Group 20

Combine with the the Explanatory Data Analysis File, this section is the next one.

Formal Data Analysis

In this section, several formal statistical models will be conducted using a Generalized Model Analysis to infer the relationships between variables.

A. Model Fitting

This should be conducted on the EDA section, so after combined it, this should be removed.

A.1 Poisson Model

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

```
Call:
glm(formula = time_at_shelter ~ animal_type + month + year +
    intake_type + chip_status, family = "poisson", data = data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               578.124443  78.038304   7.408 1.28e-13 ***
animal_typeCAT              2.765699   1.000367   2.765  0.00570 **
animal_typeDOG              2.875463   1.000126   2.875  0.00404 **
animal_typeWILDLIFE         2.609189   1.008668   2.587  0.00969 **
```

```
month                        -0.030088   0.005061  -5.945 2.76e-09 ***
year                         -0.286813   0.038683  -7.414 1.22e-13 ***
intake_typeOWNER SURRENDER   -0.735566   0.040129 -18.330  < 2e-16 ***
intake_typeSTRAY             -0.517652   0.036797 -14.068  < 2e-16 ***
chip_statusSCAN NO CHIP       0.014802   0.027678   0.535  0.59278
chip_statusUNABLE TO SCAN    -0.472796   0.069586  -6.794 1.09e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10754  on 1464  degrees of freedom
Residual deviance: 10266  on 1455  degrees of freedom
AIC: 14316

Number of Fisher Scoring iterations: 5
```

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:

$$\begin{aligned}\log(\text{Expected Count of Time at Shelter}) = {} & 578.124443 \\ & + 2.765699 \times \text{AnimalTypeCat} \\ & + 2.875463 \times \text{AnimalTypeDog} \\ & + 2.609189 \times \text{AnimalTypeWildlife} \\ & - 0.030088 \times \text{Month} \\ & - 0.286813 \times \text{Year} \\ & - 0.735566 \times \text{IntakeTypeOwnerSurrender} \\ & - 0.517652 \times \text{IntakeTypeStray} \\ & - 0.014802 \times \text{ChipStatusScanNoChip} \\ & - 0.472796 \times \text{ChipStatusUnableToScan}\end{aligned}$$

where,

- AnimalTypeCat, AnimalTypeDog, AnimalTypeWildlife are indicator variables for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category AnimalTypeBird.
- Month and Year are numerical variables representing the month and year of intake.

- IntakeTypeOwnerSurrender and IntakeTypeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeTypeConfiscated.
- ChipStatusScanNoChip and ChipStatusUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipStatusScanChip.

Model Diagnostics and Assumptions Checking

The analysis utilized a Poisson regression model to investigate the impact of factors such as, including **`animal_type`**, **`month`**, **`year`**, **`intake_type`**, and **`chip_status`**, 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 *overdispersion*. 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 overdispersion 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 = 8.1017, p-value = 2.711e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  8.661377
```

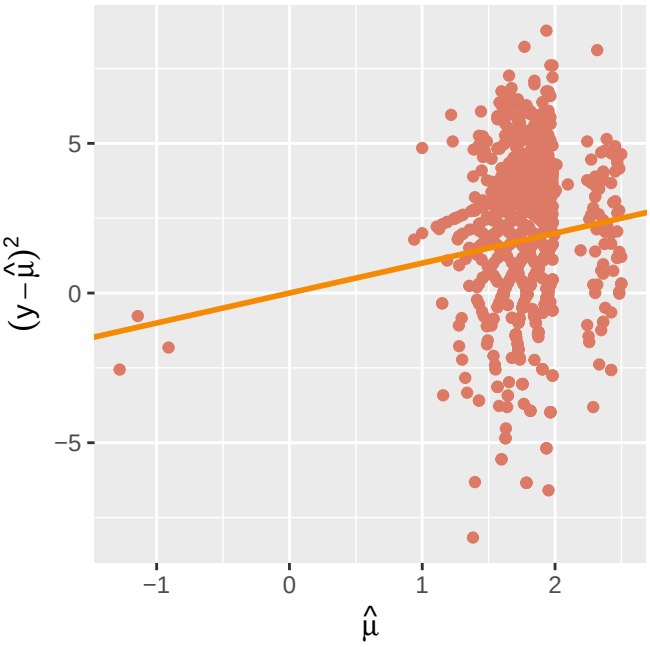


Figure 1: Residuals Plot for Assessing Dispersion in Poisson Regression

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 8.52551 which is substantially greater than 1, indicating a significant level of overdispersion 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 1. It suggests that as the predicted values increase, the variance of the residuals also increases (as indicated by the upward trend), which is a classic sign of overdispersion. Overdispersion is when the variance is greater than the mean, which often occurs with count data.
- Given this evidence of overdispersion, 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 10754 on 1464 degrees of freedom.
- Residual Deviance: This is the goodness of fit of the model that includes predictors (`animal_type`, `month`, `year`, `intake_type`, `chip_status`). It is 10266 on 1455 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 (10266) is quite high compared to the Chi-square distribution with the 1455 - degrees of freedom, which might indicate that the model does not fit the data perfectly. This could be a sign of overdispersion or that the model is missing some key explanatory variables.

Residuals Analysis

- Plotting residuals vs. fitted values

  Based on the plot Figure 2 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 overdispersion in the data where the variance exceeds the mean.

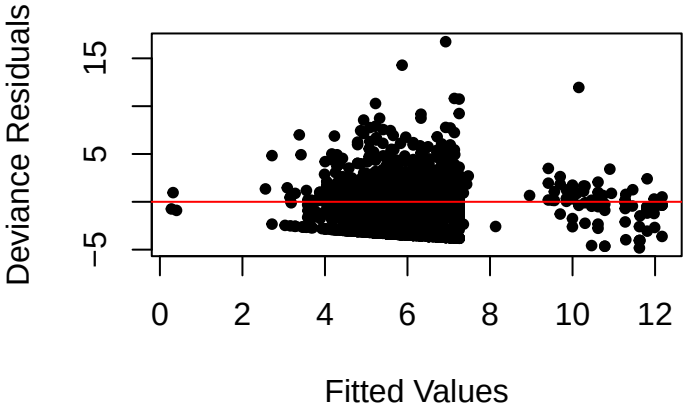


Figure 2: Residuals vs. Fitted Values Plot for Poissron Regression Model

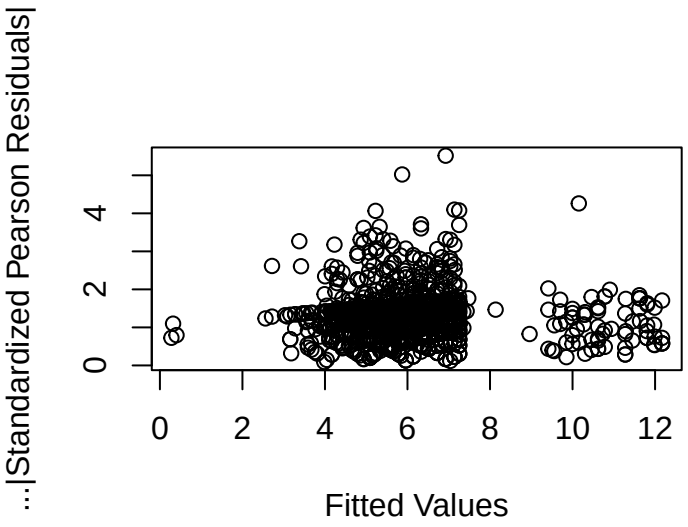


Figure 3: Scale-Location Plot (Spread vs. Level Plot)

- Scale location plot (spread vs. level plot)

  The plot Figure 3 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 overdispersion). Meanwhile, we can see that there is a concentration of residuals around the lower fitted values, with residuals spreading out as the fitted values increase. It reveals several points that stand out from the main cluster, particularly at the mid-range of fitted values, indicating potential outliers or influential observations. This pattern suggests possible overdispersion in the data and will be advisable to consider model alternatives like the negative binomial regression.

- Residual vs. leverage

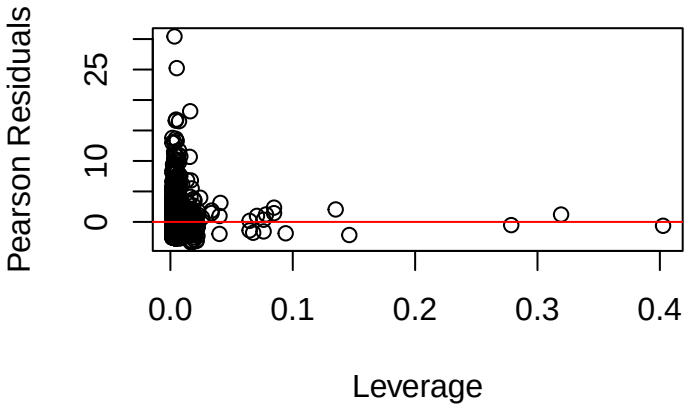


Figure 4: Residuals vs. Leverage Plot

  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 4. 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, high leverage, or both 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 overdispersion 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 overdispersion 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 overdispersion, is likely a prudent step to improve model fit and ensure more reliable inference from the model.

A.2 Negative Binomial Model

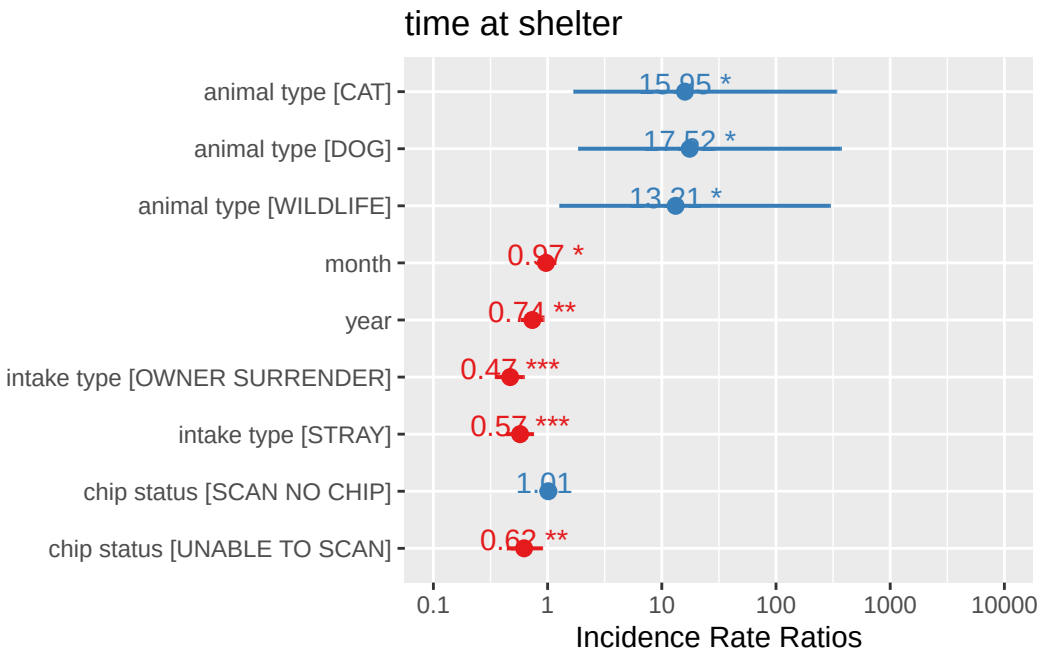
It will be conducted the Negative Binomial Model which is a good alternative model when the assumption of equidispersion (mean equals variance) in Poisson regression doesn't hold. The result was shown by:

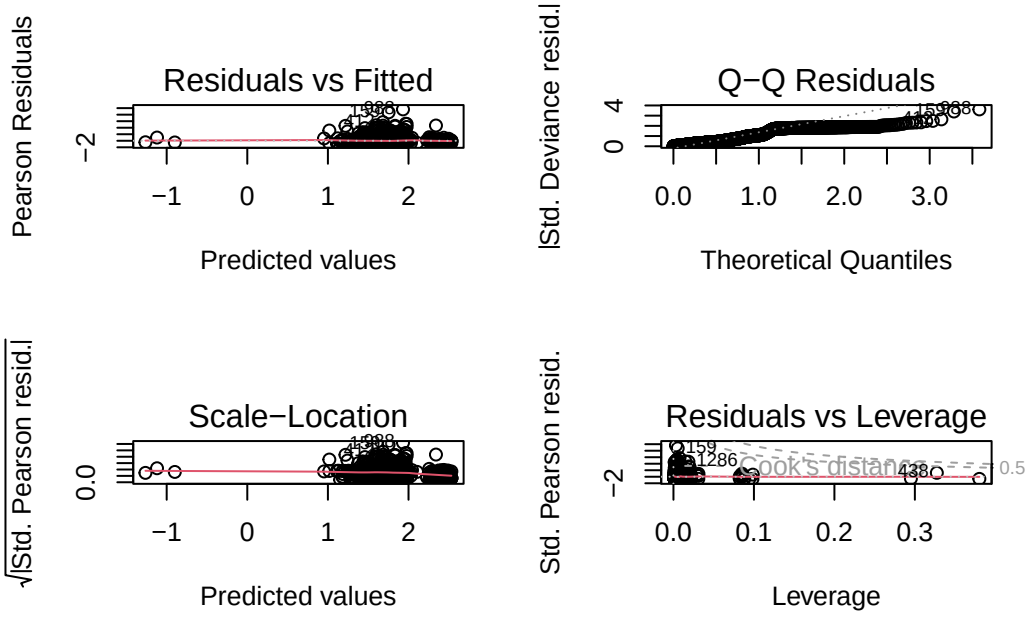
```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + month + year +
    intake_type + chip_status, data = data, init.theta = 0.7625369942,
    link = log)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                620.40667  232.62169   2.667  0.00765 **
animal_typeCAT               2.76946    1.19772   2.312  0.02076 *
animal_typeDOG               2.86336    1.19617   2.394  0.01668 *
animal_typeWILDLIFE          2.58130    1.24807   2.068  0.03862 *
month                       -0.03283    0.01507  -2.178  0.02938 *
year                        -0.30775    0.11532  -2.669  0.00761 **
intake_typeOWNER SURRENDER  -0.75490    0.14319  -5.272 1.35e-07 ***
intake_typeSTRAY            -0.55483    0.13728  -4.042 5.31e-05 ***
chip_statusSCAN NO CHIP      0.01468    0.08355   0.176  0.86054
chip_statusUNABLE TO SCAN   -0.47382    0.18025  -2.629  0.00857 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7625) family taken to be 1)

    Null deviance: 1743.0  on 1464  degrees of freedom
Residual deviance: 1687.1  on 1455  degrees of freedom
AIC: 8352.4

Number of Fisher Scoring iterations: 1

              Theta:  0.7625
          Std. Err.:  0.0348

 2 x log-likelihood:  -8330.4070
```

$$
\begin{aligned}
\log(\text{Expected Count of Time at Shelter}) = 620.40667 & \\
& + 2.76946 \times \text{AnimalTypeCat} \\
& + 2.86336 \times \text{AnimalTypeDog} \\
& + 2.58130 \times \text{AnimalTypeWildLife} \\
& - 0.03283 \times \text{Month} \\
& - 0.30775 \times \text{Year} \\
& - 0.75490 \times \text{IntakeTypeOwnerSurrender} \\
& - 0.55483 \times \text{IntakeTypeOwnerStrayStary} \\
& + 0.01468 \times \text{ChipStatusScanNoChip} \\
& - 0.47382 \times \text{ChipStatusUnableToScan}
\end{aligned}
$$

where,

- AnimalTypeCat, AnimalTypeDog, AnimalTypeWildlife are indicator variables for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category AnimalTypeBird.
- Month and Year are numerical variables representing the month and year of intake.
- IntakeTypeOwnerSurrender and IntakeTypeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeTypeConfiscated.
- ChipStatusScanNoChip and ChipStatusUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipStatusScanChip.






```
[1] 14316.41
```

```
[1] 8352.407
```

```
Single term deletions
```

```
Model:
time_at_shelter ~ animal_type + month + year + intake_type +
    chip_status
            Df Deviance    AIC F value    Pr(>F)
<none>           1687.1 8350.4
animal_type  3   1694.8 8352.1  2.2228   0.08373 .
month        1   1691.8 8353.1  4.0458   0.04446 *
year         1   1694.3 8355.6  6.1796   0.01303 *
intake_type  2   1718.9 8378.2 13.7054 1.268e-06 ***
chip_status  2   1694.8 8354.1  3.3072   0.03689 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

B. Model Selection

C. Model Interpretation and Conclusion