Impact of business categories and endorsement on their license durability after 2000*

An analysis of business licences issued by the Toronto Municipal Licensing and Standards.

Geunchul Shin

March 13, 2024

Municipal Licensing & Standards (ML&S) issues licences to various type of businesses and trades in Toronto city. This licence dataset is used for finding the impact of various factors influencing licence durations. The data set is analysed and compared for the period 2000 and beyond. Evidently, the linear model suggests that the category `HOLISTIC CENTRE` has the most significant negative impact on issue duration, while the number of endorsements does not significantly affect issue duration. The model explains only a small proportion of the variability in issue duration, indicating that other factors beyond the predictors included in the model may influence issue duration.

Table of contents

*Code and data are available at: https://github.com/geunchulshin/Business_licences_analysis

1 Introduction

The Municipal Licensing & Standards (ML&S) department grants licenses to a variety of businesses and trades within the city of Toronto Toronto Elections (2023). To investigate the trends in the distribution of licences, their duration and permits across different categories, a thorough investigation was carried out for all the licenses given since 2000. This choice allows for a comprehensive exploration of licensing and permitting distribution patterns, considering the potential influence of electoral on regulatory activities and administrative decision-making within the municipal landscape in the last 20 years or more. The Court Services Division of the City of Toronto provides administrative support to the Tribunal. Unless otherwise specified, public hearings are held every Thursday. The Tribunal makes rulings free from commercial or political influence Toronto Tribunal (2023).

In this paper, it is being considered that there are numerous factors, including both legislative and larger social-economic dynamics, that play a role in licencing choices. Those factors could be related to regulatory compliance, public safety and welfare, applicant qualification, land use and zoning, economic impact, environmental considerations, and social/political priorities. Hence, the difference in the pattern could be because of those influential factors, that lead to change in the policy decisions.

There are some categories for which most applications are received for permits and licenses. Building permits are for construction, film permits for making movies, lottery licenses for charitable events, marriage licenses for weddings, and taxi licenses for driving taxis City of Toronto (Year Accessed). These permits help ensure things are done safely and responsibly, making sure activities like building, filming, fundraising, getting married, or driving taxis follow the rules that keep everyone safe and organized in the community. The research question arises, “Does licence duration differ as per category and number of endorsements the company currently holds?”. In order to find the answer to this research question, an effort is being made to find the impact of business categories and endorsement on the licence duration.

For the analysis, the null hypothesis is defined as business category and endorsement numbers do not exert a significant influence on license duration. In other words, the analysis assumes that variations in license duration cannot be attributed to differences in business categories or endorsement numbers. Testing this hypothesis is crucial for determining whether observed differences in license durations are statistically meaningful or simply due to random chance. The result shows that there are a few categories which have a significant impact on the licence term while the other being having almost zero impact.

Our primary estimand is the difference in average license duration across different business categories and the average change in license duration associated with each additional endorsement.

The purpose of this paper is to find which factors or categories impact the licence durability. The remainder of this paper is structured as follows. R Core Team (2022) is used in this report to clean, visualize and model fitting. In future this report can be helpful to analyze and consider other factors impacting licencing. Section 2 elaborate on the data-set used for the analysis. Section 3 section explains about the linear model fitted on the data-set. Section 4 shows the result of the analysis. Section 5 explains and discusses the results in detail. Section 6 concludes the analysis i.e. the findings of this paper.

2 Data

2.1 Data Description

The data set used in this analysis is obtained from the City of Toronto's Open-Data Toronto Library Gelfand (2022). The data set is entitled with the name "Municipal Licensing and Standards - Business Licences and Permits" City of Toronto (Year Accessed). Along with that, the following packages are also used, ggplot2 for plotting Wickham (2016), tidyverse for data frame manipulation Wickham et al. (2019), knitr for pdf rendering Xie (2014), Goodrich et al. (2022) package is used to fit the model, Gabry and Mahr (2024) is used to plot the diagnostics and kableExtra for styling tables Zhu (2021).

2.2 Methodology

The raw data set contains 32000 observations each having 18 features. However, for the analysis, I only require a few of the features. The final data-set contains 4 variables i.e. Category, number of endorsements, year and licence duration. I also performed feature engineering to generate a couple of features. These will be discussed further in detail. A sample of the cleaned data set is shown in Table 1.

In the data cleaning process, several steps were undertaken to refine and prepare the dataset for analysis. Initially, a set of specific columns deemed unnecessary for the analysis, such as

Table 1: Sample of cleaned Municipal Licensing & Standards data set (Issue Duration in Months)

Category	Num Endorsements	Issued year	Issue duration
PRIVATE TRANSPORTATION COMPANY	1	2018	10.61104
PRIVATE TRANSPORTATION COMPANY	1	2017	15.04599
PRIVATE TRANSPORTATION COMPANY	1	2018	45.53219
PRIVATE TRANSPORTATION COMPANY	1	2017	71.41919
PRIVATE TRANSPORTATION COMPANY	1	2020	33.31143
PRIVATE TRANSPORTATION COMPANY	1	2021	20.00657

"_id," "Licence No.," and others related to contact details and record updates, were identified and removed. Additionally, to enhance clarity and ease of analysis, columns containing text-based information, such as "Conditions" and "Endorsements," underwent processing. A custom function was applied to calculate the number of words in each entry of the "Endorsements" column. Furthermore, date-related columns such as Issue_date and Cancel_date were modified to extract the year and month of issuance and cancellation, and a new column, "Issue_duration," was created to represent the duration between issuance and cancellation in months. Finally, several columns containing redundant or sensitive information were dropped to streamline the dataset for subsequent analysis. These meticulous data-cleaning steps ensure a more focused, structured, and standardized dataset, laying the foundation for the subsequent stages.

2.3 Visualization

The table shows category wise percentage of licenses issued during this period. The overall distribution is shown here Table 2. Figure 3 shows the box-plot of licence duration category-wise.

Table 2: Licenses issued during all years

Category	Percentage(%)
ADULT ENTERTAINMENT CLUB	0.1675120
BILLIARD HALL	0.5583733
BOATS FOR HIRE	0.0139593
BODY RUB PARLOUR	0.4094737
BOWLING HOUSE	0.1302871
CARNIVAL	0.0465311
CIRCUS	0.0139593
DRIVE-SELF RENTAL OWNER	1.7728351
DRIVING SCHOOL OPERATOR (B)	2.5405984
HOLISTIC CENTRE	8.3569866
LIMOUSINE SERVICE COMPANY	1.0329906
PERSONAL SERVICES SETTINGS	1.4936485
PLACE OF AMUSEMENT	1.2935647
PRIVATE PARKING ENFORCEMENT AGENCY	0.7538039
PRIVATE TRANSPORTATION COMPANY	0.0418780
RETAIL STORE (FOOD)	79.3076171
SMOKE SHOP	1.3726676
TAXICAB BROKER	0.4280862
TAXICAB OPERATOR	0.1163278
TEMPORARY SIGN PROVIDER	0.1488995

3 Model

By fitting a linear model to the license dataset, we can identify which factors, such as business category and number of endorsements, have the most significant impact on license durations, as by understanding the factors that influence license durations for businesses and trades in Toronto is crucial for efficient municipal operations and supporting local businesses. Linear models offer a straightforward approach to understanding relationships between variables, providing interpretable coefficients that quantify the impact of predictors on the response variable. Additionally, they require fewer assumptions compared to more complex models, making them suitable for a wide range of data-sets and facilitating easier interpretation of results. The linear model that we fitted is defined as below:

$$Issue\_duration = \beta_0 + \beta_1 * Category + \beta_2 * Num_Endorsements + \epsilon$$

where:

- Issue_duration is the response variable.
- β_0 is intercept term.
- β_1 and β_2 are the coefficients associated with the predictors (Category and Num_Endorsements).
- ϵ is the error term.

As we also fit the log linear model, when fitting a log-linear model, the equation structure changes to accommodate the logarithmic transformation of the response variable or predictor variables as below:

$$log(Issue\_duration) = \beta_0 + \beta_1 * Category + \beta_2 * log(Num_Endorsements) + \epsilon$$

where:

- $log(Issue\_duration)$ is the natural logarithm of the response variable.
- $log(Num\_Endorsements)$ is the natural logarithm of the predictor variable.
- β_0 is intercept term.
- β_1 and β_2 are the coefficients associated with the predictors (Category and Num_Endorsements).
- ϵ is the error term.

4 Results

4.1 Exploratory Analysis

Comparing the data set, it appears that certain categories, such as "DRIVE-SELF RENTAL OWNER" and "PRIVATE TRANSPORTATION COMPANY," consistently have lower endorsement frequencies, reflecting potential operational trends or regulatory considerations. The variations in endorsement numbers between the two time periods underscore the dynamic nature of endorsement activities across different business categories, suggesting changes in regulatory focus or industry dynamics over time.

4.2 Linear model

The analysis of license duration for businesses and trades in Toronto reveals interesting insights as shown in Figure 1 and plots of linear model (Figure 5, Figure 6, Figure 7, Figure 8). The linear model suggests that certain business categories, such as 'HOLISTIC CENTRE', have a significant impact on license duration, with shorter durations observed in this category compared to others. However, the number of endorsements does not seem to have a significant effect on license duration, as indicated by its non-significant coefficient and p-value. Despite these findings, the model only explains a small proportion (approximately 1.3%) of the variability in license duration, suggesting that other factors not included in the model may also influence license durations. The relatively high residual standard error of around 40.26 indicates some level of deviation of observed license durations from the predicted values by the model. While the model is statistically significant overall, with a very low p-value ($< 2.2e\text{-}16$), its limited explanatory power underscores the complexity of factors influencing license durations beyond those captured in the analysis.

4.3 Log-Linear model

The log-linear model suggests that while certain business categories still exhibit significant impacts on license duration (as shown in Figure 9, Figure 10, Figure 11, Figure 12), the effects are somewhat attenuated compared to the linear model. Notably, businesses categorized as 'HOLISTIC CENTRE' continue to demonstrate a significant negative impact on license duration. However, the logarithmically transformed number of endorsements does not appear to significantly influence license duration, similar to the findings of the linear model. Overall, the model explains only a small proportion (approximately 1.4%) of the variability in license duration, indicating the presence of other unaccounted factors. Despite the relatively high residual standard error of around 0.9323, the model remains statistically significant overall, as evidenced by the very low p-value ($< 2.2e\text{-}16$).

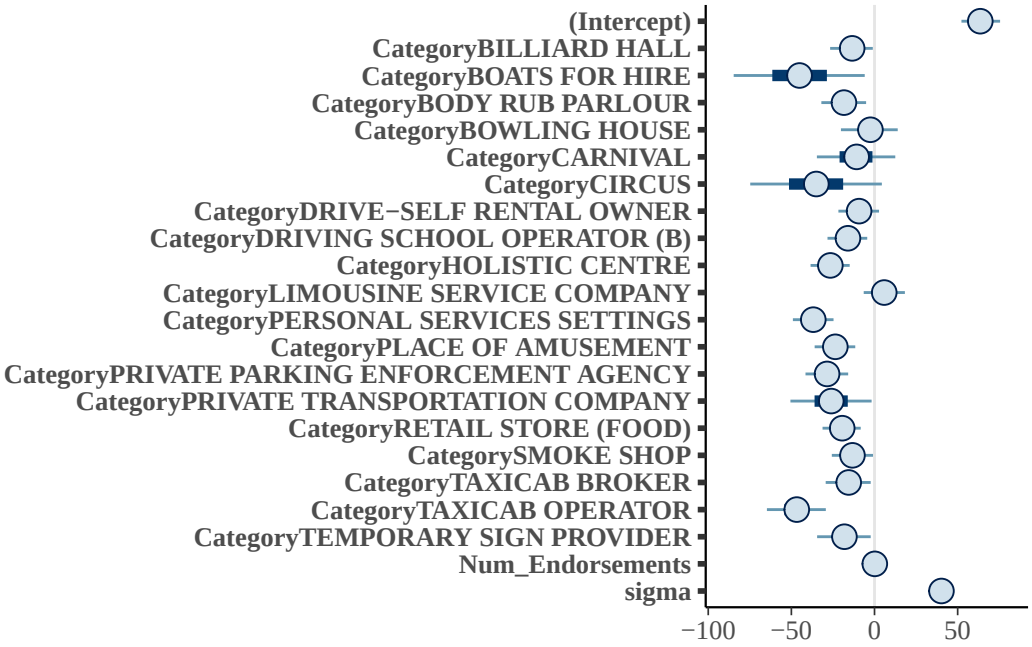

Figure 1: Factors impacting licence duration

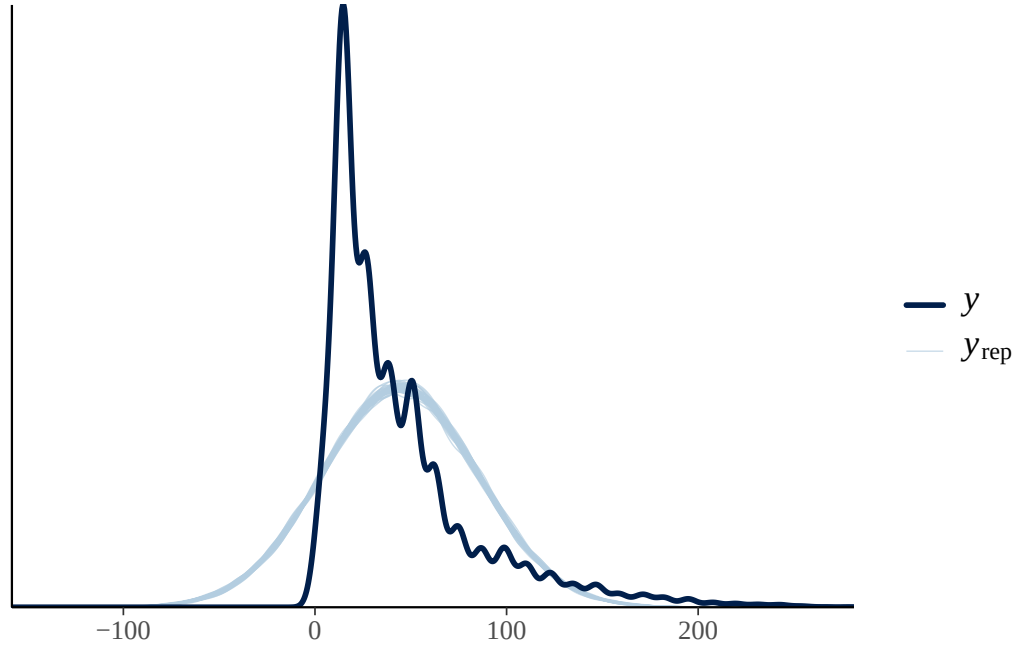

Figure 2: Prior vs Posterior Probability

5 Discussion

5.1 Linear and Log-linear model summary

Analyzing licenses and permits in Toronto shows interesting trends in how businesses operate. Some types of businesses, like those dealing with temporary signs, taxi brokering, and private transportation, seem to have longer-lasting licenses.

The analysis of license durations using both linear and log-linear models provides valuable insights, but neither model alone is strong enough to definitively prove or disprove the null hypothesis. While the models offer statistical significance for certain predictors like business categories, they only explain a small proportion of the variability in license duration, suggesting the presence of other unaccounted factors.

Comparing the two models, the log-linear model incorporates logarithmic transformations of the response variable and predictor, offering a different perspective on the relationships. However, it does not necessarily perform better than the simple linear model, as both models have similar limitations in explaining the variability in license durations.

Additionally, the results from the stan_glm model provide insights into the factors influencing license durations in a straightforward manner. The estimated coefficients indicate the average effect of each predictor on license duration. For instance, a positive coefficient suggests that an increase in the predictor is associated with a longer license duration, while a negative coefficient suggests the opposite.

In this analysis, the intercept represents the average license duration when all predictors are at their reference levels. The coefficients for different business categories indicate how each category affects license duration compared to the reference category. For example, the 'HOLISTIC CENTRE' category has a negative coefficient, indicating that businesses in this category tend to have shorter license durations compared to the reference category.

Similarly, the coefficient for 'Num_Endorsements' indicates the average change in license duration for each additional endorsement. The standard deviation (sd) provides information about the variability in the estimated coefficients.

Additionally, the fit diagnostics provide information about the model's performance and convergence. The mean_PPD represents the average posterior predictive distribution of the outcome variable, giving an indication of how well the model fits the data. The MCMC diagnostics, including Monte Carlo standard error (MCSE), effective sample size (n_eff), and potential scale reduction factor (Rhat), assess the reliability of the parameter estimates and the convergence of the Markov chain Monte Carlo (MCMC) sampling process as shown in Figure 2.

Overall, these results help us understand the factors influencing license durations and assess the model's performance in capturing these relationships.

The insights from models help both the government and businesses make informed decisions and understand the factors shaping licensing and rules in Toronto.

5.2 Weaknesses

Linear models have drawbacks. They may not capture all relevant factors influencing license durations, leading to residual variability in predictions. Additionally, the models rely on certain assumptions, such as linearity and homoscedasticity, which may not always hold true in real-world scenarios.

These observations point towards dynamic changes in licensing dynamics, possibly influenced by evolving regulatory frameworks or industry trends. This comparison underscores the importance of adapting licensing policies to the evolving landscape of businesses, ensuring that regulatory frameworks remain responsive to the needs and dynamics of various industry sectors. Hence, more variables and non-linear models can be used to get the strong influencing factors.

5.3 Future Directions

Moving forward, future research could explore additional predictors or alternative modelling approaches to better understand and predict license durations. Incorporating more comprehensive datasets and considering nonlinear relationships may improve model performance. Additionally, qualitative research methods could provide deeper insights into the complex factors influencing license durations, complementing the quantitative analysis provided by the models.

6 Conclusion

In summary, the analysis of licensing data in Toronto unveils dynamic trends in both the duration and distribution of licenses across diverse business categories. In examining license durations, two models were used: linear regression and log-linear regression. The linear regression model showed how business categories and endorsement numbers affect license duration, as it estimates and fits diagnostics, offering insights into parameter reliability and model convergence. These models collectively shed light on factors influencing license duration and provided a comprehensive understanding of their relationships. To mitigate the effect of non-linearity, the log-linear model analyzed the same variables but on a logarithmic scale. Both models assume linearity of the data, however, the results show data have some non-linearity, hence in the future more complex models, which can handle non-linearity may be more useful.

From the results, we can infer that license duration is influenced by both business category and endorsement numbers, albeit to varying degrees. The coefficients in the models provide

insight into the strength and direction of these effects. For instance, categories like "HOLISTIC CENTRE" have a significant negative impact on license duration, while others may have weaker or even negligible effects. Similarly, the impact of endorsement numbers appears to be minimal, with coefficients close to zero. However, precise quantification of these effects requires careful interpretation of the coefficients and consideration of the model's overall explanatory power.

7 Appendix

7.1 Data Visualisations

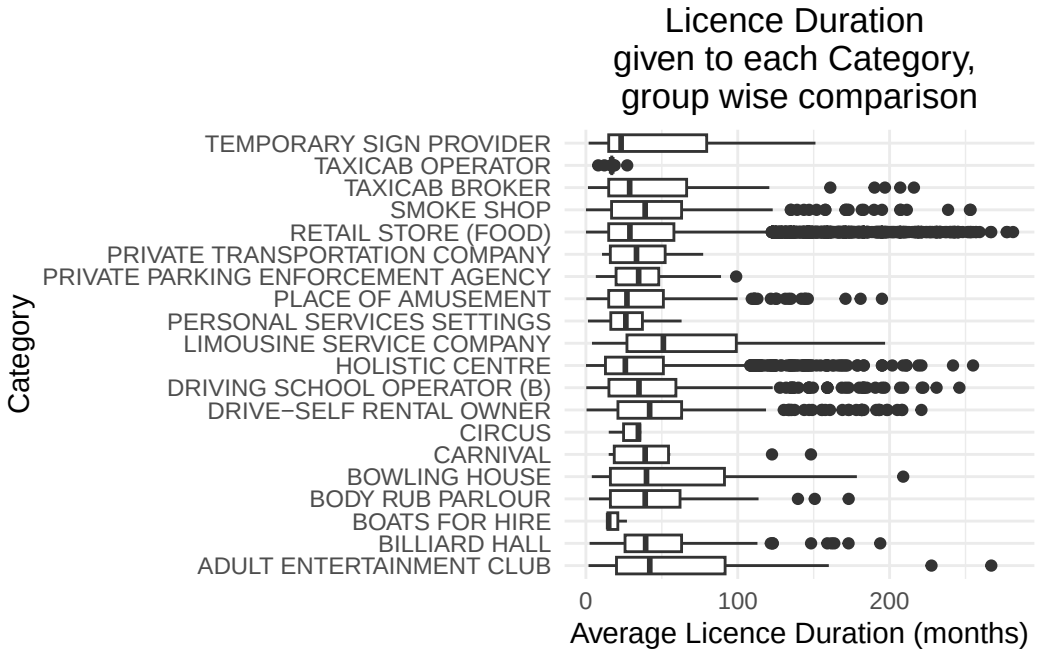

Figure 3: Licenses duration category wise

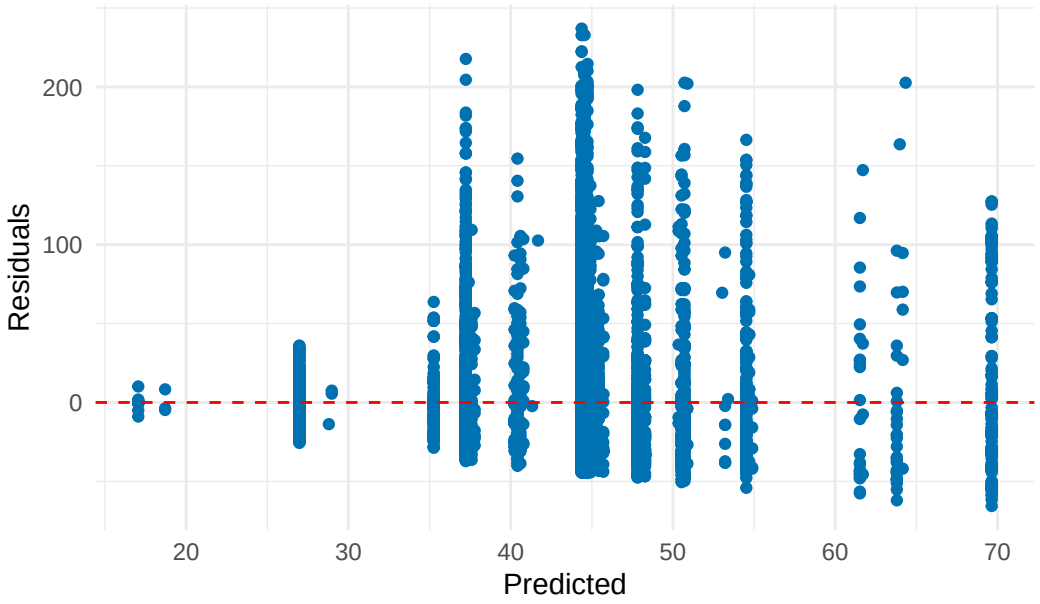

Figure 4: Predicted vs Residual plot Linear model

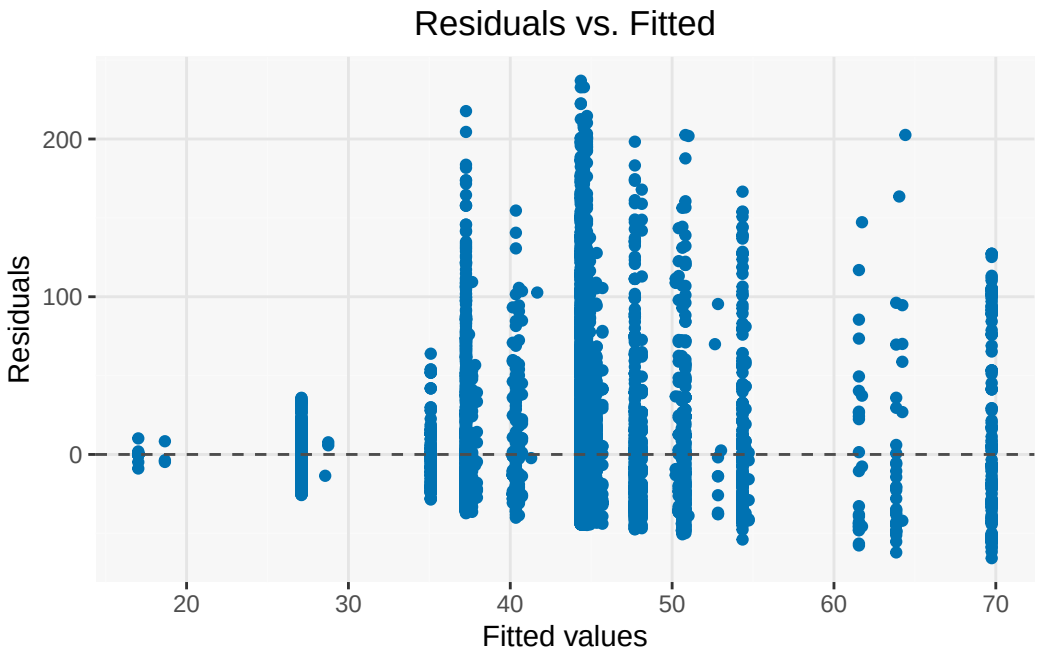

Figure 5: Diagnostic plots of linear model

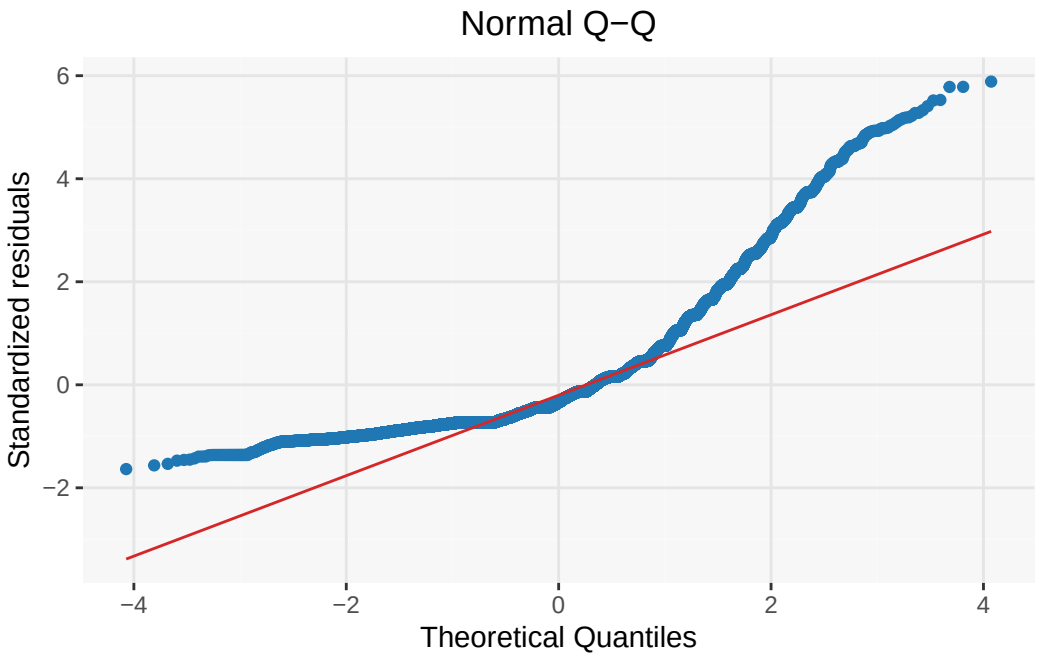

Figure 6: Diagnostic plots of linear model

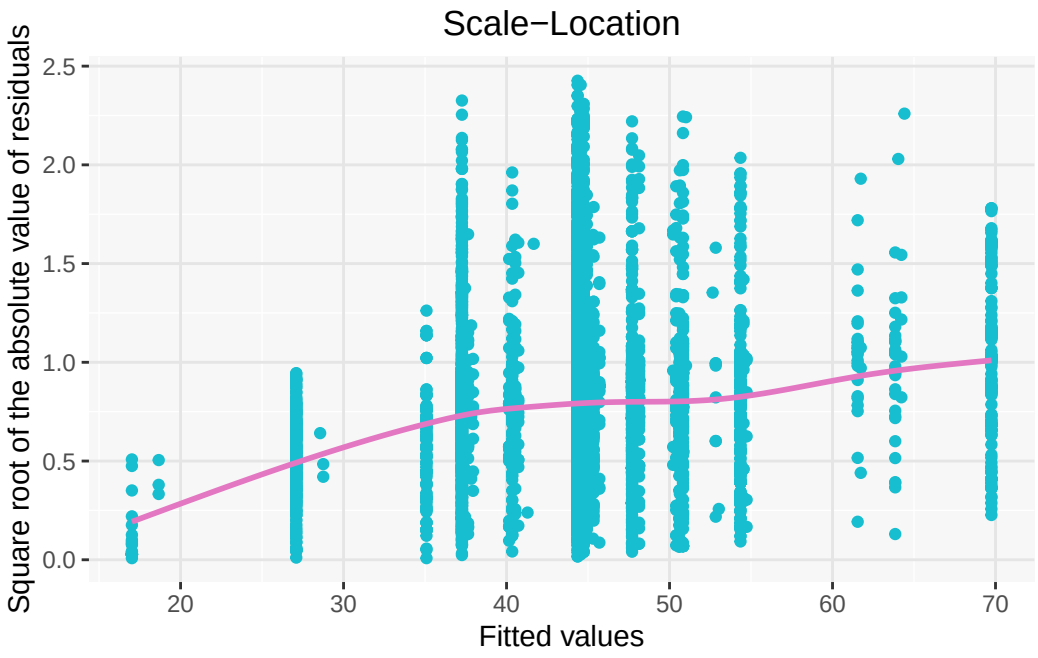

Figure 7: Diagnostic plots of linear model

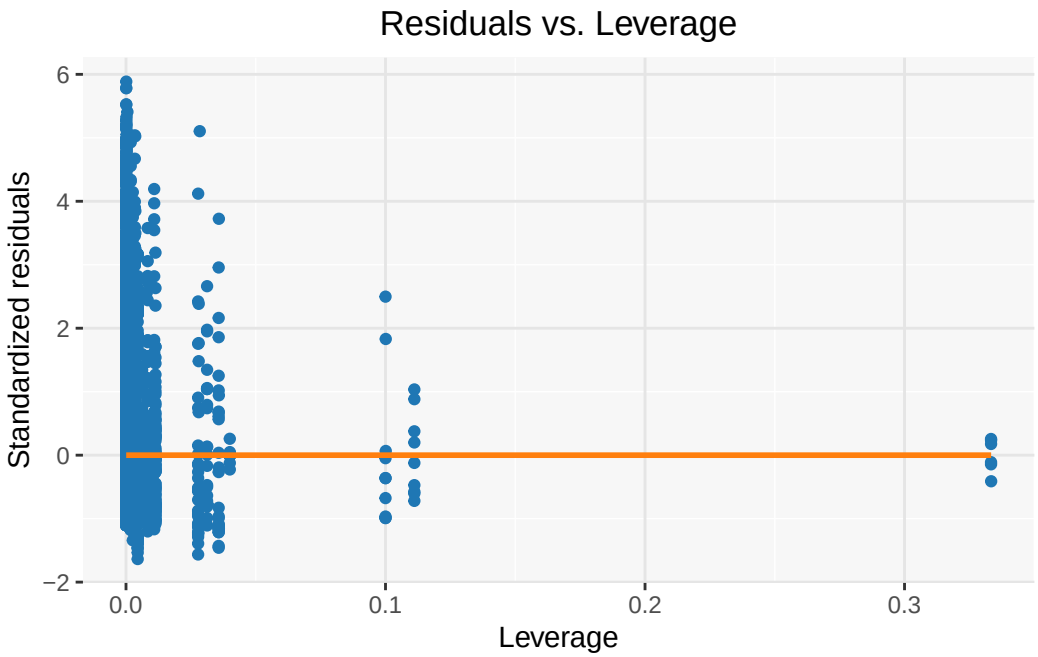

Figure 8: Diagnostic plots of linear model

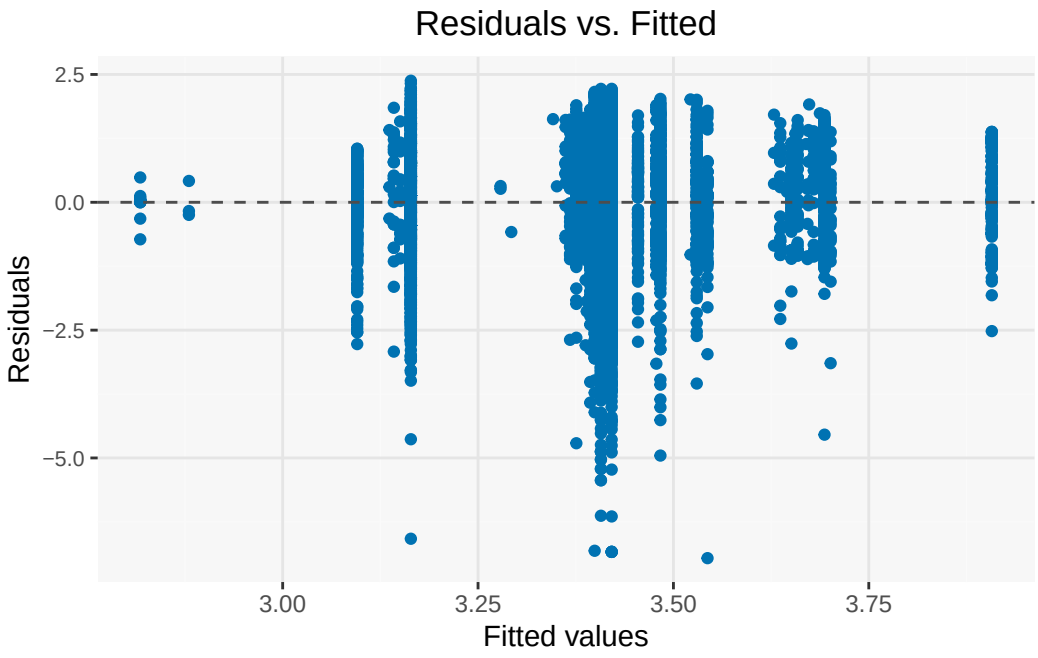

Figure 9: Diagnostic plots of log linear model

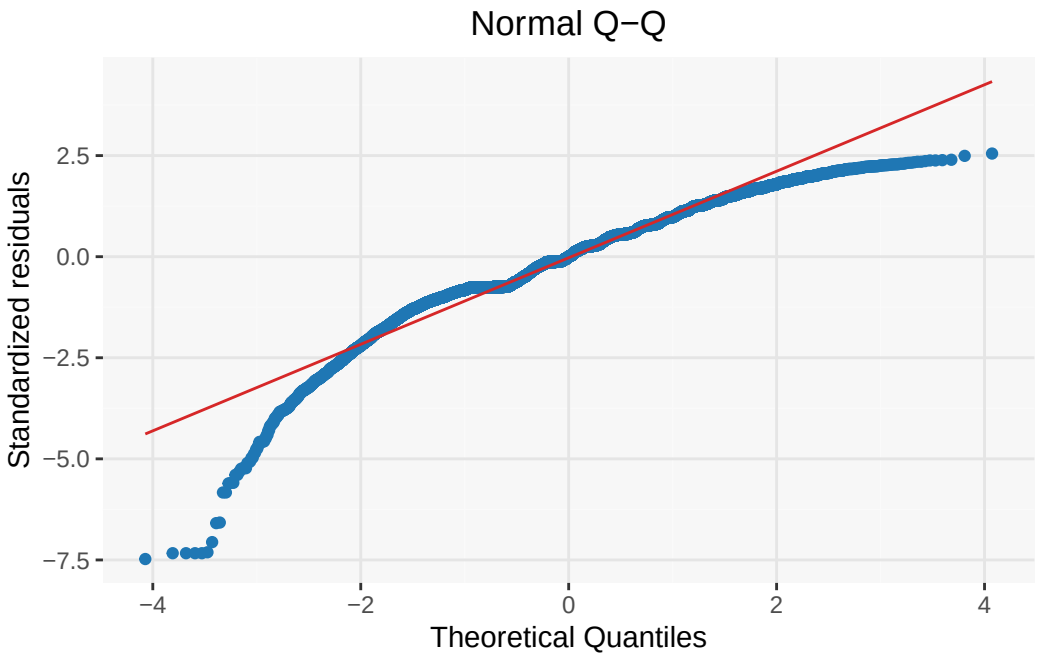

Figure 10: Diagnostic plots of log linear model

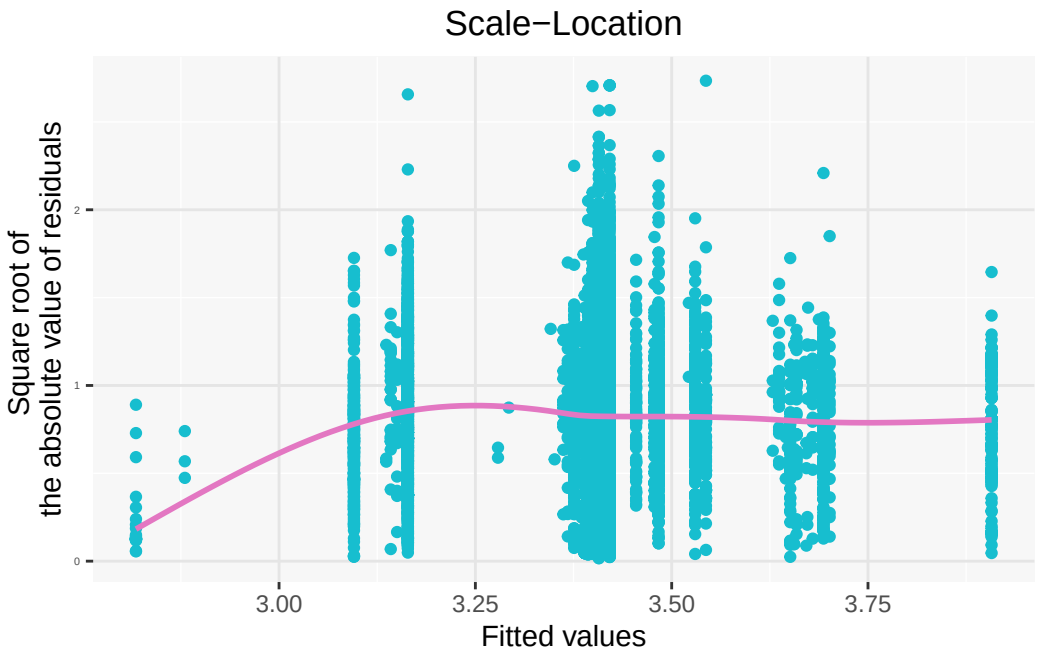

Figure 11: Diagnostic plots of log linear model

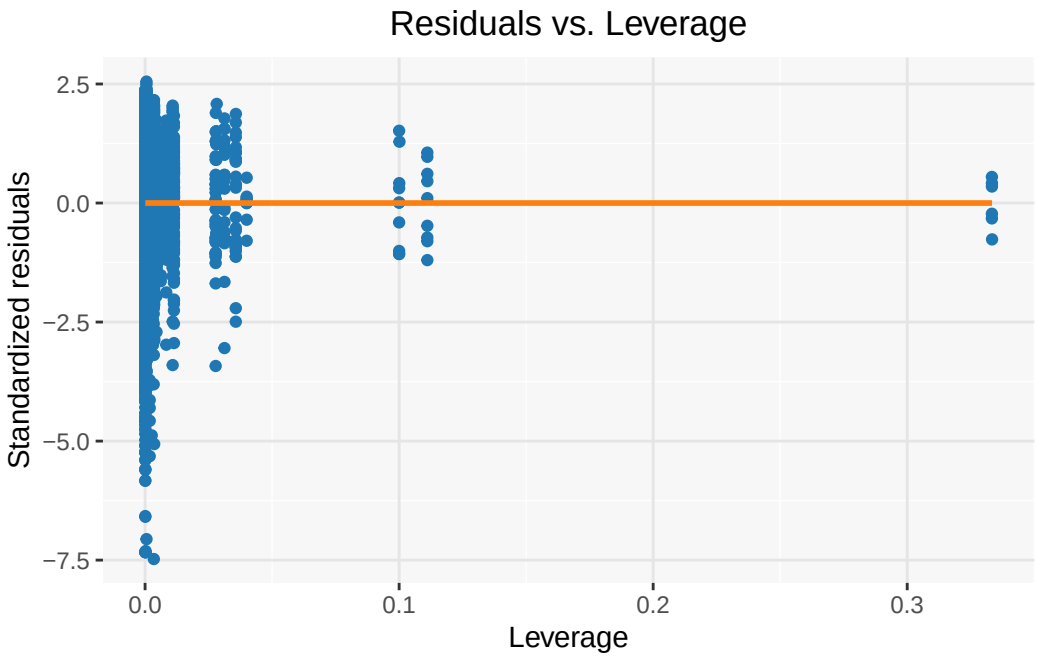

Figure 12: Diagnostic plots of log linear model

References

City of Toronto. Year Accessed. "Municipal Licensing and Standards Business Licences and Permits." https://open.toronto.ca/dataset/municipal-licensing-and-standards-business-licences-and-permits/.

Gabry, Jonah, and Tristan Mahr. 2024. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto Elections. 2023. "2022 Municipal Election Report on Accessibility: City of Toronto." https://www.toronto.ca/wp-content/uploads/2023/01/96e9-Toronto-Elections-2022-Accessibility-Report.pdf.

Toronto Tribunal. 2023. "About the Toronto Licensing Tribunal." https://www.toronto.ca/services-payments/permits-licences-bylaws/toronto-licensing-tribunal/about-the-tribunal/.

Wickham, Hadley. 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.

Xie, Yihui. 2014. *Knitr: Elegant Graphics for Data Analysis*. In Victoria Stodden, Friedrich Leisch; Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.