

Influence on days in shelter

Group 20

Table of contents

1	Introduction	2
2	Description of the Dataset	2
3	Exploratory Data Analysis	3
3.1	Numerical Variables	3
3.2	Categorical Variables	5
3.3	Identifying the Impact and Independence of Predictors on Shelter Stay Duration	9
4	Formal Data Analysis	11
4.1	Model Fitting	11
4.1.1	Poisson Model	11
4.1.2	Negative Binomial Model	17
4.2	Model Selection	21
5	Conclusion	22

1 Introduction

A study is conducted using dataset from the Dallas animal shelter, aiming to uncover the factors affecting how long animals remain at the shelter before a final decision on their outcome is made. The insights gained may help improve animal welfare and shelter efficiency.

2 Description of the Dataset

Each of the 5 datasets contain a variety of information relating to each animal admitted to the shelter.

Table 1: The Description of the Dataset

Variable	Description
type	The type of animal admitted to the shelter
month	Month the animal was admitted
year	Year the animal was admitted
intake	Reason for the animal being admitted
outcome	Final outcome for the admitted animal
chip	Did the animal have a microchip with owner information
duration	Days spent at the shelter between being admitted and the final outcome

Table 1 presents details about the variables in dataset. It is noteworthy that our dataset is complete, with no missing values across variables. We focus on the duration animals stay at the shelter before reaching their final outcome, which is quantified by the variable ‘duration’.

3 Exploratory Data Analysis

3.1 Numerical Variables

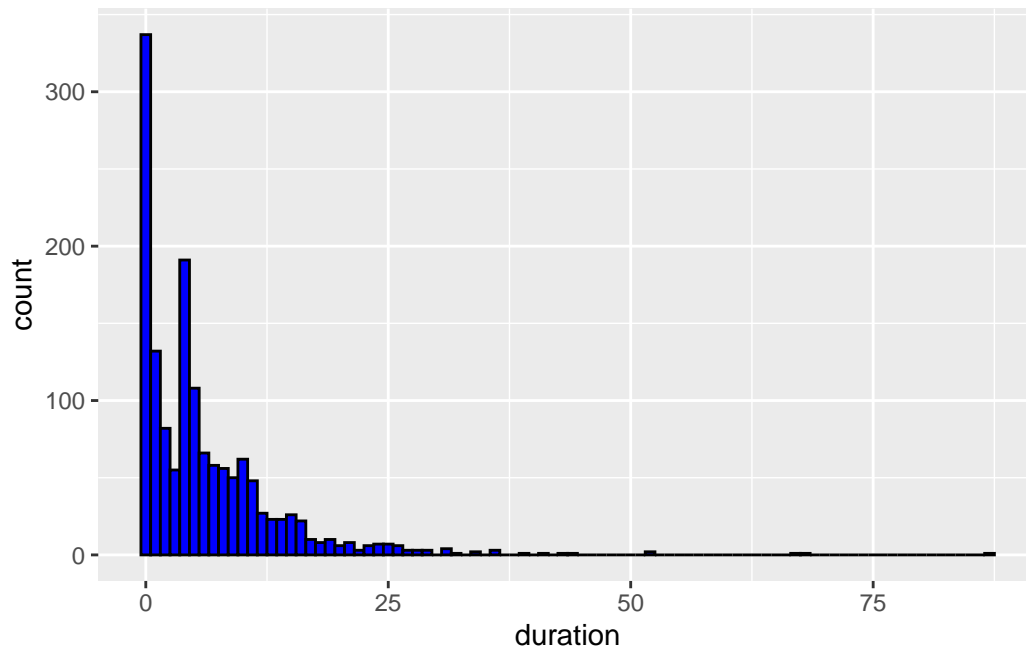


Figure 1: The Distribution of the Duration Variable

As seen in Figure 1, most animals stay in the shelter for a short period of time, while a few animals stay in the shelter for a long time. Although the histogram itself does not display outliers, the long tail suggests that there are relatively few cases that have stayed in the shelter for a very long time, which may be outliers.

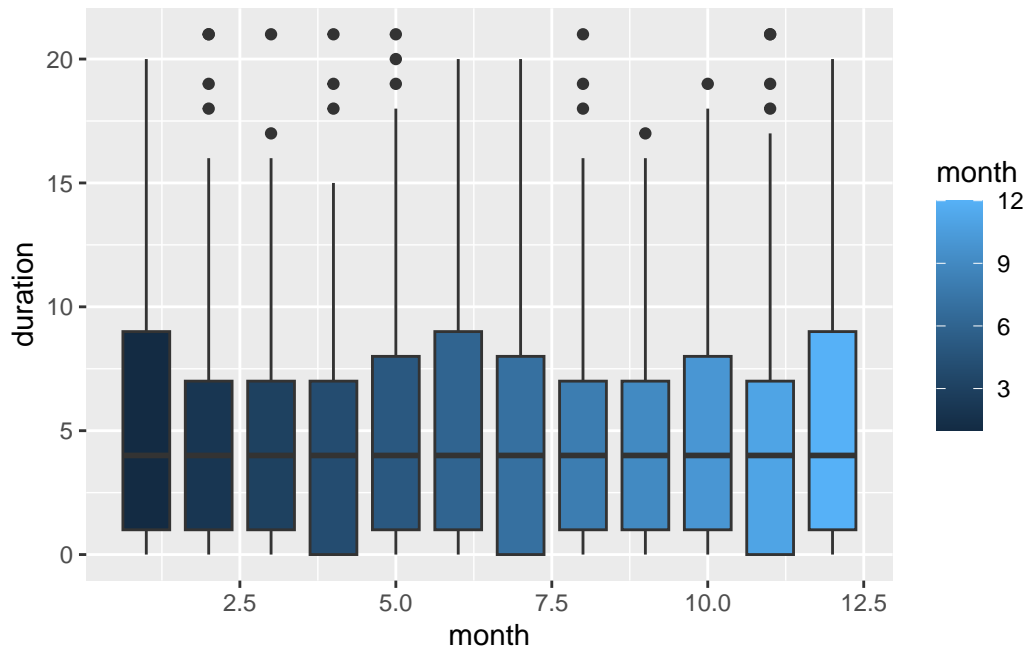


Figure 2: The Distribution of Month vs Duration Variable

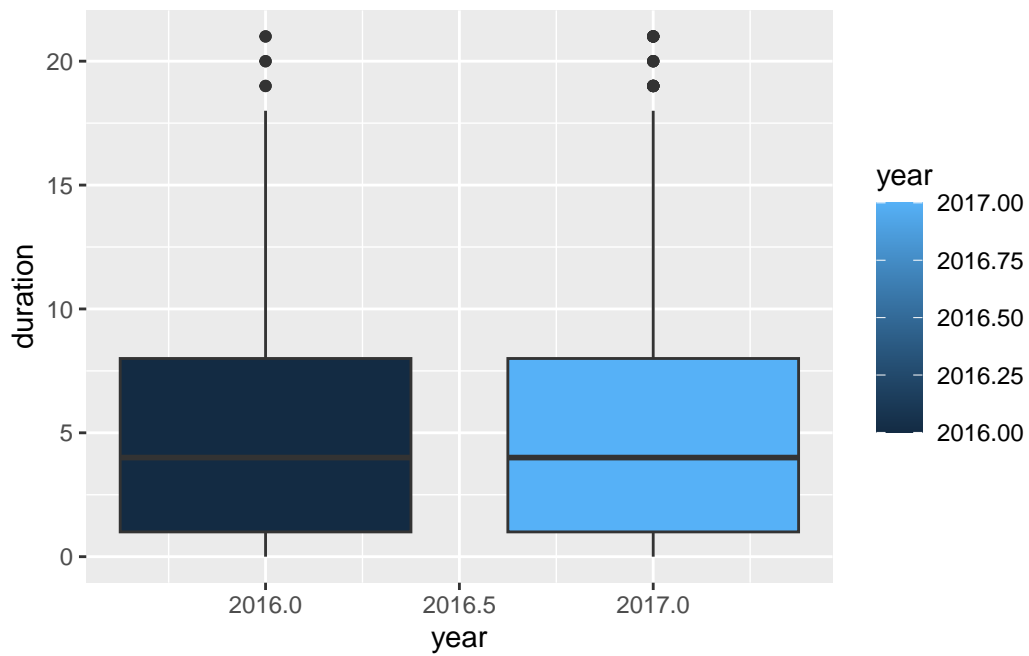


Figure 3: The Distribution of Year vs Duration Variable

Upon close examination of the distribution of shelter stay duration across different months and years, as depicted in Figure 2 and Figure 3, it becomes apparent that the duration exhibits minimal variability in relation to these temporal factors. The box-plots demonstrate that the median duration of stays, as well as the interquartile ranges, do not display significant shifts throughout the months or between the years under consideration. This lack of pronounced fluctuation suggests that the months and years do not markedly influence the length of time animals spend in the shelter. In light of this observation, we may prudently opt to exclude the variables pertaining to months and years from the predictive modelling process. This decision is based on a data-driven rationale, seeking to streamline the model by focusing on predictors that demonstrate a more substantial effect on the outcome variable, thereby enhancing the model’s explanatory power and parsimony.

3.2 Categorical Variables

Table 2: The Distribution of Animal Types

Var1	Freq
BIRD	0.0020478
CAT	0.2075085
DOG	0.7822526
WILDLIFE	0.0081911

In the analysis of our shelter data, we have examined the distribution of animal types and noted that certain categories, specifically birds and wildlife, represent a very small fraction of our dataset—comprising just 0.2% and 0.8% respectively based on Table 2. Given their minimal representation, these categories do not contribute substantially to the variance within our dataset and, consequently, are unlikely to significantly influence the outcomes of our statistical models. To enhance the robustness of our analysis and to focus on the most impactful factors, we have made the decision to exclude these observations from further analysis. This methodological choice is rooted in the principle of statistical significance which posits that only variables with enough data to demonstrate a discernible pattern should be included in the model to ensure accurate and reliable results.

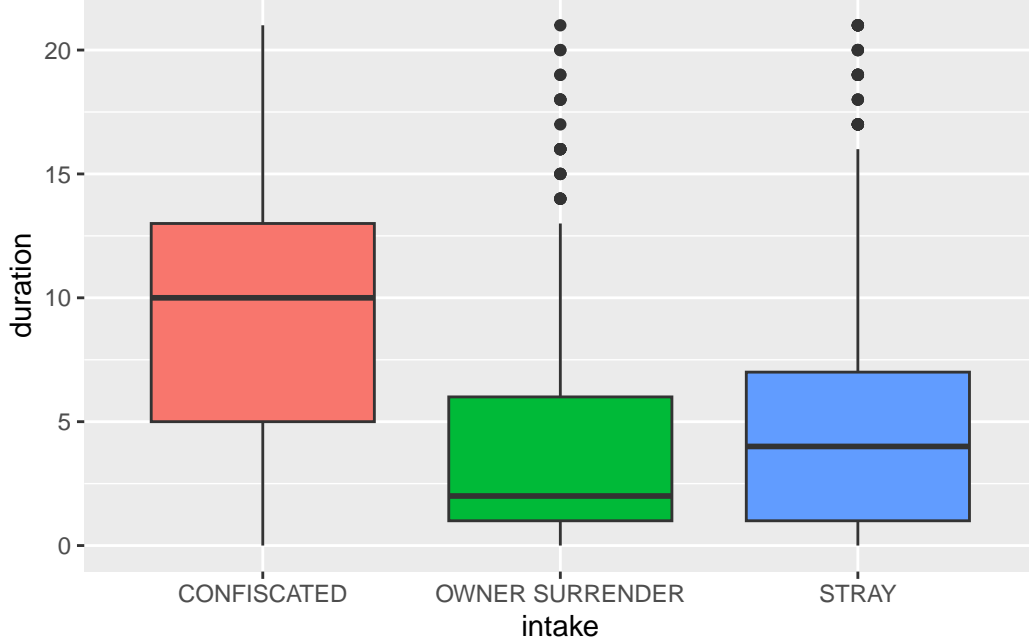


Figure 4: The Distribution of the Intake Variable

Then, upon examining Figure 4, which illustrates the distribution of shelter stay durations across different intake reasons, we observe distinct patterns in the duration variance for each category. The presence of outliers, particularly evident in the categories of ‘OWNER SURRENDER’ and ‘STRAY’, warrants a closer analysis. Outliers can significantly influence statistical analyses, potentially skewing results and leading to misinterpretations. To ensure the integrity of our analysis and to focus on representative patterns, we will proceed to judiciously remove these outliers. This removal will be conducted in a manner that respects the underlying intake reason, thereby preserving the validity of our analysis while enhancing the model’s predictive accuracy. The decision to exclude these data points is grounded in statistical best practices that seek to maintain the robustness of the model and the reliability of the conclusions drawn from it.

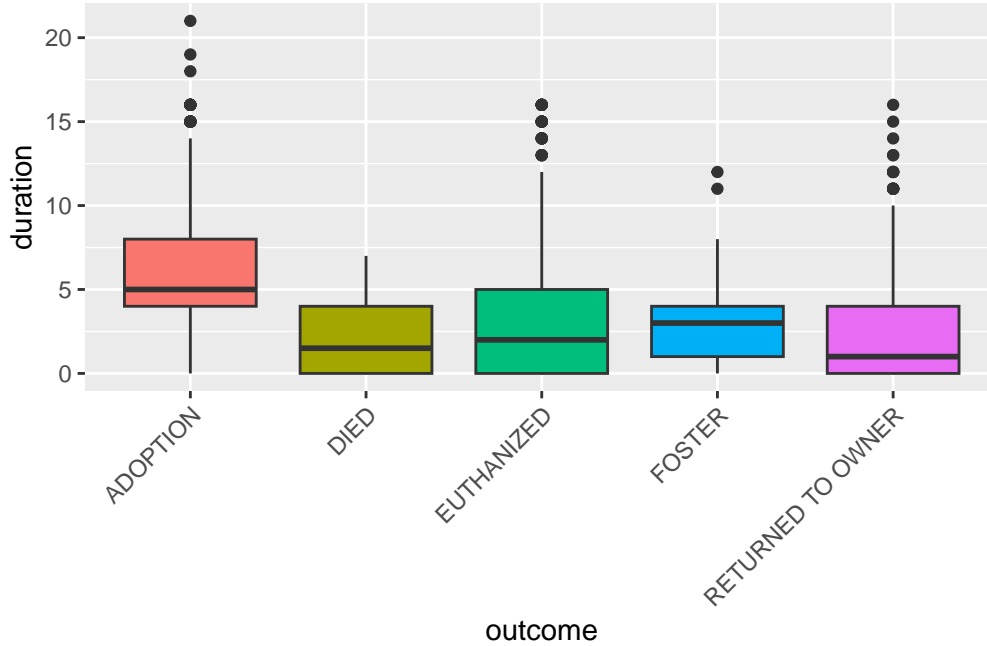


Figure 5: The Distribution of the Outcome Variable

Similar to Figure 5 which presents the distribution of shelter stay duration with respect to the outcomes of the animals' stays, offering valuable insights into the variations across different final outcomes. We notice outliers in the 'ADOPTION', 'EUTHANIZED', 'FOSTER', and 'RETURNED TO OWNER' categories. These are indicated by individual points that reside beyond the upper whisker of their respective box-plots, suggesting significantly longer stays than the typical cases within those categories. Such extreme values can disproportionately influence statistical models and skew the interpretations of the data. To maintain the rigor of our analysis and to avoid these biases, we will remove these outliers.

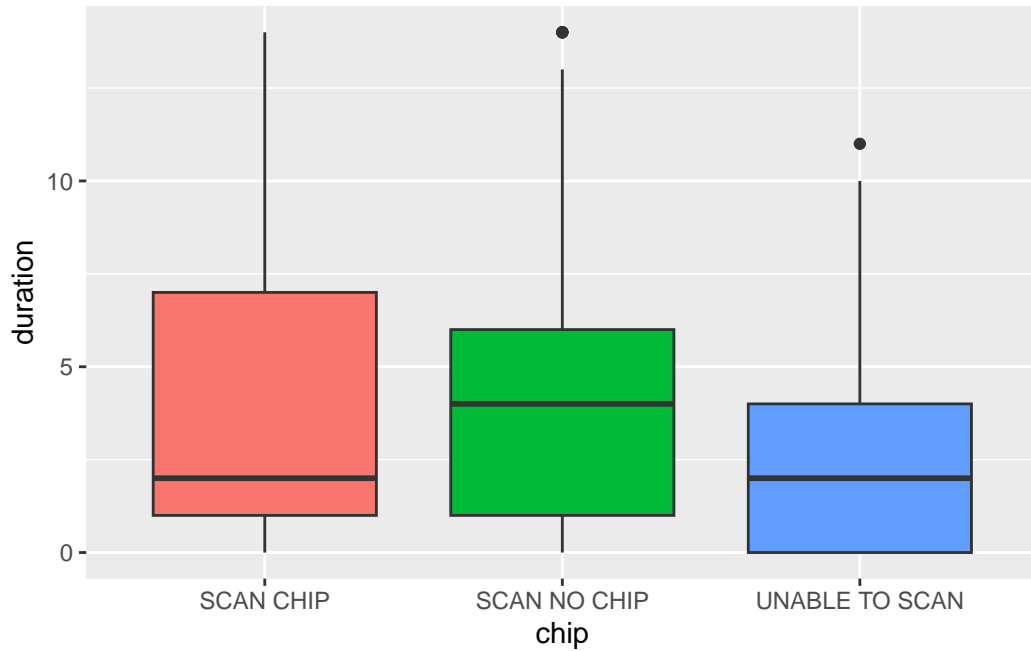


Figure 6: The Distribution of Chip Variable

Moving to the next variable in Figure 6, we are presented with the lengths of stay based on the microchip status of animals within the shelter. It is clear that for the categories — animals without a microchip (SCAN NO CHIP), and animals that could not be scanned (UNABLE TO SCAN) — there are data points that reside notably above the upper whisker, which classifies them as outliers. These points represent atypically long shelter stays, and while they could reflect actual cases, they have the potential to disproportionately affect the statistical analysis and lead to skewed interpretations.

3.3 Identifying the Impact and Independence of Predictors on Shelter Stay Duration

Table 3: Chi-Square Test Results for Association with Duration of Shelter Stay

Variable	P_Value
type	4.4e-06
intake	0.0e+00
outcome	0.0e+00
chip	0.0e+00

The Chi-Square test, a statistical method used to evaluate the association between categorical variables, reveals significant relationships between various factors and the duration of shelter stay. As indicated in Table 3, the P-Value for each categorical variable—type, intake, outcome, and chip status—is below the conventional significance level of 0.05, demonstrating a statistically significant association with the duration variable. These results suggest that the type of animal, the reason for intake, the outcome of the shelter stay, and the presence of a microchip are all factors that correlate significantly with the length of time an animal spends in the shelter. This statistical evidence underpins the importance of these variables in understanding and predicting shelter stay duration, making them essential components for inclusion in any analytical model that seeks to capture the dynamics of animal shelter stays.

Table 4: Multicollinearity Assessment of Predictor Variables

	GVIF	Df	$\text{GVIF}^{(1/(2 \times \text{Df}))}$
type	1.129852	1	1.062945
month	2.287867	1	1.512570
year	2.280442	1	1.510113
intake	1.211155	2	1.049059
outcome	1.357479	4	1.038943
chip	1.140194	2	1.033343

The Generalized VIF (GVIF) are diagnostic measures that assess the extent of multicollinearity in regression analyses when predictor variables are correlated which can obscure individual variable effects and inflate standard errors. Meanwhile the $\text{GVIF}^{(1/(2 \times \text{Df}))}$ adjustment normalizes the GVIF value by the degrees of freedom associated with the categorical predictor, providing a more comparable metric across variables with different numbers of categories.

In Table 4, the $\text{GVIF}^{(1/(2 \times \text{Df}))}$ values for ‘month’ and ‘year’ are close to 1.51 which indicates some degree of multicollinearity, though not at a level traditionally deemed excessive. Nevertheless, it is essential to consider the potential impact on the model. If the ‘month’ and ‘year’

variables do not contribute critical information beyond what other predictors offer, or if their inclusion complicates the model without substantial analytical benefit, it may be judicious to consider their transformation or exclusion.

Meanwhile exhibited from Figure 2 and Figure 3, the similarity in the distribution of ‘duration’ across all months and between the years 2016 and 2017 indicates that time-related trends may not be strong influencing factors for this dataset and also these variables may not be significant predictors of ‘duration’ in the shelter. Given the goal of model parsimony and clarity, omitting ‘month’ and ‘year’ could simplify the model without sacrificing explanatory power or predictive accuracy. This approach is consistent with the desire to focus on variables that provide clear, actionable insights into shelter stay duration and can be defended from both a statistical and practical standpoint. Thus, the combination of acceptable GVIF values and the boxplot observations lead to a reasonable conclusion to exclude ‘month’ and ‘year’ from further modeling efforts.

4 Formal Data Analysis

Moving forward, the formal analysis will pivot to GLM to model shelter stay duration with these significant predictors. This will allow for a more nuanced understanding of how each variable contributes to the duration of stays, taking into account the potential confounding effect of multicollinearity. The GLM approach will also provide a methodological scaffold to isolate and quantify each predictor's impact. The overarching objective of this formal analysis is not just to enhance predictive accuracy but also to inform more strategic and targeted interventions that can reduce shelter stays.

4.1 Model Fitting

4.1.1 Poisson Model

Since the response variable represents the count data, the Poisson regression model is a starting point because it is the simplest model for the type of the data. It was conducted by this result:

Before presenting and interpreting the derived equation from our GLM analysis, it is important to note that validating the model is underlying assumptions is a critical step. Ensuring the assumptions hold true bolsters the reliability of our findings. Here's the equation based on the initial analysis, subject to further validation:

$$\begin{aligned}\log(\text{Expected Count of Time at Shelter}) = & 2.48068 \\ & + 0.18080 \times \text{TypeDog} \\ & - 1.11596 \times \text{IntakeOwnerSurrender} \\ & - 0.63153 \times \text{IntakeStray} \\ & - 0.99932 \times \text{OutcomeDied} \\ & - 0.70378 \times \text{OutcomeEuthanized} \\ & - 0.69878 \times \text{OutcomeFoster} \\ & - 1.31990 \times \text{OutcomeReturnedToOwner} \\ & - 0.14291 \times \text{ChipScanNoChip} \\ & - 0.38348 \times \text{ChipUnableToScan}\end{aligned}$$

where,

- TypeDog is the indicator variable for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category TypeCat.

- IntakeOwnerSurrender and IntakeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeConfiscated.
- OutcomeDied, OutcomeEuthanized, OutcomeFoster, OutcomeReturnedToOwner are indicator variables for the outcome, taking the value 1 if the condition is true and 0 otherwise with the baseline category OutcomeAdoption.
- ChipScanNoChip and ChipUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipScanChip.

The significance of a Generalized Linear Model (GLM) can be assessed by examining the p-values associated with the coefficients of the model, which indicate the strength of evidence against the null hypothesis that the corresponding coefficients are equal to zero. A small p-value (in this case, compared to the significance level 0.05) indicates strong evidence against the null hypothesis, suggesting that it is unlikely to observe such a significant effect if the predictor really had no impact on the response variable.

Conversely, a large p-value suggests insufficient evidence to reject the null hypothesis, indicating that the predictor may not have a significant effect on the response variable.

Derived from the model, all variables are significant predictors of the time an animal spends in the shelter.

4.1.1.1 Model Diagnostics and Assumptions Checking

The analysis utilized a Poisson regression model to investigate the impact of factors such as, including AnimalType, IntakeType, OutcomeType and ChipStatus on the duration of stay in a shelter. To ensure the robustness of our model, we conducted a thorough diagnostic evaluation using several methods. We checked the dispersion parameter to evaluate the presence of overdispersion, which would violate the Poisson assumption of equal mean and variance. We also examined the deviance to assess the model's goodness-of-fit, with a value near the degrees of freedom indicating a well-fitting model. Additionally, we conducted a residual analysis, including plotting residuals against fitted values and creating Q-Q plots of standardized residuals, to detect any systematic patterns that might indicate model misspecification. Together, these diagnostics help validate our model and confirm the reliability of our conclusions.

Dispersion

In a Poisson generalized linear model (GLM), the dispersion parameter is assumed to be fixed at 1. This assumption is crucial because the Poisson distribution is characterized by its mean being equal to its variance, a property known as *equidispersion*. When the observed variance is greater than the mean, the data exhibit over-dispersion. This is more common in practice and can arise from various sources, such as unobserved heterogeneity among observations,

excess zeros, or violations of the Poisson model's assumptions (such as events not occurring independently).

Then, the hypothesis being tested relates to the dispersion of the data with a significance level 5% is stated below:

- Null Hypothesis (H0):

The null hypothesis posits that the true dispersion parameter equals 1. This means the data follow a Poisson distribution accurately, where the mean and variance of the count data are equal (equidispersion). The model is adequately specified, and there's no extra variability in the data beyond what the Poisson model accounts for.

- Alternative Hypothesis (Ha):

The alternative hypothesis suggests that the true dispersion parameter is greater than 1, indicating over-dispersion in the data. Overdispersion occurs when the observed variance in the count data is greater than what the Poisson model would predict based on the mean.

Overdispersion test

```
data: glm_poisson
z = 14.143, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
2.584389
```

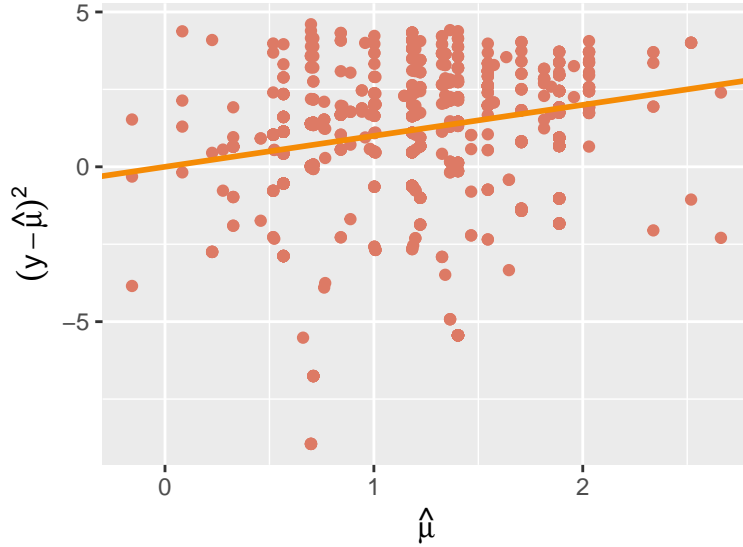


Figure 7: Analysis of Overdispersion in Poisson Regression Model

Interpretation:

- p-value is extremely small, below the significance level, which indicates that the result is highly statistically significant. Hence, this supports the alternative hypothesis that the true dispersion is greater than 1.
- The sample estimated - dispersion is 2.584389 which is substantially greater than 1, indicating over-dispersion in the data.
- Meanwhile, a common way to assess dispersion in a Poisson model is through a plot of the residuals which is displayed in Figure 7. It suggests that as the predicted values increase, the variance of the residuals also increases, which is a classic sign of over-dispersion. Over-dispersion is when the variance is greater than the mean, which often occurs with count data.
- Given this evidence of over-dispersion, it would be prudent to consider alternative models that can accommodate the extra variability, such as a Negative Binomial regression model which will be conducted later.

Deviance:

Deviance is used to quantify the difference between a fitted model and a perfect model (a saturated model that fits the data exactly). The deviance essentially quantifies the discrepancy between the observed data and the values predicted by the model under the assumption that the model is correct. A lower deviance indicates a better fit of the model to the data.

Hence, based on the generated model result, we got:

- Null Deviance: this represents the goodness of fit of a model that includes only the intercept (no predictors). It is 5122.4 on 1449 degrees of freedom.
- Residual Deviance: This is the goodness of fit of the model that includes predictors (AnimalType, IntakeType, OutcomeType and ChipStatus). It is 3759.4 on 1440 degrees of freedom.

The residual deviance is used to assess the fit of the model to the data. For a well-fitting model, the residual deviance should be close to the degrees of freedom (relatively low). Here, the residual deviance (3759.4) is quite high compared to the Chi-square distribution with the 1440 - degrees of freedom, which might indicate that the model does not fit the data perfectly. This could be a sign of over-dispersion or that the model is missing some key explanatory variables.

Residuals Analysis

- Plotting residuals vs. fitted values

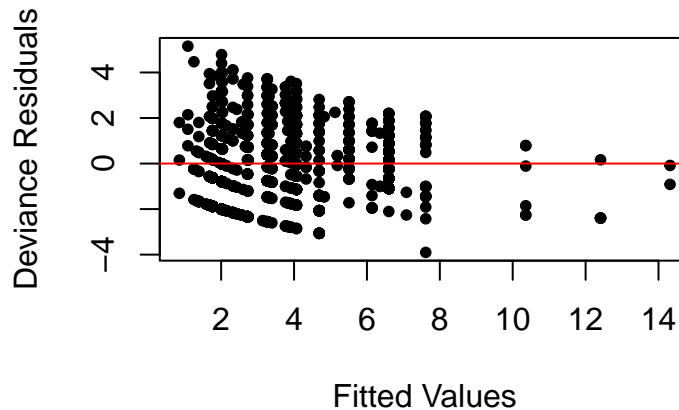


Figure 8: Residuals vs Fitted Values

Based on the plot Figure 8 The residuals plot here shows that as the fitted values increase, the spread of the residuals also increases. The increase in the spread of residuals as the fitted values increase suggests that the variance of the residuals is not constant. This pattern indicates potential over-dispersion in the data where the variance exceeds the mean.

- Scale location plot (spread vs. level plot)

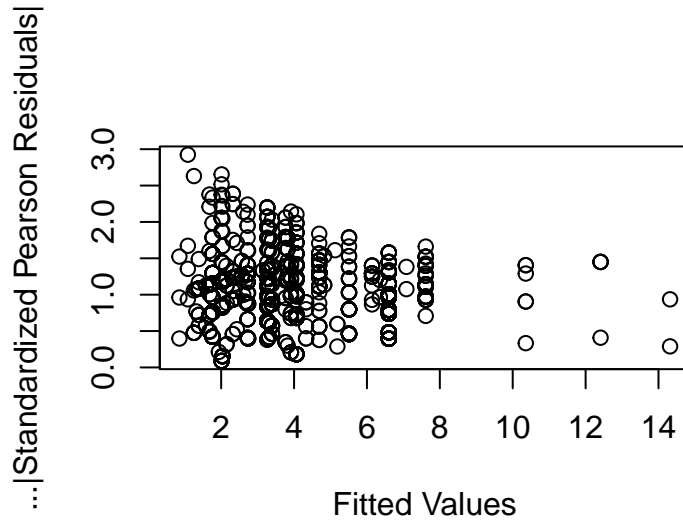


Figure 9: Scale-Location

The plot Figure 9 represents a scatter plot of the square root of the absolute standardized Pearson residuals versus the fitted values from a Generalized Linear Model (GLM). It will give the information about the move of the variance across different levels of the mean, making it easier to see whether there is a consistent spread of residuals across all counts or whether the spread increases with the count (which would indicate over-dispersion).

This pattern suggests possible over-dispersion in the data and will be advisable to consider model alternatives like the negative binomial regression.

- Residual vs. leverage

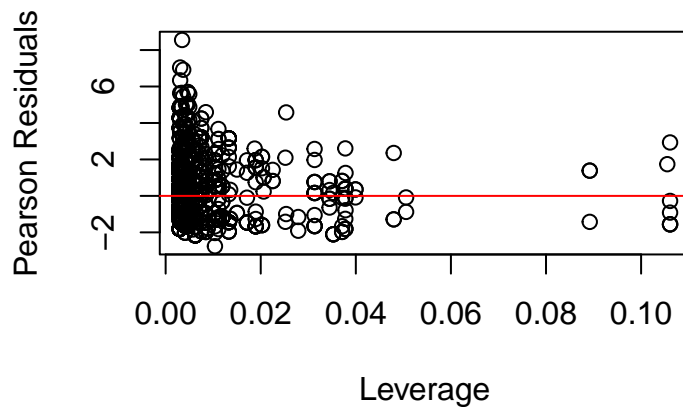


Figure 10: Residuals vs Leverage

Most of the data points cluster at the left side of the plot, suggesting that these observations have lower leverage and smaller residuals, shown by the plot Figure 10. Then, they don't have an undue influence on the model which seems the model fits well for the majority of the data.

However, the presence of points with high residuals and high leverage in the right side suggests there are some exceptions where the model's fit is not as good. Hence, it would be that while the model appears to provide a good fit for most of the data, there are specific observations that require further investigation to ensure the model's robustness and to potentially improve its accuracy.

When diagnostic plots provide different insights, it is crucial to consider the overall evidence and the specific research context. If over-dispersion is a consistent concern across multiple diagnostics (like the Residuals vs. Fitted Values Plot and the Scale-Location Plot), it often outweighs indications from other plots that the model might be adequate for most data points. Therefore, even if the Residuals vs. Leverage Plot suggests the model is mostly fitting well, the evidence of over-dispersion from the other plots is a strong indicator that the Poisson model might not be the best choice. Exploring alternative models like the Negative Binomial, which can accommodate over-dispersion, is likely a prudent step to improve model fit and ensure more reliable inference from the model.

4.1.2 Negative Binomial Model

It will be conducted the Negative Binomial Model which is a good alternative model when the assumption of equidispersion (mean equals variance) in Poisson regression doesn't hold. The result was shown by:

$$\begin{aligned} \log(\text{Expected Count of Time at Shelter}) = & 2.58882 \\ & + 0.22863 \times \text{TypeDog} \\ & - 1.23810 \times \text{IntakeOwnerSurrender} \\ & - 0.75541 \times \text{IntakeStray} \\ & - 0.99159 \times \text{OutcomeDied} \\ & - 0.75098 \times \text{OutcomeEuthanized} \\ & - 0.67674 \times \text{OutcomeFoster} \\ & - 1.39817 \times \text{OutcomeReturnedToOwner} \\ & - 0.15081 \times \text{ChipScanNoChip} \\ & - 0.40916 \times \text{ChipUnableToScan} \end{aligned}$$

where,

- TypeDog is the indicator variable for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category TypeCat.
- IntakeOwnerSurrender and IntakeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeConfiscated.
- OutcomeDied, OutcomeEuthanized, OutcomeFoster, OutcomeReturnedToOwner are indicator variables for the outcome, taking the value 1 if the condition is true and 0 otherwise with the baseline category OutcomeAdoption.
- ChipScanNoChip and ChipUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipScanChip.

Derived from the model, all variables are significant predictors of the time an animal spends in the shelter.

4.1.2.1 Model Diagnostics and Assumptions Checking

To ensure the reliability and validity of our Negative Binomial regression model findings, it is imperative to rigorously examine the underlying assumptions through a series of diagnostic checks.

Over-dispersion Check

Attained from the model summary, we can draw the result that the parameter Theta: 2.189 along with a relatively small standard error Std.Err: 0.16 which indicates that the model has identified and is accounting for over-dispersion in the data. It suggest that the Negative Binomial model is a suitable choice for the data, given the presence of over-dispersion as a positive outcome in terms of assumptions checking for the model.

Residual Deviance

Deviance in GLMs is a measure of the goodness-of-fit of a model. It is based on the likelihood function and compares two models between the fitted model and a reference model. The lower the deviance, the closer the fitted model is to the reference model in terms of likelihood. Comparing the residual deviance to the degrees of freedom helps assess if the fitted model is adequately fitting the data without overfitting. If the residual deviance is close to the degrees of freedom for the fitted model, it suggests that the model is adequately fitting the data.

Then, in the model summary, attained the Residual Deviance: 1795.2 with the 1440 degrees of freedom and the Null Deviance: 2316.9 with the 1449 degrees of freedom, this shows how much better the model fits the data compared to the model with only the intercept.

Residual Analysis

In evaluating the adequacy of our Negative Binomial regression model, a series of residual diagnostic plots were examined, including Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage, each offering insights into different aspects of model fit and underlying assumptions.

- Residuals vs. Fitted Values Plot

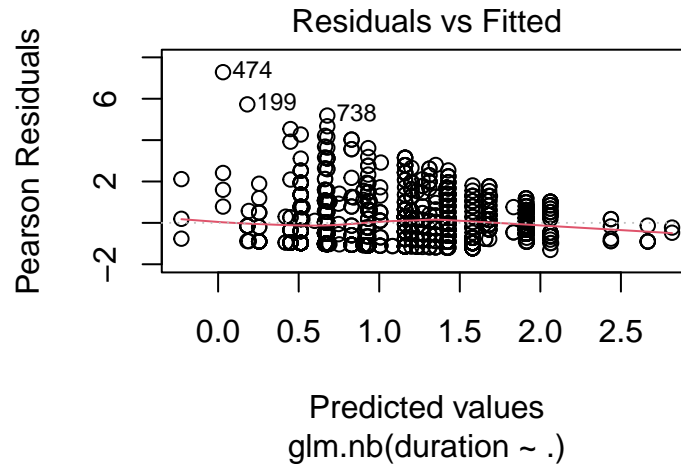


Figure 11: Residuals vs Fitted Values

Shown in the Figure 11, the residuals seem to increase slightly in variance with smaller fitted values, suggesting potential mild heteroscedasticity. However, since this is count data being modeled with a Negative Binomial regression, some of this is to be expected and the model accounts for it. There may be some potential outliers, but their influence would need further investigation, possibly with additional diagnostics like Cook's distance.

- Normal Q-Q Plot:

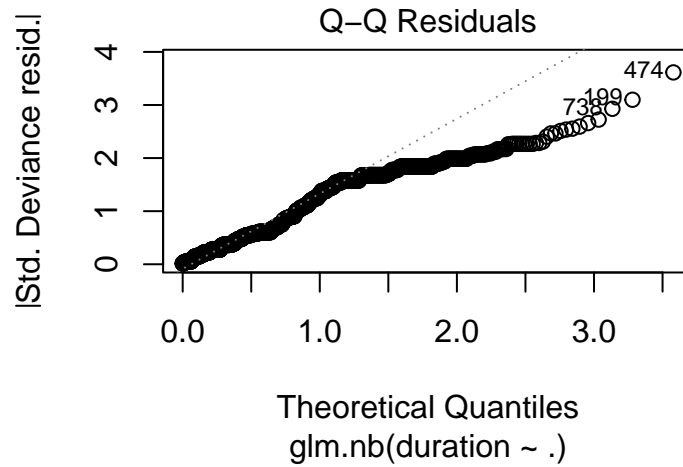


Figure 12: Normal Q-Q Plot

In the Q-Q plot Figure 12, the residuals deviate from the line at both ends, indicating that the residuals might not be following a normal distribution, which is a common issue for count data and is often the reason for using a Negative Binomial model instead of Poisson.

- Scale-Location Plot

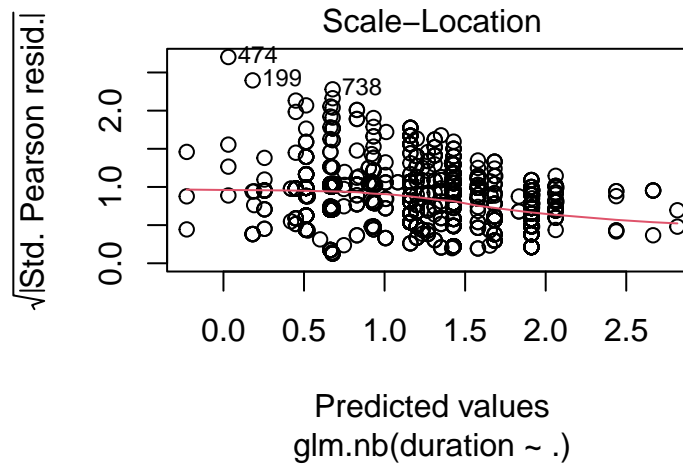


Figure 13: Scale-Location

The red line, captured on the Figure 13, shows some fluctuation but does not exhibit a clear or strong trend which indicates potential mild heteroscedasticity.

- Residuals vs. Leverage

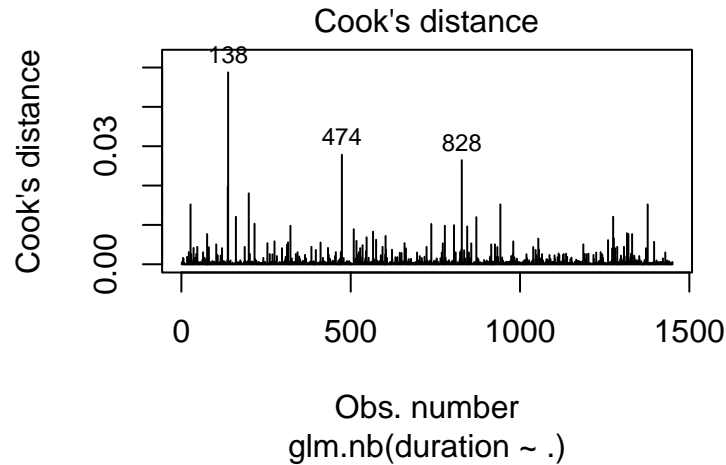


Figure 14: Residuals vs Leverage

Observations with higher Cook's distance values are worth examining because they might be outliers or influential data points that could unduly affect the model's coefficients, predictions, and overall fit. Based on the Figure 14, most observations have a Cook's distance close to zero, indicating they have little influence on the regression model. However, there are a few observations, labeled with their observation numbers, that stand out with higher Cook's distance values yet not necessarily a problem for the fitting model.

Summary of Model Diagnostic

In conclusion, the model appears to perform well in terms of fitting the central tendency and dispersion of the data, as suggested by the lack of systematic patterns in residuals and the appropriate handling of over-dispersion.

4.2 Model Selection

We employ a backward elimination process guided by the Akaike Information Criterion (AIC) to identify the most parsimonious model that adequately explains the variation in the time animals spend at the shelter. This approach systematically removes the least significant predictors from the full model, optimizing the balance between model complexity and fit to the data.

```
Start:  AIC=7440.01
duration ~ type + intake + outcome + chip
```

	Df	Deviance	AIC
<none>		3759.4	7440.0
- chip	2	3785.3	7461.9

```

- type      1    3786.3 7464.8
- intake    2    4167.9 7844.5
- outcome   4    4944.4 8617.0

```

Start: AIC=6756.64

```
duration ~ type + intake + outcome + chip
```

	Df	AIC
<none>		6756.6
- chip	2	6763.7
- type	1	6769.3
- intake	2	6902.7
- outcome	4	7128.5

The output of our backward stepwise selection process, as implemented via the stepAIC function, has been instrumental in guiding us toward the most parsimonious model to explain the duration of stays in animal shelters. This rigorous procedure evaluated both Poisson and Negative Binomial distributions, with the resulting AIC values serving as our yardstick for model comparison. Notably, the stepAIC procedure indicated that the initial model, which garnered an AIC of 6756.64, is the most efficient configuration, as further removal of predictors did not yield a lower AIC value. This finding aligns with the principle of parsimony, which underscores the importance of model simplicity without sacrificing explanatory power.

Upon comparing the AIC of the refined Negative Binomial model against the Poisson model, the former, with a substantially lower AIC, clearly emerged as the superior fit. The Negative Binomial model's AIC of 6756.64 stands in stark contrast to the Poisson model's AIC of 12216.73, reinforcing the suitability of the Negative Binomial distribution for our count data, which exhibits signs of overdispersion. This compelling evidence has led us to select the Negative Binomial model for in-depth analysis, ensuring that our investigative framework is not only robust and statistically sound but also retains clarity and interpretability for drawing meaningful conclusions.

5 Conclusion

Our investigation into the Dallas Animal Shelter Dataset has identified critical factors affecting the duration of animals' shelter stays. Through rigorous application of Generalized Linear Models, we have navigated the complex landscape of animal shelter data, addressing both its strengths and limitations.

Key Findings:

- The Negative Binomial Model, selected for its fit and parsimony, outperformed the Poisson Regression Model by effectively handling over-dispersion in the data.
- Significant predictors such as animal type, intake reasons, outcomes, and microchip status have been quantified, offering insights into their roles in influencing shelter stay durations.
- Interventions targeting these factors could lead to improvements in shelter operations and animal welfare, particularly promoting microchipping and optimizing intake and adoption processes.

Implications for Practice: Our findings suggest actionable strategies for shelters, including targeted campaigns for specific animal types and enhancing microchipping practices to facilitate faster reunification or adoption, ultimately aiming to shorten shelter stays and improve outcomes for animals.

In light of these insights, shelters can better strategize resource allocation and management practices, potentially leading to more effective and compassionate outcomes for the animals in their care. While our model provides a strong foundation, we acknowledge that further research could incorporate additional nuanced data to enhance predictive capabilities and provide even deeper insights into shelter dynamics.