Predictions of Occupied Beds at Faculties Providing Daily Shelter And/Or Overnight Services Using Multivariate Linear Regression Implies That Modelling This Situation Using Other Models Would Be Better*

Kevn Shao

December 14, 2024

Recently, the Canadian government expresses interest in continuous investment in government faculties providing daily shelter and/or overnight services. With the aim to evaluate the validity of predicting the occupied beds at a related faculties at a day, which can reflect the demand of such services at such locations, a multivariate linear regression model is constructed, with the city to Downtown Toronto, and the gender and age set to Mixed Adults. Five predictor variables are used, including two numerical variables, occupancy rate and capacity, and three categorical variables, including classification, service types, and program types. The final results showed likely violations of linearity assumptions such as homoscedasticity and normality of errors, and suggests that we may need to consider either proceed deeper in the linear regression model by adding nonlinear terms for predictor variables, identifying and deleting influential points, and/or consider the weighted-least squares (WLS) method, or consider switching to other more-flexible model types.

Table of contents

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

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 refugees 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, including the use of a box-cox transformation during the process, we can find that the model violates many linearity assumptions, such as linear relationship, normality of errors, and homoscedasticity (constant variance). The new model's residuals versus fitted and qq plots implicate that influential points should be identified and eliminated, and nonlinear transformations should be put on predictor variables. It can then be concluded that further steps can be made to the current multivariate linear regression model, but also more-flexible alternative model types, such as generalized additive models (GAMs), can be implemented to fulfill the same purpose with more precise final results. More detailed limitations and discussions to this result 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; Section 5 includes discussions, such as implications and future steps, of our model, and how our results can in turn be furthermore improved; Section A is about analysis of observational data and surveys, and is the appendix of this paper.

2 Data

2.1 Overview

The dataset "Daily Shelter & Overnight Service Occupancy & Capacity" (Support Services (2 December 2024)) was sourced from the dataset with same name found in Open Data Toronto. Note that the version of the dataset used for this paper is updated to December 13, 2024. The paper's in-depth analysis and preliminary data analysis are all completed using the statistical programming language R (R Core Team (2023)). The book 'Telling Stories with Data' (Alexander (2023)) was also used to complete this paper.

Packages used to complete the paper and/or the scripts include ggplot2 (Wickham (2016)), tidyverse (Wickham et al. (2019)), opendatatoronto (Gelfand (2022)), janitor (Firke (2023)), dplyr (Wickham et al. (2023)), readr (Wickham, Hester, and Bryan (2024)), knitr (Xie (2014)), kableExtra (Zhu (2024)), arrow (Richardson et al. (2024)), testthat (Wickham (2011)), car (Fox and Weisberg (2019)), here (Müller (2020)), and MASS (Venables and Ripley (2002)).

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_rate
Shelter	Base Program - Refugee	Emergency	8	8	100.00
Warming Centre	Winter Programs	Emergency	19	23	82.61
Warming Centre	Winter Programs	Emergency	21	48	43.75
Shelter	Base Shelter and Overnight Services System	Transitional	93	93	100.00
Shelter	Base Shelter and Overnight Services System	Transitional	35	35	100.00
24-Hour Respite Site	Base Shelter and Overnight Services System	Emergency	77	78	98.72
24-Hour Respite Site	Winter Programs	Emergency	120	120	100.00
Shelter	Base Shelter and Overnight Services System	Emergency	62	62	100.00
Warming Centre	Winter Programs	Emergency	41	46	89.13
Shelter	Base Shelter and Overnight Services System	Transitional	5	5	100.00

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	50.39825
standard_deviation	34.13035
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 (Support Services (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	4164

Table 3: Count for each service type

	Beds Occupieds in a Given Day
Warming_Centre	401
Respite	2039
Cont_Space	359

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	5575
Transitional	1388

From Table 4, we can see that there are much more emergency locations (around 5600) 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" (Support Services (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	1316
Base_Shelter	4942
Temp_Refugee	157
Winter_Programs	548

From Table 5, we observe that locations with the program area ‘Base Shelter and Overnight Services System’ has most observations (around 5000). Locations with the program area ‘Base Program - Refugee’ has 1316 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	50.89229
standard_deviation	34.13498
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	98.630649
standard_deviation	6.029715

Table 7: Summary Statistics of Occupancy Rate

	Occupancy Rate in a Given Day
maximum	100.000000
minimum	2.040000

From Table 7, we can conclude that occupancy rate of locations are all pretty much near 100, with a relatively low standard deviation of 6 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 created a basic multivariate linear regression model with the set predictor and response variables. However, we notice that by official descriptions by Open Data Toronto (Support Services (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	0.0000000	0.0000000
R-Squared	0.9988529	1.0000000
Residual Standard Error	1.1566916	0.0007451

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.

Moreover, also note that since the occupancy rate of a shelter is equal to the number of people living in that shelter ("count") divided by the capacity of that shelter. Moreover, when the Toronto city government builds a shelter with a higher capacity, this typically means that they forecast the demand of shelters at this location, so the number of people living in that shelter on average is also typically higher. Therefore, the two variables should have a logical relationship, and so adding an interaction term between the two is also logical.

Now, the next 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.

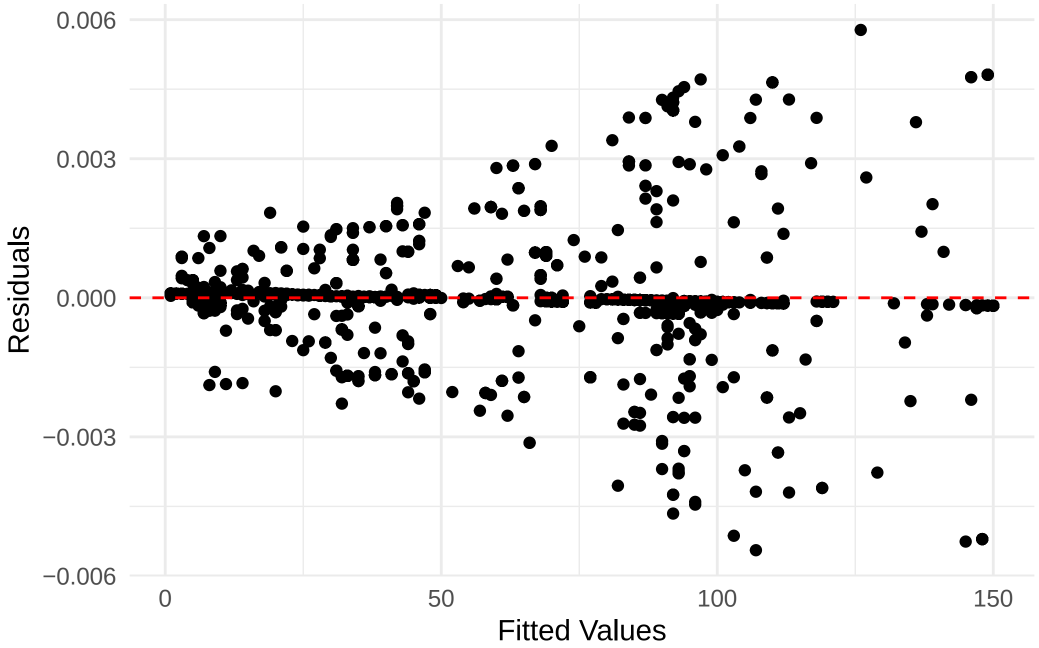

Figure 1

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, or constant variance, 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 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 data points, 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.

Our next step is to determine the specific transformation that should be done to the response variable using the box-cox transformation method. Before we do that, we want to check one more time whether to use to use the pre-logged model or the original model. We do that by comparing the qq plots between the two models. Figure 3 is the qq plot for the pre-logged model, and Figure 4 is the qq plot for the original model.

The two qq plots both have bad performance, both clearly violating normality of errors assumptions, and with the previous analysis of the residuals standard errors and p-values, we still eliminate the original model and choose the model with interaction terms but before we log the response variable. Now we proceed with the Box-Cox transformation, update the dataset and model, and then create the residuals versus fitted (Figure 5) and qq plot (Figure 6) for this new model.

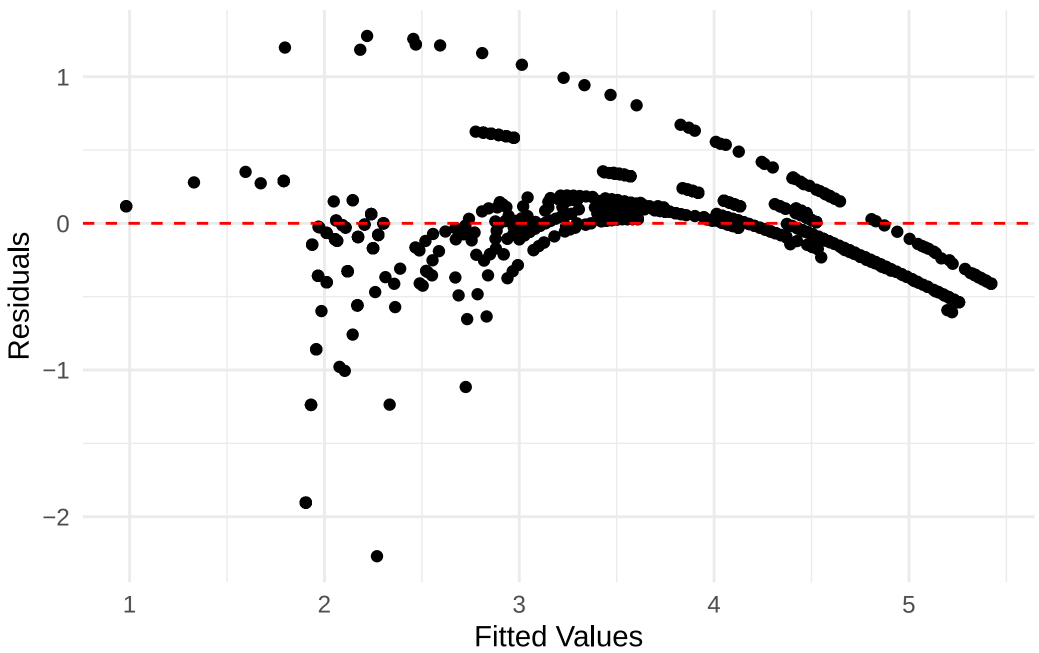

Figure 2

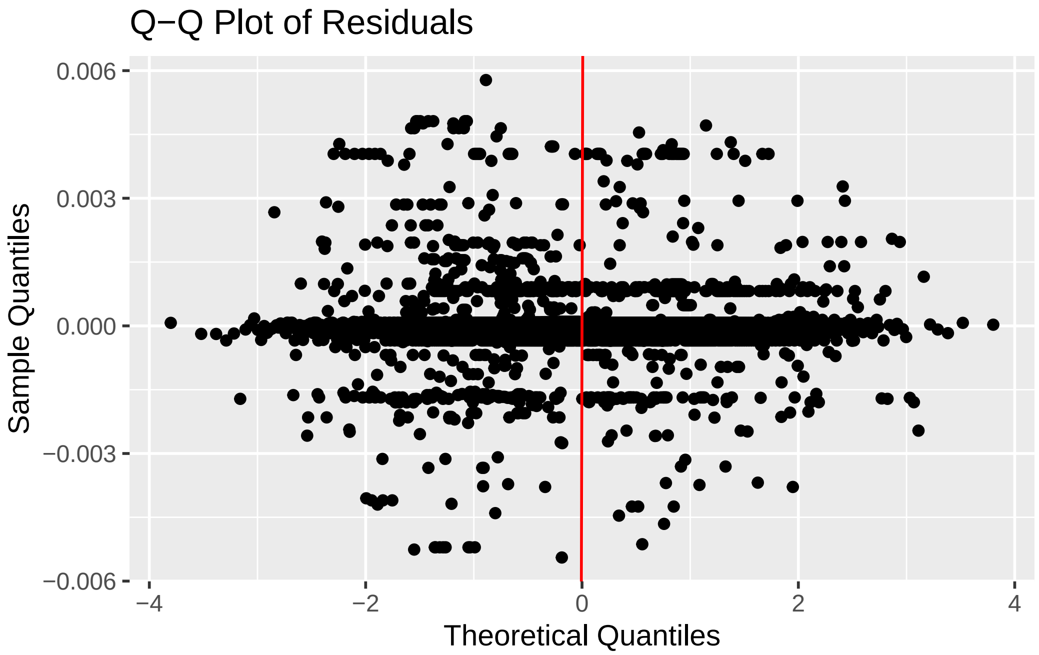

Figure 3

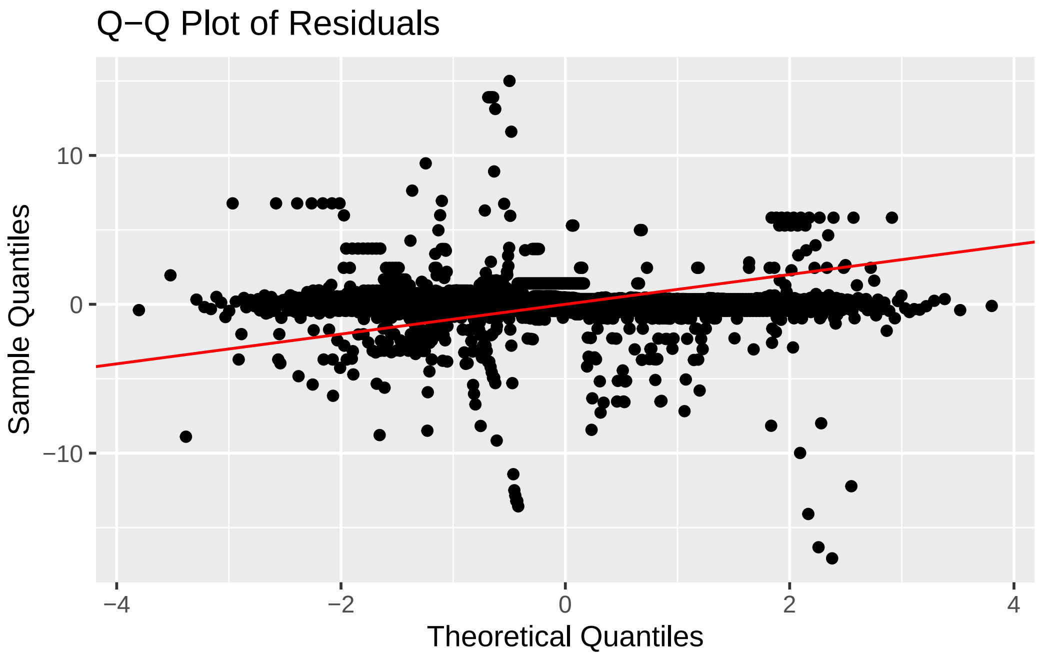

Figure 4

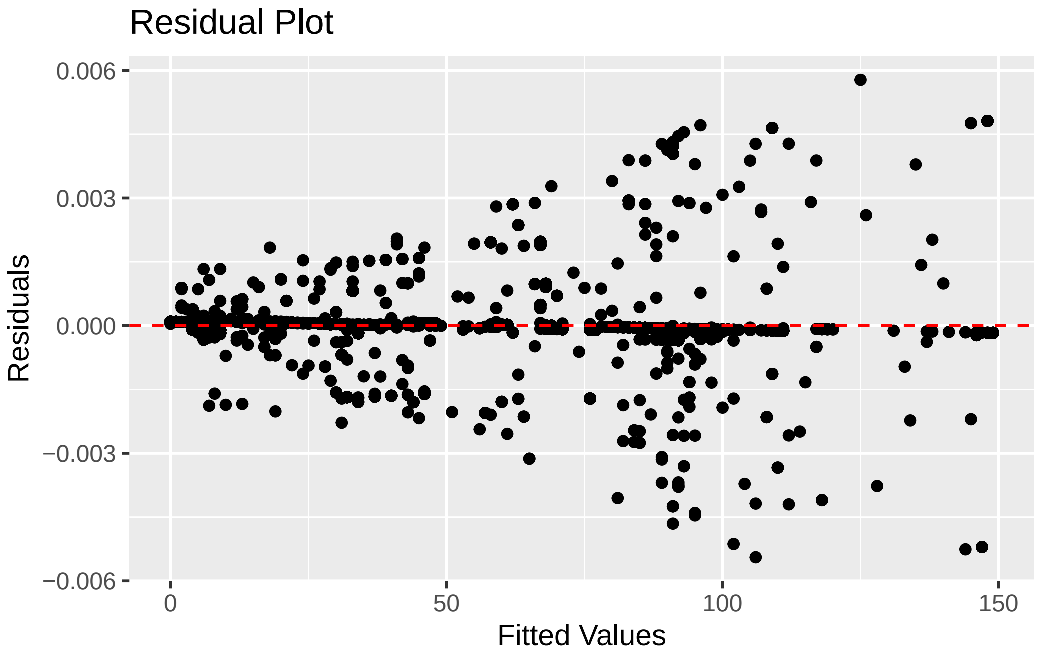

Figure 5

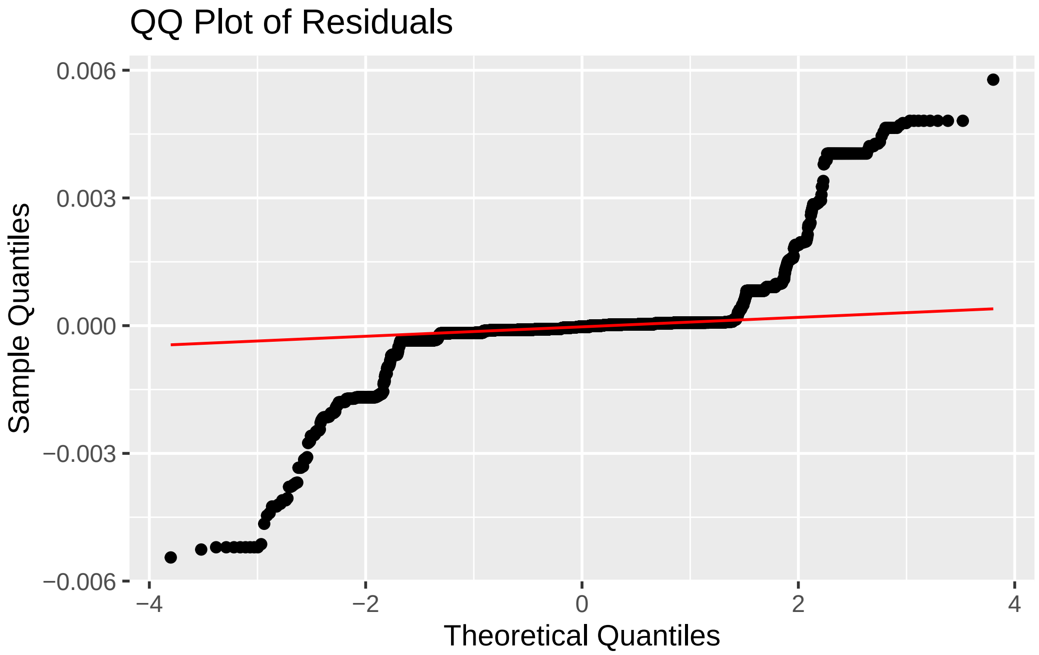

Figure 6

The qq plot of the new model after completing the box-cox transformation (Figure 6) has improved compared to the previous one, with at least most data points following the trend marked by the red line, but heavy tails can still be seen on both ends, the one with the negative theoretical quantiles and the one with the positive theoretical quantiles. This suggests possible outliers on both ends when we only look at the qq plot itself.

Now we observe the residuals versus fitted plot for the new model after the box-cox transformation (Figure 5). This plot data points, with the x-axis representing fitted values and the y-axis representing residuals, displays a few semi-circle patterns, which suggests even more severe violations of homoscedasticity, or constant variance. Therefore, I should consider adding polynomial terms at this point, but the method for deciding the specific polynomial terms is over the scope of my knowledge, so I just stop here and save the current model as our temporary final model. Also note that from this residuals versus fitted plot, we see that the new model also violates the linearity relationship assumption, but since it at least provides a possibility to add polynomial terms to our predictor variables, I am on the right track, so I still save this new model.

Note that this 'Model Setup' is only a brief summary of what I did in the model section, and the '05-model_data' file in the scripts section and the file in the models sections contains more details of the model derivation process. Please read the 'README' file for more detailed information of the structure of this Github repository.

4 Results

Our results are summarized in Table 9.

Table 9: Coefficients and Intercept for the Linear Regression Model

	Coefficient
(Intercept)	-1.0007658
service_typeShelter	-0.0000763
service_typeTop Bunk Contingency Space	0.0000213
service_typeWarming Centre	0.0001159
program_areaBase Shelter and Overnight Services System	-0.0001007
program_areaTemporary Refugee Response	-0.0001041
program_areaWinter Programs	-0.0002357
classificationTransitional	0.0002749
capacity	0.0000255
occupancy_rate	0.0000075
capacity:occupancy_rate	0.0099998

To this point, we can conclude that using multivariate linear regression models to predict the demand of shelters at a location, aimed to precisely control the amount of investment on shelters by Canadian government is a complicated approach, since it includes necessary further steps such as adding polynomial terms, checking influential points, and/or using weighted least squares (WLS) methods. Therefore, we must admit that switching to more flexible models such as generative additive models (GAM) may be a better choice to proceed furthermore.

5 Discussion

Although the derived results of the constructed multivariate linear regression model violates many linearity assumptions, including homoscedasticity, linear relationship, and normality of errors, and I stop since it is out of scope of my current knowledge, I can still discuss further steps and discussions of previous results.

At this point, please also note that it is also a challenge to predict the average demand of specific shelters (the response variable), since after we proceed with the box-cox transformations, we should not interpret the results by back-transforming again, so it would be hard for predicting the specific values of our response variable.

5.1 Next Steps

5.1.1 Polynomial Terms

As mentioned in the previous sections, from the residuals versus fitted plot of the newest model (Figure 5), to violate less of linearity assumptions such as linear relationship and homoscedasticity, what I should first do is to add nonlinear relationships to the predictor variables. Such approaches may include adding polynomial terms, such as capacity + (capacity)^2, logging our predictors, such as transforming capacity to log(capacity), and transforming our predictors using exponential terms, such as from capacity to exp(capacity).

In my newest linear regression model, transforming my predictors to nonlinear ones is logical. This is mainly due to the corresponding residuals versus fitted plot, which shows a clear pattern between the residuals versus fitted values instead of scattered, pattern-less data points.

5.1.2 Checking and Removing Influential Points

From the qq plot of the newest linear regression model after the box-cox transformation (Figure 6), we see a heavy tail on the end with positive theoretical quantiles. This may suggest possible outliers, which are data points that have extreme values on the response variable, and/or leverage points, which are data points that have extreme values on the predictor variables. Therefore, the best approach to consider that takes in consideration of both types of points is to determine and remove the influential points, which takes in the most extreme values of both by determining which observations have unusual large impacts on the fitted regression model itself. This determination and elimination process can be proceeded by either calculating DFBETA values or Cook's distance, both methods using reasonable set thresholds.

In this model derivation process, the step of identifying and eliminating influential points was not proceeded, since it should be done after adding nonlinear terms on predictor variables (and this step exceeded my scope of knowledge)

5.1.3 Next Steps

Other steps that I could complete in the future include the following.

I can first consider weighted less squares (WLS) after completing the procedures described above to account for addressing heteroscedasticity. The new model may not entirely eliminate violations of homoscedasticity, or constant variance, since this approach is best to apply when variance increases or decreases systematically with fitted values, but it at least diminishes this violation.

After this, the final step for constructing the multivariate linear regression model would be using the AIC(Akaike Information Criterion)/BIC(Bayesian Information Criterion) to determine the best model of all numbers of predictors, ranging up to the current most-complex model. Note that our model is predictive, since we want to predict the demand of shelters ("count"), and there is a possibility of a larger scope in the future, so we will choose AIC in this case, but this is also part of further steps.

However, we must also consider to use an alternative model instead of multivariate linear regression model to address the relationship between the predictor and response variables. For example, higher-end models such as generalized additive models (GAMs), as well as machine-learning methods such as decision trees and random forests are all more-complicated ones that provide a more flexible approach to address the relationship.

5.2 Further Steps After Completing The Above Possible Procedures

5.2.1 Enlargening The Scope

Another way to possibility improve the constructed multivariate linear regression model is to enlarge the scope. In this linear regression model, the chosen city was set to be Downtown Toronto, which is only one part of Greater Toronto Area (GTA), and so is an even smaller part of Canada, which is the world's second largest country. To make the model more representative of Canada as a whole, and to make it more useful for predicting the needed number of beds for a location per day, we can enlargen the scope to the entire GTA, or possibly the entire province of Ontario, which may make results of the model more effective for predicting the amount of investment needed in these facilities, or to make good use of current investment budgets on such facilities providing daily shelter and overnight services.

Moreover, in this linear regression model, the gender and age that a specific location provides service to is also set to 'Mixed Adults'. To make the linear regression model even more effective and accurate, one can also enlarge this scope to contain all genders and ages by directly making it a categorical variable. One can also consider making independent linear regression models, or even make use of the more-complicated Bayesian model, for each of the following specific types of genders/ages of groups of people provided by locations, further increasing precision.

5.2.2 Making The Date a Predictor Variable

In this derived multivariate linear regression model, date is not a predictor variable, and the model is based on a dataset that only considers data from January 1 to December 13, 2024. However, one could make it a continuous predictor variable to increase precision and/or accuracy. For example, if the Canadian government decide to invest in such facilities that provide daily shelter and/or overnight service on November 16, 2026, one could make use of the newly-derived function, and put the date as the date predictor variable, and get a rough prediction of how many beds would be occupied at that specific date and location for a specific facility, aiding them to effectively use investments to improve the entire shelter system.

5.3 Further Discussions

Using the current multivariate linear regression model, one can get an estimate of occupied beds at a random day at a specific location of such a facility. By simply considering the amount of facilities providing daily shelter and overnight service, one can get an estimate of the total occupied beds at a random at a specific location of all such facilities in Downtown Toronto. After estimating the size needed for each bed, one can get the total area of facilities needed in Downtown Toronto, which gives the Canadian government a sense of which facilities needs to be renovated, and whether there are any new facilities needed to be built at locations due to shortage of space and/or old facilities that can be stopped using due to excessive space. Actually, if found useful after a time period, this approach can be used to the entire GTA(Greater Toronto Area), and can further expand to the entire Ontario, and later, the entire Canada in the future if proved to be efficient enough.

A Appendix A: Analysis of Observational Data and Surveys

Here is just some analysis of my constructed multivariate linear regression model, as well as some of my understandings of this provided dataset.

As I have been analyzing the cleaned data all the time, I am only being in touch with a subset of the Open Data Toronto that is about daily shelter and overnight service occupancy and capacity that I have chosen. Therefore, I may have stepped into the Simpson's Paradox, which occurs "when we estimate some relationship for subsets of our data, but a different relationship when we consider the entire dataset" (Alexander (2023)). The Simpson Paradox is an instance of the ecological fallacy, which basically occurs when we try to conclude some characteristics of an individual based on their group. This paradox is also certain to be possible to happen on my analysis, since I only chose the city Downtown Toronto, whereas the original dataset is about information of same facilities, but for the entire Greater Toronto Area (GTA). For example, if for the current data about the city Downtown Toronto, the residual plots show a clear conical shape, showing a violation of the linearity assumption homoscedasticity (constant variance), then I would naturally try to log the response variable to adhere more to this assumption. After this approach, I follow the other typical procedures to construct a final multivariate linear regression model. Then, at this moment, if I then generalize the model for all cities in Greater Toronto Area (GTA), an issue may occur if for certain other cities, no conical shape appears in other residual plots, then there may exist a possibility that for these cities, when I log the response variable, the homoscedasticity assumption is then unexpectingly violated. If the frequency of these cities are too high, the model I constructed for the city Downtown Toronto may not be valid as a whole.

A similar paradox Berkson's paradox may also occur similarly. Berkson's paradox occurs when "we estimate some relationship based on the dataset that we have, but because the dataset is so selected, the relationship is different in a more general dataset" (Alexander (2023)). Take the cleaned version of my chosen dataset again, the current multivariate linear regression model is only relatively valid, and is not valid to the highest extent, due to the possible violation of the linearity assumptions of homoscedasticity and/or normality of errors. However, this may be not be the case for the entire dataset which has the scope of the entire Greater Toronto Area (GTA), so the model between the same predictor and response variables may be more valid, less valid, entirely valid, or not valid at all, depending on the key characteristics of variables of other subsets based on the scope of other cities in GTA.

Let's also analyze some information about the data collection process of the same Open Data Toronto dataset named 'Daily Shelter & Overnight Service Occupancy & Capacity'. According to the official website page of the dataset on Open Data Toronto, the updated information about shelter and overnight service programs, including the program's operator, location, classification, occupancy, and capacity, which are all "administered by TSSS(Toronto Shelter and Support Services" (Support Services (2 December 2024)). However, we cannot make sure that the administration of TSSS is certainly free of mistakes, so some parts of the dataset may also be done with convenience sampling, which is a type of non-probability sampling when researchers collect data which are easiest to accest from their perspective. If this hypothesis holds true, then even if the entire dataset used, the model can be still not representative of all the related facilities in the Greater Toronto Area (GTA).

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman & Hall/CRC. https://tellingstorieswithdata.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

"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.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.