How might winter seasons and shelter clienteles influence shelter capacity conditions?

An analysis of the association between shelter capacity conditions and winter seasons as well as shelter clienteles.

Weijia Song（1004043689）

19/12/2020

Table of Contents

Key Words

Propensity Scroe Matching, Logistic Regression, Shelter Occupancy Condition, Observational Study, Casual Inference

Abstract

Shelter's occupancy conditions may be affected by shelter clienteles and winter seasons. This study uses logistic regression model combined with propensity score matching method to predict the relationship between shelter occuapncy conditions and shelter clienteles as well as winter seasons. The result shows that during winter seasons, shelter capacity is more likely to be exceeded. This study is based on observational data.

Introduction

According to the Toronto Street Needs Assessment 2018, 8715 homeless people lived in the Toronto area. Most of the homeless people gathered in the Toronto-East York region, including downtown Toronto (2018). Moreover, the number of homeless people in Toronto had been increasing over the years. The count of homeless in 2018 was 1.8 times higher than that in 2013 (2018). The growing numbers of homeless population led me to consider homeless people's circumstances: whether the shelter system in Toronto can satisfy homeless people's needs?

Data shows that 6 percent of the homeless population were living outdoors, and 29 percent shifted between indoor shelters and outdoors. The current shelter system may not satisfy the increasing homeless population. Shelter occupancy condition becomes an issue of concern because the unsheltered homeless population might increase the risk of disease transmission and the cost of city expenditure. A report about California homeless points out that outdoor homeless people have a higher risk of being exposed to community disease; there are also cases of harassment by homeless people (2017). Therefore, the city should resettle unsheltered homeless people as soon as possible. Also, a study of the shelter clientele shows that about 10 percent of the homeless were youth, and about 49 percent of the homeless were women. This led us to consider whether the clienteles of a shelter are related to its occupancy conditions, as well. Winter seasons are undoubtedly the most challenging seasons for homeless. About 51 percent of the homeless population used winter respite service, which aims to help unsheltered homeless get through winter seasons (2013). Thus, whether or not seasons and shelter clienteles are associated with shelter occupancy conditions is crucial. Because it allows the policymakers to determine whether there should be more shelter service during the winter seasons, and whether there should be more shelters for specific clienteles. In this study, the hypothesis is that there is a strong association between shelter occupancy conditions and clienteles and winter seasons. If the hypothesis is established, the government should consider adding more shelter beds during winter seasons and increasing funds or budget regarding particular homeless clienteles.

This is an observational study, and the data set in this project is about daily shelter occupancy in Toronto, which was obtained from the Open Data Toronto Catalogue. There are four variables of interest, "Occupancy," which is the number of homeless in a shelter; "Capacity," which is the maximum space available in the shelter; "Occupancy date," which is the date of the data

observed; and “Sector,” which is the clienteles of the shelter. This study will use propensity score matching to analyze if a casual inference is existed. A more detailed illustration of the data and the variable will be presented in the data description part. The simulation of data and the models will be discussed in the model section. In the discussion part, I will discuss the results of the study, the weaknesses of the project, and future steps that can be taken.

Methodology

Data

The Daily Shelter Occupancy Data aims to collect data of all functional shelters in the Toronto Area. The data is obtained from Open Data Toronto Catalogue. This dataset can help policymakers making a more proper decision regarding shelter programs in the Toronto Area. This is an observational study with 39381 observations. All the data was collected every day at 4 am, from January to December, in 2019. The population was all functioning shelters in the Toronto area from January 1 to December 31, 2019. The information of Violence Against Women shelters was excluded from the dataset. This choice is reasonable because such shelters should be kept confidential to prevent victims from being located.

There are 13 variables in this dataset. Variables include basic information such as city, street, occupancy date and postcode information, and shelter information such as sector, capacity and occupancy. Here is an overview of the original dataset.

Table 1: The Overview of City Shelter Occupancy Data

_id	OCCUPANCY_DATE	ORGANIZATION_NAME	SHELTER_NAME
1	2019-01-01T00:00:00	COSTI Immigrant Services	COSTI Reception Centre
2	2019-01-01T00:00:00	COSTI Immigrant Services	COSTI Reception Centre
3	2019-01-01T00:00:00	COSTI Immigrant Services	COSTI Reception Centre
4	2019-01-01T00:00:00	COSTI Immigrant Services	COSTI Reception Centre
5	2019-01-01T00:00:00	Christie Ossington Neighbourhood Centre	Christie Ossington Men’s Hostel
6	2019-01-01T00:00:00	Christie Ossington Neighbourhood Centre	Christie Ossington Men’s Hostel

SHELTER_ADDRESS	SHELTER_CITY	SHELTER_PROVINCE	SHELTER_POSTAL_CODE
100 Lippincott Street	Toronto	ON	M5S 2P1
100 Lippincott Street	Toronto	ON	M5S 2P1
100 Lippincott Street	Toronto	ON	M5S 2P1
100 Lippincott Street	Toronto	ON	M5S 2P1
973 Lansdowne Avenue	Toronto	ON	M6H 3Z5
973 Lansdowne Avenue	Toronto	ON	M6H 3Z5

FACILITY_NAME	PROGRAM_NAME	SECTOR	OCCUPANCY	CAPACITY
COSTI Edward Hotel (Families)	COSTI Edward Hotel Refugee Family	Families	543	628
COSTI Edward Hotel (Singles)	COSTI Edward Hotel Refugee Singles	Co-ed	12	13
COSTI Radisson Hotel	COSTI Radisson Hotel Family Program.	Families	757	902
COSTI Reception Centre	COSTI Reception Ctr CITY Program	Co-ed	16	16
Christie Ossington Men's Hostel	Christie Ossington Men's Hostel	Men	81	82
Christie Ossington South Hostel	Christie Ossington Men's Hostel South	Men	32	32

For the interest of this study, there was a data cleaning process. I used "tidyverse" package to clean the data. The cleaned data, including three variables, was created from original dataset.

The first variable is a dependent variable called "if exceed." Based on the original dataset, "occupancy" and "capacity" would be the primary focus in this study because this paper would like to know whether the shelter's capacity is overloaded. Therefore, a new variable, called "if exceed," was generated, and the output will be "1" if "occupancy" is more than "capacity", "0" if "occupancy" is less or equal than "capacity."

"Winter or not" is the second variable, which is a independent variables. I used the variable "occupancy date" from the original data to define this new variable. Occupancy dates can be used to predict capacity conditions because seasons might influence the need for shelter. Toronto's winter is long and extremely cold; therefore, homeless people might be more likely to seek shelter during winter seasons. Hence, this paper would like to know whether or not winter seasons might influence shelter occupancy conditions. According to Toronto's Winter Respite Service, the cold and winter seasons range from November 15 to April 15 next year (2013). Therefore, if the occupancy date is in the range between November 15 to April 15, the observation under variable "winter or not" will be encoded as 1. The observation that falls outside the date range will be encoded as 0.

Another variable chosen is "sector," which represents the clienteles of the shelter. The shelters were grouped into five categories: women, men, co-ed, youth and family. The occupancy conditions of each sector might be different due to various reasons. Therefore, understanding the relationship between shelter clientele and shelter occupancy is helpful for other researches.

Below is an overview of the cleaned data.

Table 2: The Overview of the Cleaned Data

_id	if_exceed	winter_or_not	SECTOR
1	0	1	Families
2	0	1	Co-ed
3	0	1	Families
4	0	1	Co-ed
5	0	1	Men
6	0	1	Men

Model

This study uses logistic model because it is suitable for analyzing relationship between binary dependent variable and a grouped of independent variables.

Here is the general model of the logistic model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n x_n$$

(Equation 1)

- $x_1 \ldots x_n$ represent the independent variables.
- p represents the probability of the event of interest occuring
- β_0 is the y-intercept of the model
- β_n is the coefficient of variable n, representing the one-unit change in dependent variable after one-unit change in x_n

In this study, whether or not a shelter's capacity is exceeded (if exceed) is the dependent variable. Therefore, I use logistic model to estimate the effects of winter seasons and shelter clienteles on shelter occupancy conditions. So, three variables were selected to build the model: "if exceed", "winter or not" and "sector". A more detailed interpretation of the variables is explained in data section. When fitting the variables into the logistic model, the dependent variable is "if exceed", and two independent variables are "winter or not" and "sector."

Hence, the logistic regression model for this study will be:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \ * \ winter\ or\ not \ + \beta_2 \ * \ sector$$

(Equation2)

- p represents the probability of a shelter capacity being exceeded
- β_1 is coefficient of the variable “winter or not”,
- β_2 is the coefficient of the variable "sector.

Propensity Score Matching

Propensity Score Matching is used in this study for a more precise prediction. A propensity score means the probability of a certain unit being assigned to the treatment. Then a treatment group and control group will be generated. The propensity score matching ensures that at least one unit in the treatment group and one unit in the control group with similar propensity scores (2019). The purpose of the propensity score matching is to eliminate bias that happened in an observational study.

This project is based on an observational study. It is observational because a experiment is impossible to achieve. For example, the variable “winter or not” is impossible to simulate in an experiment because there is no method to change the Toronto’s season. Therefore, the study uses propensity score matching to get a more precise prediction.

This study will treat some shelters with winter seasons, and treat others without winter seasons, then we will see what happened to the occupancy conditions. Therefore, the treatment group will be variable “winter or not”; and “if exceed” will be the outcome of interest. The propensity score will be the probability of a shelter’s capacity being exceeded. For each shelter that was actually exceeded, there is a shelter was not exceeded, both with a similar propensity score. Then the study will match all the shelters through the matching process. Finally, we can check whether there is a casual inference between winter seasons and shelter occupancy conditions.

The results of models will be interpreted in the results section.

Results

The result of this study is that, during winter seasons, shelter capacities are more likely to be exceeded. This paper will illustrate results in the following parts.

Results of the Logistic Model

Table 4 shows the summary of the logistic model

	(1)
(Intercept)	-21.631
	(355.581)
winter_or_not	0.154 **
	(0.057)
SECTORFamilies	20.935
	(355.581)
SECTORMen	12.164
	(355.582)
SECTORWomen	-0.000
	(473.664)
SECTORYouth	-0.000
	(514.389)
nobs	39381
null.deviance	15588.017
df.null	39380.000
logLik	-3650.399
AIC	7312.797

BIC	7364.283
deviance	7300.797
df.residual	39375.000
nobs.1	39381.000

*** p < 0.001; ** p < 0.01; * p < 0.05.

The output shows the coefficients of β_0 (y-intercept), β_1 (coefficient of variable 'winter or not'), and β_2 (coefficient of variable 'sector'), and their corresponding p-value. The coefficient of variable "winter or not" is positive (0.154), implying that during winter seasons, the shelters' capacity are more likely to be exceeded. For the variable "sector", the families and men sectors have a positive coefficient, representing that shelters with clientelts of families and men are more likely to be overloaded; while the negative coefficeint shows that shelters with clienteles of women and youth has less than 0 oppotunity to be overloaded.

However, looking at the coefficients is not enough. The p-value representing the probability against the null hypothesis. In statistics, the null hypothesis proposes that there is no relationship between independent variables and dependent variables. In this study, the null hypothesis is that winter seasons and shelter clienteles are not related to shelter occupancy conditions. If the p-value is more than 0.05, the null hypothesis holds; if the p-value is less than 0.05, the null hypothesis is being against.

The significance of the variable "winter or not" is 0.0064, which is less than 0.05; therefore, there is a strong evidence to reject the null hypothesis. Thus, there is a significant relationship between the independent variable "winter or not" and the dependent variable "if exceed". However, the p-values of the variable "sector" are more than 0.05, which are very high. Thus, we cannot conclude that there is a significant relationship between the independent variable "sector" and the dependent variable "if exceed."

To better improve the accuracy of the model, a backward selection is needed. The process of backward selection removes the least significant variable each time and re-run the model. Since the variable "sector" is not significant, I removed this variable.

Therefore, the new logistic model becomes:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \ * \ winter\ or\ not$$

(Equation 3)

Table 5 shows the summary of new model

	(1)
(Intercept)	-3.005 ***
	(0.031)
winter_or_not	0.132 **
	(0.047)
nobs	39381
null.deviance	15588.017
df.null	39380.000
logLik	-7790.050
AIC	15584.100
BIC	15601.263
deviance	15580.100
df.residual	39379.000
nobs.1	39381.000

*** p < 0.001; ** p < 0.01; * p < 0.05.

The new model summary shows a high significance of the association between the independent variable “winter or not” and the dependent variable “in exceed.” This new model is more accurate than the former one. Hence, in this study, the model with one variable is more appropriate than the model with two variables.

In conclusion, there is a significant relationship between winter seasons and shelter occupancy conditions, and during winter seasons, shelters are more likely to be overloaded. While there is no association between shelter clienteles and shelter occupancy conditions.

Results of Propensity Score Matching

In the last section, the conclusion is that the shelters are more likely to be overloaded during the winter seasons. However, this study is based on an observational study, so there might be some bias in the data set. To get a more accurate result, I conducted the propensity score matching process. The matching process had been illustrated in the Model section. This section would be focused on the result of the propensity score matching.

There are 16207 observations in winter seasons. The treatment group is winter, so we need to match those variables, there will be 32414 observations in the matched data set.

Table 6 shows the table of propensity score matching

	(1)
(Intercept)	-0.002
	(0.003)
winter_or_not	0.004 *
	(0.002)
SECTORFamilies	0.354 ***
	(0.004)
SECTORMen	0.000
	(0.003)
SECTORWomen	0.000
	(0.003)
SECTORYouth	0.000
	(0.004)
N	32414

R2	0.319
logLik	9184.563
AIC	-18355.127

*** p < 0.001; ** p < 0.01; * p < 0.05.

The result shows a significant relationship between winter seasons and shelter occupancy conditions, saying that shelters during the winter seasons are more likely to exceed capacities.

In conclusion, this study estimates that during the winter seasons, shelters are more likely to be overloaded. This is based on the analysis of the proportion of shelter capacity being exceeded modeled by logistic regression, which accounted for seasons and shelter clienteles, and a combination of propensity score matching.

Discussion

Summary

The study is interested in whether winter seasons and shelter clienteles affect shelter capacity conditions. The original data set is obtained from the Open Data Toronto Catalogue. Package "tidyverse" was used to clean the data and generate new variables. The cleaned data contains three variables: the shelter's capacity condition (if exceed), the season (winter or not), and the clienteles of the shelter(sector).

This study used a logistic regression model to estimate the association between independent variables "winter or not" and "sector", and dependent variable "if exceed." The model's output shows a strong relationship between winter seasons and shelter occupancy conditions. To get a more accurate result, this study conducted propensity score matching process. Treatment is "winter," and the interest of outcome is whether the shelter is overloaded.

Conclusion

The final logistic regression model shows that shelter capacities during the winter seasons are 0.132 times more likely to be exceeded (p-value = 0.00479). That is to say, during the winter

season, there is a 13% chance for a shelter to be overloaded. And the propensity score matching shows that shelter capacities during winter seasons are 0.0042 times more likely to be exceeded (p-value = 0.038), which means that a shelter has a 0.42% chance to be overloaded during the winter season.

The results are different; however, since the logistic regression model is based on observational data, while the propensity score matching process created a treatment group that can eliminate bias, propensity score matching is more accurate. Therefore, this study concludes that shelters' capacity in the Toronto area has a 0.42 percent chance to be exceeded during winter seasons.

Weakness & Next Steps

There are several weaknesses in this study.

Though there is a relationship between the winter seasons and shelter occupancy condition, the association is weak. There might be other variables that can also influence the shetler's occupancy conditions. For example, a report regarding shelters showed that use of beds could affect the shelter's capacity. For instance, in 2012, about 1381 shelter beds were unavailable because some homeless people occupied the bed for a long term (2013). So, the improper use of beds can be a factor leading to the overload of the shelter capacity.

Another weakness is that this study only analyzed the influence of winter seasons on shelter's occupancy conditions. However, other seasons might also influence the situation. For example, during summer seasons, where temperatures are too high, the homeless might more likely seek shelter. So further research can focus on other seasons.

In the future, the data collection can focus on more aspects of shelter, such as the use of bed and the facility condition of a shelter. More detailed data can help to get a more accurate result in the data analysis process. Also, an in-depth analysis regarding reasons can be conducted. The result is helpful for policymakers who hope to a more detailed shelter program plans in different seasons.

References

Homelessness in California. (2017). Retrived from http://www.auditor.ca.gov/reports/2017-112/summary.html

Toronto Street Needs Assessment 2018 Results Report. (2018). Retrieved from https://www.homelesshub.ca/sites/default/files/attachments/99be-2018-SNA-Results-Report.pdf

Stephanie. (2019, August 08). Propensity Score Matching: Definition & Overview. Retrieved December 18, 2020, from https://www.statisticshowto.com/propensity-score-matching/

Update on Shelter Occupancy and the Quality Assurance. Review of Shelter Access(2013) https://www.toronto.ca/legdocs/mmis/2014/ex/bgrd/backgroundfile-66029.pdf

http://mgimond.github.io/ES218/Week03a.html

Github Link

Code and data supporting this analysis is available at:

https://github.com/wei-jia99/STA304_Final-Project.git