My title*

My subtitle if needed

Kevn Shao

December 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/KevinShao1357/Toronto_Daily_Shelter_And_Overnight_Service_Modelling

1 Introduction

Canada provides shelters for people in need, such as refugees, the homeless, and people experiencing transitions in living spaces, and the country is continuing to build new shelters for such residents. On April 16, 2024, the Canadian government announced a funding of over 250 million dollars over two years to address homelessness and encampment all over the country. In 2022, Canada announced a target to eliminate homelessness by 2030. Canada Mortgage and Housing Corporation also allocated a total of 420 million dollars to account for homelessness ("Homelessness and Social Housing Fudnding" (n.d.)). This implicates that the Canadian government will continuously invest in improving current shelters and building new shelters around the country. In 2024, Canada's Minister of Immigration, Refugees, and Citizenship Marc Miller also announced that the country would accept 27,000 new refguees this year, and this trend is likely to continue in the next years (Miller (n.d.)). We can conclude that more shelters are likely to be built in Canada in the future, and the same trend is also valid for the GTA(Greater Toronto Area).

To find a way to maximize efficiency of the investment on shelters, I want to evaluate the effect of predicting then number of beds (which reflects the necessary size of shelters need be built, expanded, or renovated) by constructing a multivariate linear regression model based on variables such as specific types of shelters and their services, as well as occupancy rates and averaged number of occupants per day. We will limit our scope of prediction to the GTA(Greater Toronto Area), since it is the largest city in Canada, and is also the scope of our chosen dataset. We also set the city of where the shelter is located to Downtown Toronto, and also set the classification of shelters based on gender and age to mixed adult, which are all the ones of their category with the highest frequency. As mentioned in the previous paragraph, we also chose our response variable to be number of beds, a choice between shelters grouped by rooms or shelters grouped by beds. The reasoning behind these choices are explained in the 'Measurement' of the data section (Section Section 2).

After finishing the derivation of the final multivariate linear regression model, we can find that the model satisfies most linearity assumptions, but its residual versus fitted plot shows that the homoscedasticity (constant variance) assumption is violated. Also, from the qq plot, we can also see that the normality of errors assumption may be violated too, although this is unsure due to the relatively ideal p-value, r-squared, and residual standard error. Overall, we can see that there is the possibility that modelling the predictor variables, such as occupied beds, occupancy rate, types of shelters, and types of operations, to predict the number of beds needed per day at a location providing daily shelter and/or overnight service, which can reflect the most-likely number of beds needed for a location providing such services, which can determine the size and number of shelters which renovation is needed, and whether new shelters should be built at locations, which may precisely control the amount of investment by Canadian government needed on facilities providing daily shelter and/or overnight service. However, there are still some limitations and further discussions to this result, which is discussed in detail in Section 5.

The remainder of this paper is structured as follows: Section 2 details the data and measurement process of the chosen dataset; Section 3 covers the multivariate linear regression model used in this paper; Section 4 provides results obtain from the derived model; and Section 5 includes discussions, such as implications and future steps, of our model, and how our results can in turn be furthermore improved.

2 Data

2.1 Overview

The dataset "Daily Shelter & Overnight Service Occupancy & Capacity" (Toronto (2 December 2024)) was sourced from the dataset with same name found in Open Data Toronto. The paper's in-depth analysis and prelimary data analysis are all completed using the statistical programming language R (R Core Team (2023)).

The primary objective of this paper is to evaluate the validity of predicting the number of beds necessary for a shelter by constructing a multivariate linear regression model, as mentioned in the introduction section. The number of beds of a shelter can reflect the necessary size of a shelter, so the general investment amounts of renovating existing shelters and building new shelters can be estimated, which can maximize efficiency of Canadian government's investments in shelters.

Table 1 provides a sample (the first ten rows) of the cleaned dataset.

Table 1

service_type	program_area	classification	count	capacity	occupancy_rat
Shelter	Base Program - Refugee	Emergency	8	8	100.0
Warming Centre	Winter Programs	Emergency	19	23	82.6
Warming Centre	Winter Programs	Emergency	21	48	43.7
Shelter	Base Shelter and Overnight Services System	Transitional	93	93	100.0
Shelter	Base Shelter and Overnight Services System	Transitional	35	35	100.0
24-Hour Respite Site	Base Shelter and Overnight Services System	Emergency	77	78	98.7
24-Hour Respite Site	Winter Programs	Emergency	120	120	100.0
Shelter	Base Shelter and Overnight Services System	Emergency	62	62	100.0
Warming Centre	Winter Programs	Emergency	41	46	89.1
Shelter	Base Shelter and Overnight Services System	Transitional	5	5	100.0

Sample Data of Cleaned Data of Daily Shelter and Overnight Service

2.2 Measurement

Since the aim of this paper is only to evaluate the validity of predicting the number of beds necessary for a shelter by constructing a multivariate linear regression model, we want to simplify the dataset as much simple as possible.

We first consider the classification of shelters based on gender and age, which is divided into adult men, adult women, mixed adult, youth and family. We find that mixed adult shelters takes the majority with the most observations, so we conclude that mixed adult shelters are most representative for checking the validity of the multivariate linear regression model for predicting number of beds for a shelter, so we filter out so that only mixed adult shelters are in the cleaned dataset.

Downtown Toronto has the most shelters in all the cities (of GTA(Greater Toronto Area)) in this dataset, so we also conclude that Downtown Toronto is the most representative of all the provided cities, so we then filter out the cleaned data so it only contains shelters in Downtown Toronto.

In this dataset, the shelters are either grouped by rooms or by beds. Since our aim is to analyze the validity of maximizing efficiency of investments of shelters using a multivariate linear regression model, considering number of beds will bring more accurate results. This is because each room can vary in size, so it is not a precise measure of the demand of shelters, which is ultimately based on the number of people, while it is more accurate to assume that one bed can accommodate one person. Therefore, we then filter out the cleaned data so it only contains shelters in Downtown Toronto.

In the original dataset, each observation of shelters vary by date. However, because we are discussing about investments in shelters, which is a long-term process, and our response variable is the number of beds necessary in a shelter, we want to get an average of the number of beds to maximize the efficiency of shelter investments by the Canadian government. Therefore, we ignore the date column, and just aim to construct a multivariate linear regression model considering all days from January 1 to December 1, which is the version of the dataset we chose to use, which is updated until December 2, 2024.

Then, we determine our predictor variables, made up of three categorical variables, which are classification, program area, and service type, and two numerical variables, which are capacity and occupancy rate. The rationale here is basically just ignoring the columns which only provide basic information, and then save all the remaining categorical variables (leaving the different types of shelters variables that are prepared to be involved in the multivariate linear regression model, after fixing the city, as well as general age and gender of the shelters) and numerical variables that are logical to predict the number of beds. Finally, we rename the variables to make them easier to understand, and we get our final cleaned dataset.

2.3 Outcome variables (And Estimand)

The outcome variable, or response variable, for this multivariate linear regression model, corresponds to the variable 'count' in the cleaned data set, which in context of the dataset, is the number of occupied beds during this day of a specific location that provides daily shelter and/or overnight service. By averaging this variable's statistics over about a year, we may get a representative value of the number of beds necessary for a shelter with the given statistics of the predictor variables. Note that the variable 'count' should be less or equal to the current capacity of the corresponding shelter. Its summary statistics are presented in the following Table 2.

Table 2: Summary Statistics of Number of Beds Occupied During a Given Day

	Beds Occupieds in a Given Day
mean	50.51831
standard_deviation	34.18986
maximum	150.00000
minimum	1.00000

From the above summary statistics presented in Table 2, the mean of 'count' is around 50 to 51 beds occupied in a given day of a shelter, the standard deviation is around 34, and the maximum and minimum are 150 and 1 respectively. Here, we can see that 'count' has some outliers, and also has reasonable mean and standard deviation, and so 'count' is suitable for being the response variable of our multivariable linear regression model.

Since we are trying to constructing a multivariate linear regression model, our estimand is the response variable, which is the the number of occupied beds during a day (can be understood as the demand of beds) in a location that provides daily shelter or overnight service.

2.4 Predictor variables

As mentioned in the previous measurement section, our predictor varibles are made up of three categorical variables and two numerical variables, and are already chosen, because they are the only ones that are appropriate and logical to predict the number of occupied beds at a given day for a shelter. The following is a more specific description for each of the predictor variables.

2.4.1 Service Type

'Service Type' (dis a categorical predictor variable, representing the type of the overnight service being provided. Here, because we set the city to be Downtown Toronto, as well as age and gender to Mixed Adult, so only the following types exist:

A 'Shelter' is a facility that give people experiencing homelessness a temporary living space so that they can move into new housing. Shelters operate the entire year for 7 days a week and 24 hours a day (Toronto (2 December 2024)).

A 'Warming Centre' prevents people from experiencing extreme weather, giving people a shelter to stay in, and only opens with operation of 7 days a week and 24 hours a day during extreme weather alerts.

A '24-Hour Respite' gives people experiencing homelessness a resting place, also providing them with "meals and service referrals." It operates 7 days a week and 24 hours a day.

A 'Top Bunk Contingency Space' is not described in the dataset, but is also another type of shelter.

The following Table 3 describes the number of each service type in the cleaned data. Note that 'Respite' refers to 24-Hour Respite, and 'Cont_Space' refers to Top Bunk Contigency Space.

Table 3: Count for each service type

	Beds Occupieds in a Given Day
Shelter	4044
Warming_Centre	373
Respite	1979
Cont_Space	349

The above Table 3 clearly represents that the four service types have very different frequencies. Shelters take the most frequency, of around 4000 observations, while 24-Hour Respites take up of around 2000 observations, about half of that of shelters, and other two have the least frequency, each of around 400 observations.

2.4.2 Classifcation

The variable 'classification' corresponds to the original dataset's variable 'PROGRAM_MODEL', and are directed to either 'Emergency' or 'Transitional'. Basically, a location classified as 'Emergency' means that any people experiencing homelessness can come in without a referral, and vice versa.

Table 4 below is a table representing the counts of each of the two classifications of all the observation in the cleaned dataset.

Table 4: Count for each classification

	Counts For Each Classification
Emergency	5397
Transitional	1348

From Table 4, we can see that there are much more emergency locations (around 5400) than transitional locations (around 1400). There are around three times more emergency locations than transitional locations.

2.4.3 Program Area

'Program Area' (denoted as 'program_area' in the cleaned dataset) "indicates whether the program is part of the base shelter and overnight services system, or is part of a temporary response program". For this cleaned dataset with the set limitations, it includes the following types.

A 'Base Program - Refugee' is a program that serves refugees and other similar groups of people, and also operates the whole year.

A 'Winter Program' is a program based on the additional spaces under winter service plans. This may also add additional spaces to existing programs. In the dataset description, it is denoted as 'Winter Response'.

A 'Base Shelter and Overnight Services System' are regular programs that are set to operate the whole year.

A 'Temporary Refugee Response' is similar to that of a 'Base Program - Refugee', but instead "create spaces in the overnight services systems" (Toronto (2 December 2024)).

Table 5 below represents the counts of observations that have each program area. Note that 'Refugee' refers to 'Base Program - Refugee', 'Base_Shelter' refers to 'Base Shelter and Overnight Services System', and 'Temp_Refugee' refers to 'Temporary Refugee Response'

Table 5: Count for each program area

	Counts For Each Classification
Refugee	1276
Base_Shelter	4802
Temp_Refugee	147

Table 5: Count for each program area

	Counts For Each Classification
Winter_Programs	520

From Table 5, we observe that locations with the program area ‘Base Shelter and Overnight Services System’ has most observations (around 4800). Locations with the program area ‘Base Program - Refugee’ has 1272 observations, around one-fourth of the previous one. The remaining two locations has much less observations.

2.4.4 Capacity

The ‘Capacity’ (denoted ‘capacity’ in the cleaned dataset) is the maximum capacity of a location. The following Table 6 gives the mean, standard deviation, maximum, and minimum of ‘Capacity’.

Table 6: Summary Statistics of Variable Capacity

	The Capacity of a Location
mean	50.96738
standard_deviation	34.22653
maximum	150.00000
minimum	1.00000

From Table 6, the mean of capacity of locations is around 51 people, with an acceptable standard deviation of around 34 people. The maximum and minimum of capacity are 150 and 1 people respectively, and these statistics are all acceptable for constructing a multivariable linear regression model.

2.4.5 Occupancy Rate

The occupancy rate here (denoted as ‘occupancy_rate’ in the cleaned dataset) is basically the occupied number of beds divided by the capacity in number of beds. The occupancy rate is measured in percentage, from 0 to 100 percent.

The following Table 7 represents the summary statistics of occupancy rate.

Table 7: Summary Statistics of Occupancy Rate

	Occupancy Rate in a Given Day
mean	98.78766
standard_deviation	5.42981
maximum	100.00000
minimum	16.67000

From Table 7, we can conclude that occupancy rate of locations are all pretty much near 100, with a relatively low standard deviation of 5 percentage points, so most locations are relatively full, but we can still consider it as a predictor variable, but needs careful consideration when interpreting results.

3 Model

3.1 Model Set-Up (Including Diagnostics and Checking)

Now we start to create the multivariate linear regression model, with the set predictor and response variables. Specifically, the predictor variables in this model are made up of three categorical variables and two numerical variables. The three categorical predictor variables are service type, program area, and classification. The two numerical predictor variables are capacity and occupancy rate. The response variable is count, which is the number of occupied beds during a day for a specific location. The detailed descriptions of the predictor and response variables can be found in the 'Outcome variables (And Estimand)' and 'Predictor Variables' sections in Section 2.

We first create a basic multivariate linear regression model with the set predictor and response variables. However, we notice that by official descriptions by Open Data Toronto (Toronto (2 December 2024)) that occupancy rate is equal to count divided by capacity. Therefore, there is likely a strong relationship between the two predictor variables occupancy rate and capacity, so we add an interaction variable between occupancy rate and capacity, and construct another multivariate linear regression model.

We now have to test the validity of the two models, evaluating which one is better. To do that, we create Table 8 below, presenting a comparison between the p-value, r-squared, and standard error of the two models, with 'Original_Model' representing the first one and 'Interaction_Model' being the one with the added interaction terms.

Table 8: Comparsion of Features of Original and Interacted Model

	Original Model	Interaction Model
P-Value	0.000000	0.0000000
R-Squared	0.999042	1.0000000
Residual Standard Error	1.058960	0.0007439

From the above Table 8, both models have a p-value that is way less than the significance level of 0.05, and both have an r-squared that is extremely close to 1, meaning that a very high proportion of variance of the response variable can be explained by the predictor variables. This implicates that both multivariate linear regression models are highly valid models. However, the residual standard error of the second model with the interaction term is less than that of the first model without the interaction term, so we choose the second model with the interaction term.

Now, the final task is to check if the new model with interaction terms satisfy all linearity assumptions. To finish that, we will first graph the residual versus fitted plot, which is the Figure 1 below.

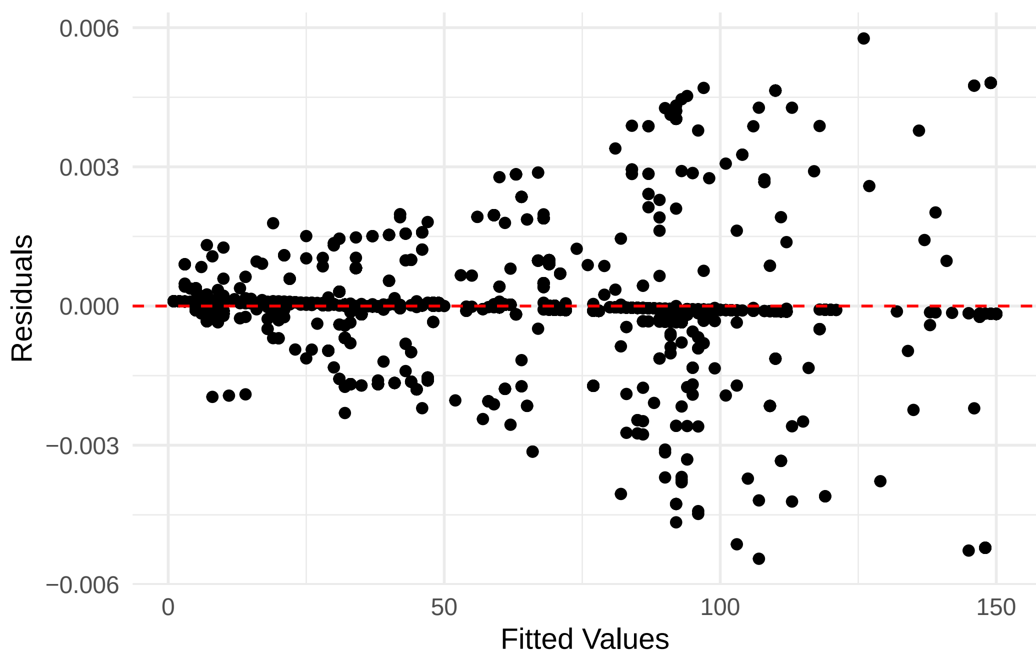

Figure 1

Here, according to the residuals versus fitted plot presented in Figure 1, the linearity of the relationship assumption is satisfied because there is no systematic pattern, since points in this plot are scattered all over the place. Homoscedasticity, is violated, since we see a conical shape, reflecting a variable variance. Therefore, we log the response variable and then graph the qq plot again, as the below Figure 2.

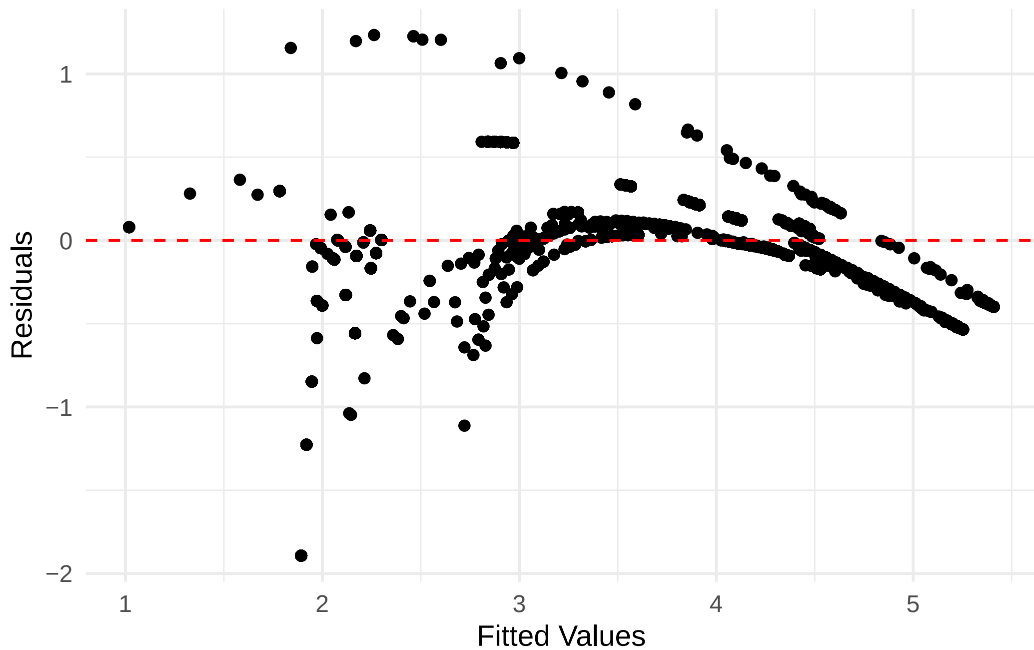

Figure 2

Figure 2 shows that now, the residual versus fitted plot shows that it both violates the linearity relationship assumption, since there is a clear pattern of the datapoints, and the homoscedasticity assumption, since the variance still varies, as seen in the graph. Therefore, the new model with logged response variable is even worse than the one that only added the interaction term, so we return the model that added the interaction terms. Now, we will continue to analyze the linearity assumptions by graphing the qq plot, as displayed in Figure 3 below.

Now, as shown in the qq plot displayed in Figure 3, there are many elements built arount y = 0, but from the overall perspective, it still has relatively no pattern. Therefore, we can still say that in a relative basis, we can still say that this model adheres to normal errors assumptions, but we must also take consideration that it does not follow the red line. However, also taking into the fact that the p-value, r-squared, and residual standard error, so we can still make the conclusion that we accept this multivariate linear regression model.

4 Results

Our results are summarized in

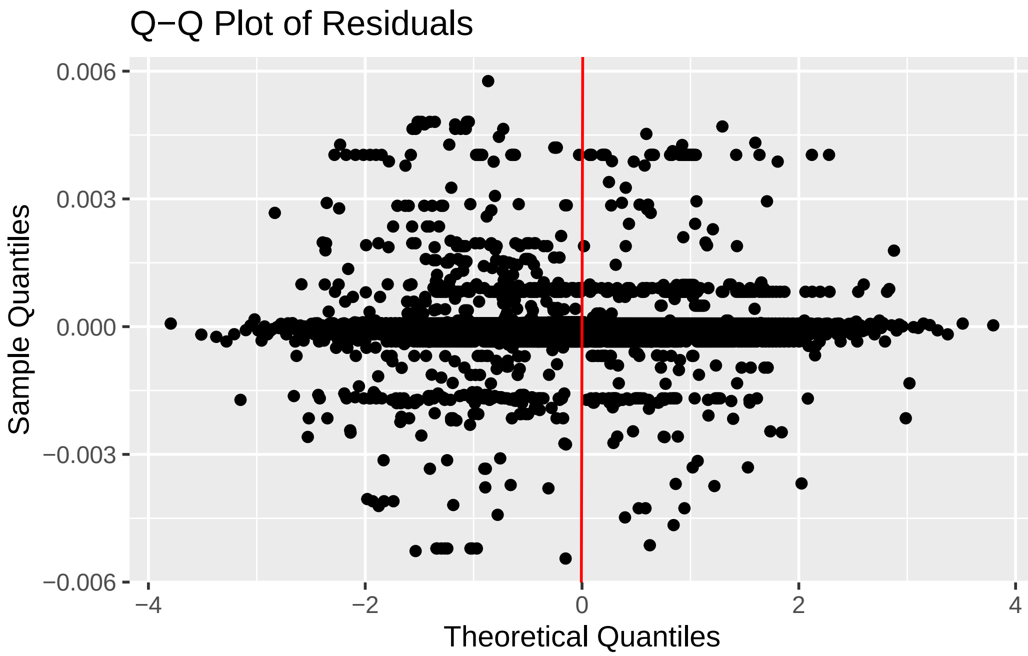

Figure 3

5 Discussion

Although our results of the constructed multivariate linear regression model do give a possibility that using statistical modelling, the number of beds needed per day at a location providing daily shelter and/or overnight service, which can help maximize efficiency of the investments on such facilities. However, there are still a few limitations that may lead in inaccuracies of our model.

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

References

"Homelessness and Social Housing Fudnding." n.d. Government of Canada. https://housing-infrastructure.canada.ca/pd-dp/parl/2024/05/huma/huma-d-eng.html.

Miller, Marc. n.d. "Canada Honours and Shows Solidarity with Refugees Worldwide." Government of Canada. https://www.canada.ca/en/immigration-refugees-citizenship/news/2024/06/canada-honours-and-shows-solidarity-with-refugees-worldwide.html.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto, Open Data. 2 December 2024. "Daily Shelter & Overnight Service Occupancy & Capacity." https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/.