Visitors who: go to parks with children/pets, or with disability will visit parks with higher frequency

Jiayue Wu

3/02/2022

Abstract

It is well-known that parks provide preservation and conservation of not only wildlife but also plant life. Many studies have indicated that parks can increase the quality of daily life. To help designers and planners improve park usage, this study's purpose and research question investigate, first, what specific factors might affect the frequency of park users visiting parks and, second, how these factors might affect visitors' frequent usage of parks. With the data obtained using an online survey from our collaborators, 1315 responses were collected. Results from the fitted model show that visitors in a group with members aged less than 12 are more likely to go to parks with higher frequency. Disabled people also tend to visit parks more frequently. Furthermore, visitors who stay at parks to relax or to play also show a higher frequency compared to visitors who are at a park for different activities. Visitors who visit parks with pets also tend to do so more frequently. Additionally, parks in the southern part of Toronto might be visited more frequently compared to parks in the northern part of Toronto. Based on the results, the author will recommend that Toronto parks involve some convenient facilities for high-frequency visitors. For example, parks can include more relaxing facilities, such as benches, that provide spaces for a group of visitors and facilities for disabled people, which will allow easy access to buildings in parks, particularly washrooms.

1. Introduction

As the demand for urban parks increases, greater attention to planning and exploring more of the parks' use and the frequency of people going to parks is required. The research question of this report is to assess the significant factors that will affect the frequency of people going to parks, either more frequently (1) or less frequently (0). The potential independent variables I collected are gender; activities visitors do; whether visitors go to parks alone, with pets, or with others; age; condition of disability; and location of the parks. Further details about the variables and the data are listed below.

Data description: Data for this report were obtained using an online survey from our collaborators. The questionnaire included questions about frequency, duration people stayed at parks, and some basic information about the respondents, such as gender and age. In addition, respondents also indicated the activities in which they engaged during their visits to parks. Among the 37 park surveys provided by our collaborators, only five park surveys include all the questions/factors in which I am interested. The parks selected and the corresponding locations in Toronto are Albert Standing Park (North); Cashman Park (South); St. James Town West Park (South); Willowdale Park (North); and Ivan Forrest Garden (South). After filtering out some multi-selected variables, the final data frame I used in the next modelling part includes 1315 responses.

Description for the variables the author selected are the following:

1. Frequency: This is the response variable I set in this study and model. This variable describes the frequency of the respondents going to parks. The original data is constructed with four levels: "Yearly", "Monthly", "Weekly", "Daily", and we transformed it into a two-level variable with either more frequently (1) or less frequently (0) for better prediction.
2. Gender: This is a characteristic variable that describes the Gender condition of respondents with three levels: "only female", "only male", and "Both male and female".
3. Activity: These two-level variables describe the activities visitors do in parks. For example, sit in parks; relax in parks; play in parks; walk in parks; picnic at parks.
4. Who: This variable describes whether visitors go to parks: Alone; with Pets; or with Others.
5. Age: This variable is the age conditions for park visitors.
6. Disability: this is a factor variable that describes the disability conditions of visitors to parks.
7. Location: This factor variable describes whether visitors located in the Northern(N) or Southern(S) of Toronto.

Outline: As for the following parts of this report:

- The Method section (part 2) indicates the candidate models and methodologies the author tried fitting to the data and the brief reasons for how the final model was selected. Furthermore, the author also checked the model's diagnostics with the assumption part and goodness of fit part.
- The Results section (part 3) indicates the results from the model and the corresponding interpretation
- The Conclusion & Discussion section (part 4) presents a more general finding of this study and corresponding suggestions. Furthermore, this part also indicates the limitations/challenges when analyzing the research question and corresponding potential solutions.

2. Method

In the model selection part, the author conducted two kinds of statistical strategies/models: (1) Ordinal Logistic Regression; and (2) Logistic Regression.

Candidate Model I: Ordinal Logistic Regression

As previously mentioned, the original response variable in this study is a four-level ordinal variable called

Frequency, after researching from multiple papers, the ordered logistic regression model could be well fit for the ordinal variable. The mathematical expression is as follows:

$$logit(P(Y \leq j)) = log(\frac{P(Y \leq j)}{P(Y > j)}) = \beta_{j0} - \beta_1 x_1 - ... - \beta_p x_p$$

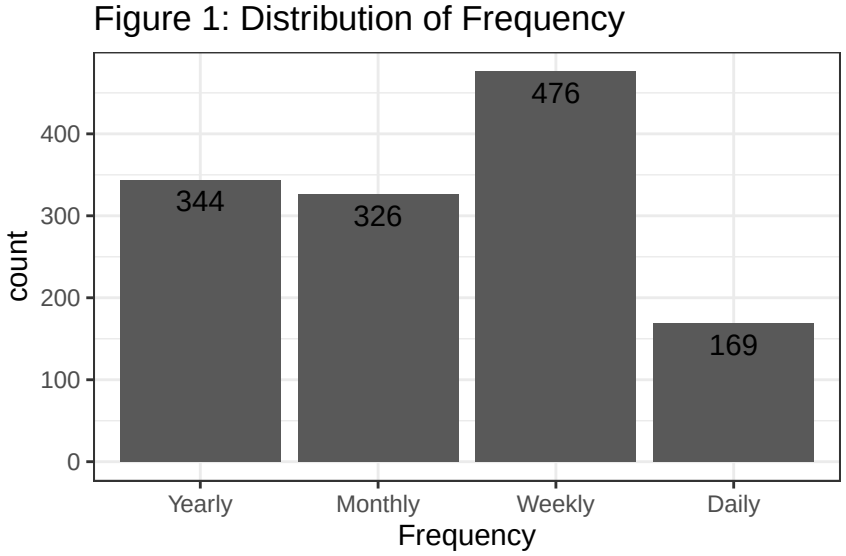
- Y is the ordinal outcome *Frequency* with 4 categories, since there are 4 levels for the variable *Frequency*.
- $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category j = 1,2,3,4. While P(Y>5) = 0 is undefined for the odds of being less than or equal to a particular category. $\frac{P(Y \leq j)}{P(Y>j)}$ is the odds of being less than or equal to a particular category.
- x_i is the indicators put in the model. β_i is the estimated coefficient for the independent variables.

However, when checking the diagnostic of the model, one of the important assumptions: *Proportional odds assumption*, i.e., the assumption of the relationship between each pair of outcome groups has to be the same, failed to satisfy. In other words, the Ordinal Logistic Regression method failed to fit the data selected for this study quite well.

Recall that the purpose of this study is to investigate what kind of visitors go to parks more frequently. The author can simply classify the frequency visitors go to parks into two levels, either more frequently or less frequently. Hence, the author tried the second method, Logistic Regression, as follows.

Candidate Model II: Logistic Regression

Firstly, the outcome variable of the Logistic Regression should be binary. Below is the distribution plot of the response variable *Frequency*. The author will classify the dependent variable into two-level by half and half by the number.



Here we can find that the count of $Yearly + Monthly = 670 \approx 645 = Weekly + Daily$. Hence, here I classify Weekly and Daily as *more Frequently*; Monthly and Yearly as *less Frequently*.

The Logistic Regression model is expressed as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

The notation and Mathematical expression for the Logistic Regression are the following:

- Y is the two-level outcome *Frequency* with higher frequency to parks (1) or less frequency to parks(0).

- p is the probability of the visitor goes to parks more frequently. $\frac{p}{1-p}$ is the odds ratio of going parks with higher frequency.
- x_i is the indicators will put in the model, i.e., Gender, Age, Activity visitors do in parks, relationship among visitors go to parks together, whether go to parks alone, disability conditions for visitors to parks, and so on. β_i is the estimated coefficient for the independent variables mentioned above.

Furthermore, despite constructing the full model by logistic regression, the author also constructed a model by conducting the AIC stepwise selection, where the AIC (Akaike Information Criteria) selected model has conducted a trade-off between the goodness of fit of the model and the simplicity of the model in case of the overfitting. More precisely, here, the author applied the both-direction AIC stepwise selection, adding predictors to the model that are statistically significantly related to the response variable while removing any predictors that no longer provided an improvement in model fit.

Model Selection: To compare the model behaviors quantitatively, the author constructed a stepwise elimination for each of the two candidate models based on Akaike information criterion(AIC) and a Bayesian information criterion(BIC). More precisely, AIC is a measure of the goodness of the models; BIC is a type of model selection among a class of parametric models with different numbers of parameters. Here the lower value of AIC or/and BIC is better. Furthermore, the likelihood ratio test is to check whether the full model offers a significant improvement in fitting the data.

After comparing the AIC and BIC and the p-value from likelihood ratio test, the author can roughly conclude that the stepwise selected model behaves better. Hence, the stepwise selected model is selected as the final model.

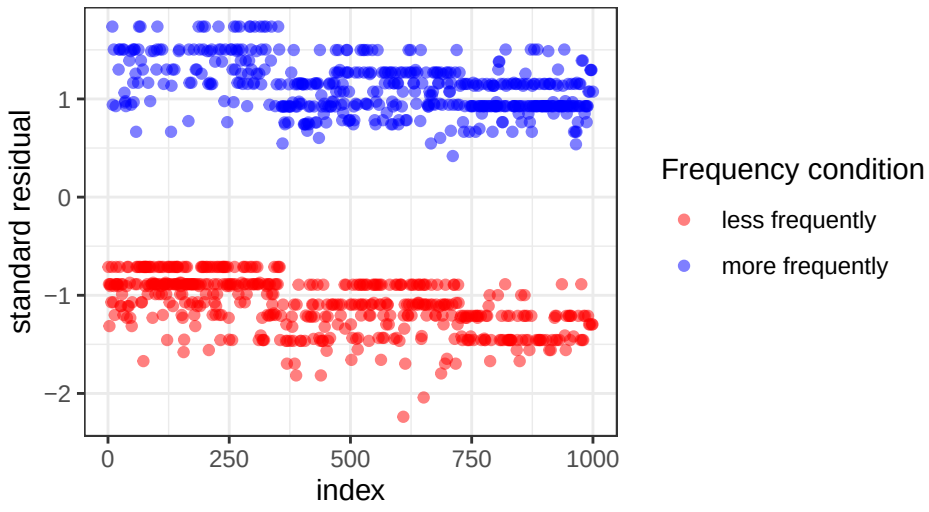
Assumption checking: The author will check the four assumptions of logistic regression as the following.

1. The outcome is a binary variable This binary dependent variable assumption is definitely true, where the outcome variable is either visit parks more frequently (1) or less frequently (0).

2. There is a linear relationship between the logit of the continuous predictor variables. Here we do not need to check the assumption, since all the predictor variables are all two-level variables instead of continuous variables.

3. There is no influential values (extreme values or outliers) For the third assumption, since extreme values in the logistic regression model can alter the model's quality, below is the residual plots conducted for the examination, where an extreme value is a point with a large residual.

Figure 2:Residual Plots (scatter plot)



Minitab flags any observation with an internally studentized residual that is larger than 2 (in absolute value)

4. There is no multicollinearity among the predictors. For the forth assumption, the collinearity assumption means that two or more independent variables cannot be highly correlated with each other. A VIF (Variance inflation factor) test might be appropriate for testing the collinearity among the independent variables.

Table 1: the GVIF for model (Relationship)

	x
`below 12`	1.45
What_Relax	1.10
What_Play	1.50
Pet	1.04
Disability	1.04
Location	1.09

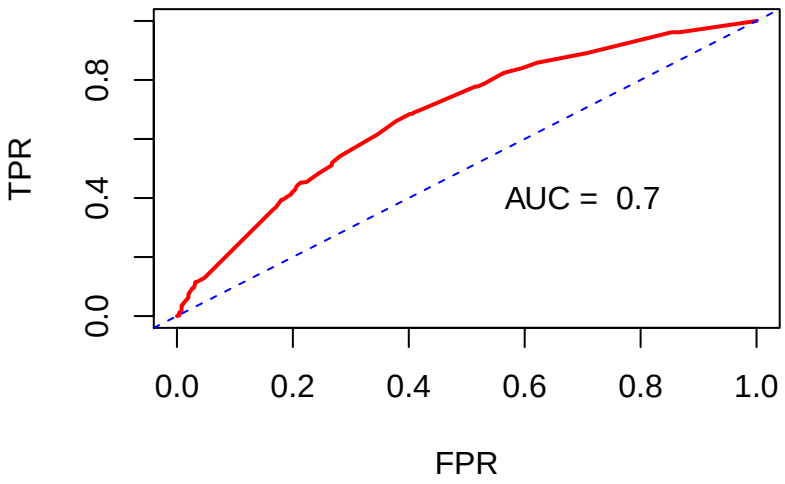
Similarly, the threshold for multicollinearity 2.5 here, i.e., VIF value > 2.5 as an indicator of multicollinearity.

Overall, the author can conclude that this model overall appears to be reasonably well-fitted for all the assumption listed above based on that all points in the residual plots within the line with 2 and all variables with VIF value less than 2.5.

Goodness of Final Model: To further investigate how this model fits the actual values, the ROC(receiver operating characteristic) curve was conducted to test the goodness of the fit.

More precisely, the ROC curve conducted here uses the 1000 training data to fit the model, then fit the 315 test data which were divided out from the original 1315 data set to see the model's performance. Here the x-axis FPR (False Positive Rate) stands for how many correct positive results occur among all positive samples available during the test; the y-axis TPR (True Positive Rate) stands for how many incorrect positive results occur among all negative samples available during the test. The best possible prediction would yield a point in the upper left corner, representing no false negatives and all true positives. Furthering, AUC (area under the curve) represents the entire two-dimensional area underneath the entire ROC curve. Here we prefer the AUC to be approximate 1.

Figure 3: ROC Curve



Here from the ROC curve, we can notice that the model has a relatively good prediction, where the AUC is 0.7, i.e., the model can correctly discriminate between the events 70% of the times, which is relatively good. However, as a rule of thumb, an AUC above 0.85 means high classification accuracy (D' Agostino, Rodgers, & Mauck, 2018). Here the AUC we have is 0.7, which might deduct a fair performance of the model we fit.

3. Results

Table 2: odds ratio coefficient & 95% CI for final model

	odds value	2.5 %	97.5 %	have evidence
(Intercept)	0.29	0.21	0.39	Yes
`below 12`	1.75	1.25	2.45	Yes
What_Relax	1.68	1.26	2.23	Yes
What_Play	2.23	1.61	3.08	Yes
Pet	2.17	1.53	3.11	Yes
Disability1	1.59	1.09	2.33	Yes
LocationS	1.7	1.28	2.26	Yes

Here the Logistic Regression model we have is:

$$\frac{p}{1-p} = 0.29 + 1.75 * (Age < 12) + 1.68 * (Relax) + 2.23 * (Play) + 2.17 * (Pet) + 1.59 * (Disability) + 1.70 * (Southern)$$

Here all the Xs are two-level variables describing the condition of the visitor, with 1 satisfies the condition and 0 does not satisfy the condition. The baseline for all the independent variables here is 0.

Here the intercept $\beta_0 = 0.29$ means that when all X conditions do not satisfy, i.e., the visitor's age is above 12; Relax, play all not included in the activity the visitor go to parks; the visitor does go to park with pets and do not have any disability; the visitors tend to visit parks located in the northern Toronto, the odds of visitor go to park with higher frequency is 0.29.

For a better interpretation, the author exponentiate the summary results and the values column represents for the odds ratio. More precisely, for some specific parameter β_i, $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X1 + ... + \beta_i X_i + ... + \beta_n X_n$, holding all other variables the same, recall that all independent variables in the model are two-level variables, for 1 unit increase in X_i, i.e., $X_i : 0 \to 1$, multiplies the odds of having the outcome by e^{β_i}

For the coefficient of independent variables, holding constant all other variables:

- For Age: The group of visitors with some of them are 12-years-old or younger, i.e., visitors go to parks with children($<$ 12-years-old) will have a 1.75 times odds ratio of visiting parks with a higher frequency than those who do not. The 95% Confidence Interval(CI) calculated is [1.25, 2.45], the interval is larger than 1, which means the we have 95% confidence that visitors with children will go to parks more frequently.
- For Activity: Visitors who go to parks and relax (play) will have a 1.68 (2.23) times odds ratio of visiting parks with a higher frequency than those who do not relax in parks. The 95% Confidence Interval(CI) for relax and play are [1.26, 2.23], [1.61, 3.08], respectively, all larger than 1. In other words, we have 95% confidence that visitors go to parks to play or relax will go there more frequently
- For pet: Visitors who go to parks with pets will have a 2.17 times odds ratio of visiting parks with a higher frequency than those who do not. The 95% Confidence Interval(CI) calculated is [1.53, 3.11], which is larger than 1. In other words, we have 95% confidence that visitors with pets will go to parks more frequently.

- Fir Disability: Disabled visitors will have a 1.59 times odds ratio of higher frequency than those who do not have a disability. The 95% Confidence Interval(CI) calculated is [1.09, 2.33], the interval is larger than 1, which means the we have 95% confidence that disabled visitors will go to parks more frequently.
- For geographic feature: Visitors who visit parks in the southern part of Toronto will have a 1.70 times odds ratio of higher frequency than the visitors in the northern part of Toronto. The 95% Confidence Interval(CI) calculated is [1.28, 2.26], the interval is larger than 1, which means the we have 95% confidence that visitors in the Southern Toronto will go to parks more frequently.

4. Conclusion:

Conclusion

Overall, from the parameters shown in the final model I selected, I can conclude that visitors are more likely to go to parks more frequently if visitors go to parks with pets/children show a higher frequency of visiting parks. Furthermore, visitors are more likely to visit parks more frequently if they go to parks to relax or to play. Visitors with disabilities will be more likely to visit parks with higher frequency compared to visitors without a disability. As for the geographic features, visitors in southern parks of Toronto will visit parks more frequently than visitors in northern parks of Toronto.

Hence, based on the results obtained above, I would suggest that parks include more facilities for relaxation, such as benches, green spaces, and pet-friendly facilities. At the same time, the parks can also include more walking areas, such as cobble roads, or more playgrounds, instead of just green spaces. In addition, parks should also include more facilities for disabled people, particularly easy access to buildings in parks, such as toilets.

The challenges and limitation when analyzing these surveys and implementing the analysis

1. Here the author just selected 5 parks from the 37 parks, the model might well-fitted for the 5 parks while might not work well for other parks.
2. The relatively low AUC value(0.7) suggest a fair prediction while not an accurate result from the model.
3. The difference among survey questions of different parks might increase bias on the visiting frequency. Different parks have different survey questions/focuses, the parks survey selected on the model may not catch the factors that affect visitors' attendance to park appropriately.
4. As for the geographic feature, most of the parks located on the center line of Toronto. It might be hard to fit or predict the parks' attendance on the west of east part of Toronto.

Potential solutions to the challenges mentioned above

1. The author should try to improve the model performance by applying other kind of statistical methods/models or improve/filter the data.
2. The author should conduct similar survey questions to different spaces for the uniformation of features exploration.
3. As for the geographic features in this report, the author need to uniformly include more data from different location of Toronto.