, and the resolution of the covariate data to only community area level is troublesome. While the sensitivity analysis does somewhat strengthen the argument for the result, it is by no means conclusive.

**Sources:**

<https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr>

<https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu>

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2>

https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

<https://data.cityofchicago.org/Community-Economic-Development/Grocery-Stores-2013/53t8-wyrc>

<https://catalog.data.gov/dataset/crimes-2001-to-present>

Victor Veitch’s examples