Analysis of Time Spent in Shelter

Group 20

```r
library(MASS)
library(stats)
library(ggplot2)
```

```r
df <- read.csv("dataset20.csv")

df$month <- as.numeric(df$month)
df$season <- cut(df$month, breaks = c(2, 5, 8, 11, 12),
                 labels = c('Spring', 'Summer', 'Autumn', 'Winter'),
                 include.lowest = TRUE)
df$season[df$month %in% c(12, 1, 2)] <- 'Winter'
df$season <- factor(df$season, levels = c("Spring", "Summer", "Autumn", "Winter"))
```

Spring is the breeding season for many animals, which may lead to a surge in the number of stray animals, putting shelters under immense pressure. As resources become strained and workloads increase, staff efficiency may decline, resulting in longer stays for animals in shelters.

Summer is a peak travel season, and as people are away from home, the demand for pet adoption decreases. Additionally, the hot and humid environment increases the risk of diseases, further prolonging the stay of animals in shelters.

In winter, the holiday season, including Christmas and New Year, may lead to an adoption surge, reducing the time animals spend in shelters.

Given the impact of these seasonal factors on animal sheltering and adoption trends, we have decided to incorporate seasonal effects as an explanatory variable

```r
full_model <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status + season + ye
```

Outcome_type is an outcome variable rather than a factor influencing the length of stay in the shelter, and therefore should not be used as an explanatory variable. In this study, we use animal_type, chip_status, intake_type, season, and year as explanatory variables, while time_at_shelter serves as the outcome variable to construct a negative binomial regression model.

```
selected_model1 <- step(full_model, direction = "backward", trace = TRUE)
```

```
Start:  AIC=8356.35
time_at_shelter ~ animal_type + intake_type + chip_status + season + 
    year

              Df Deviance    AIC
- season       3   1689.8 8353.1
<none>             1687.0 8356.4
- year         1   1689.1 8356.4
- animal_type  3   1695.1 8358.4
- chip_status  2   1694.5 8359.8
- intake_type  2   1718.7 8384.1

Step:  AIC=8353.09
time_at_shelter ~ animal_type + intake_type + chip_status + year

              Df Deviance    AIC
<none>             1686.8 8353.1
- year         1   1689.3 8353.6
- animal_type  3   1695.4 8355.7
- chip_status  2   1693.8 8356.1
- intake_type  2   1717.7 8380.0
```

To select the most appropriate explanatory variables, we employ a backward stepwise regression approach using the Akaike Information Criterion (AIC) as the evaluation standard. Specifically, we start with a full model that includes all candidate explanatory variables and iteratively remove variables with lower contributions to the model until we identify the optimal model with the lowest AIC. The results indicate that removing the season variable leads to a lower AIC value for the model. Therefore, we exclude the season variable to improve model fit.

```
summary(selected_model1)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type + 
    chip_status + year, data = df, init.theta = 0.7592685117,
    link = log)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                243.35450  154.43465   1.576   0.1151
animal_typeCAT               2.72164    1.20350   2.261   0.0237 *
animal_typeDOG               2.84387    1.20218   2.366   0.0180 *
animal_typeWILDLIFE          2.54467    1.25404   2.029   0.0424 *
intake_typeOWNER SURRENDER  -0.74503    0.14350  -5.192 2.08e-07 ***
intake_typeSTRAY            -0.54367    0.13757  -3.952 7.75e-05 ***
chip_statusSCAN NO CHIP      0.01149    0.08369   0.137   0.8908
chip_statusUNABLE TO SCAN   -0.45268    0.18023  -2.512   0.0120 *
year                        -0.12089    0.07658  -1.579   0.1144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7593) family taken to be 1)

    Null deviance: 1737.8  on 1464  degrees of freedom
Residual deviance: 1686.8  on 1456  degrees of freedom
AIC: 8355.1

Number of Fisher Scoring iterations: 1

              Theta:  0.7593
          Std. Err.:  0.0346

 2 x log-likelihood:  -8335.0890
```

After removing the season variable, we reconstructed the negative binomial regression model. The results indicate that at a 95% confidence level, the p-value of the year variable is greater than 0.05, suggesting that it is not statistically significant. Additionally, based on the AIC evaluation, the model with the year variable has an AIC of 8353.1, while the model without it has an AIC of 8353.6, indicating that removing the year variable has a minimal impact on the model. Therefore, we decide to exclude the year variable from the explanatory variables.

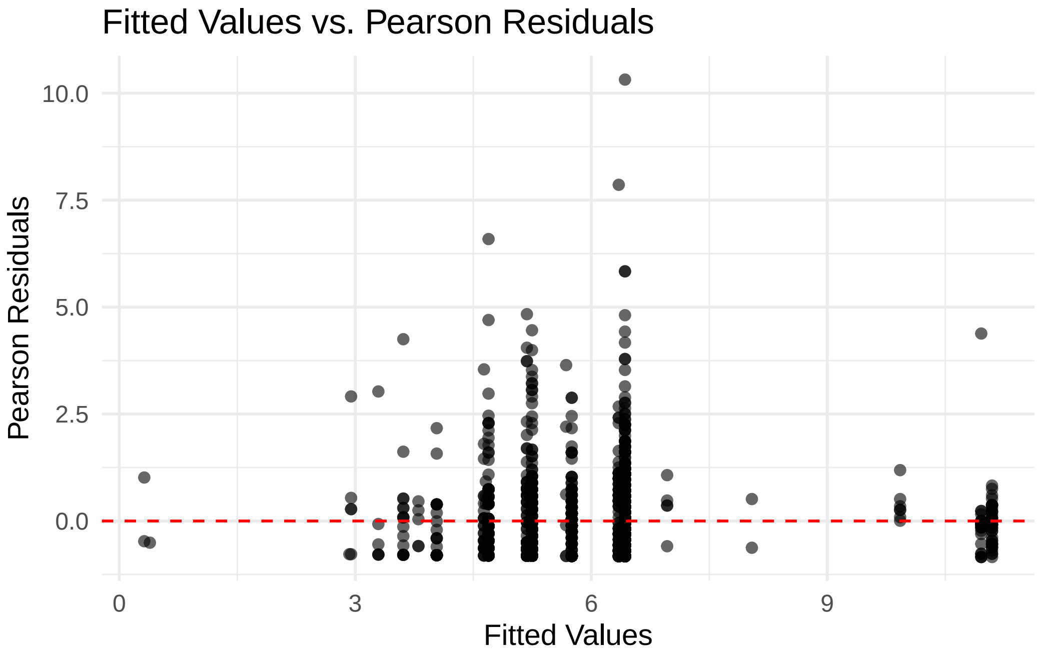
```
selected_model2 <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status,  data
summary(selected_model2)
```

```
Call:
```

```
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status, data = df, init.theta = 0.7575663959, link = log)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.41057    1.20172  -0.342   0.7326
animal_typeCAT              2.69334    1.19365   2.256   0.0240 *
animal_typeDOG              2.80462    1.19231   2.352   0.0187 *
animal_typeWILDLIFE         2.48278    1.24462   1.995   0.0461 *
intake_typeOWNER SURRENDER -0.74886    0.14364  -5.213 1.85e-07 ***
intake_typeSTRAY           -0.54570    0.13770  -3.963 7.40e-05 ***
chip_statusSCAN NO CHIP     0.01240    0.08376   0.148   0.8823
chip_statusUNABLE TO SCAN  -0.45322    0.18040  -2.512   0.0120 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7576) family taken to be 1)

    Null deviance: 1735.1  on 1464  degrees of freedom
Residual deviance: 1686.7  on 1457  degrees of freedom
AIC: 8355.6

Number of Fisher Scoring iterations: 1

              Theta:  0.7576
          Std. Err.:  0.0345

 2 x log-likelihood:  -8337.5990
```

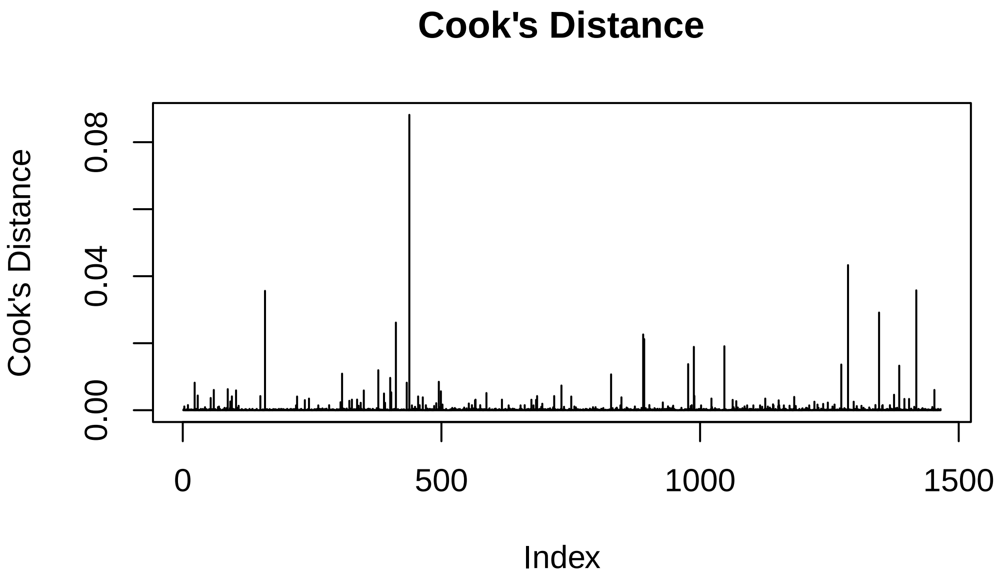
Finally, we select animal_type, chip_status, and intake_type as explanatory variables, with time_at_shelter as the outcome variable to construct a negative binomial regression model.

```r
df$residuals <- residuals(selected_model2, type = "pearson")
df$fitted_values <- fitted(selected_model2)
ggplot(df, aes(x = fitted_values, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Fitted Values vs. Pearson Residuals",
       x = "Fitted Values",
       y = "Pearson Residuals")
```



The plot shows some points with Pearson Residuals > 6, indicating the possible presence of outliers in the data. These outliers may have a significant impact on the model, requiring further investigation to determine whether any adjustments or modifications to the model are necessary.

```r
plot(cooks.distance(selected_model2), type="h",
     main="Cook's Distance", ylab="Cook's Distance")
abline(h = 1, col = "red", lty = 2)
```



The highest Cook's Distance in the plot appears to be less than 0.1, which is far below 1, indicating that there are no particularly severe high-influence points. However, some points still exhibit a relatively large influence. Therefore, in the subsequent steps, we will identify these high-influence points and attempt to remove the outliers before refitting the model to assess their impact.

```
df_cleaned <- df[-which(cooks.distance(selected_model2) > 4 / nrow(df)), ]
final_model <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status, data = df_
summary(final_model)
```

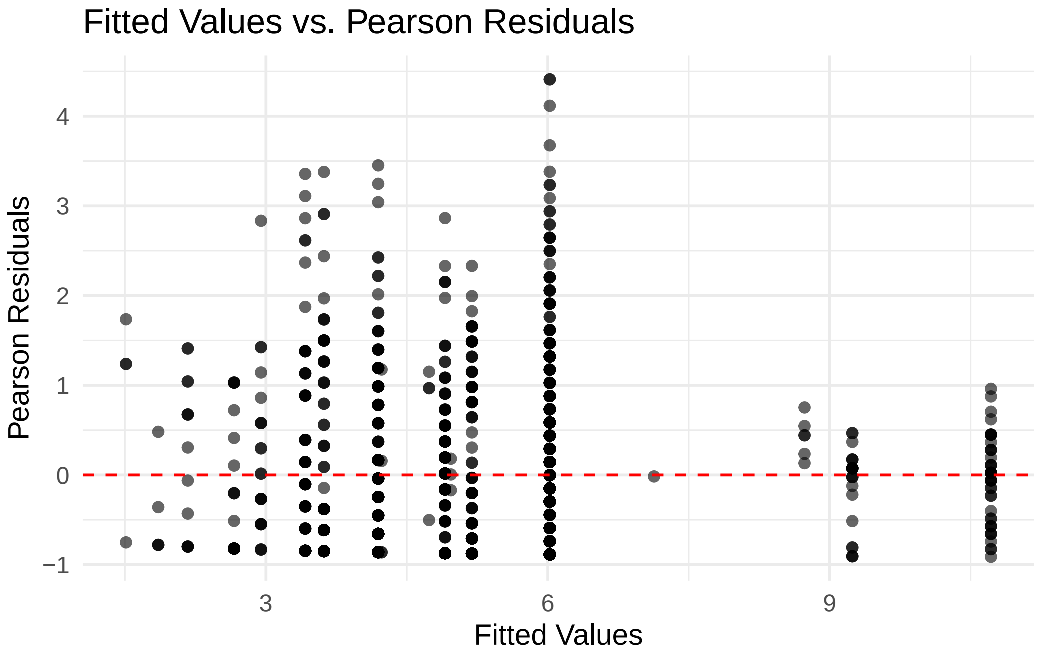
```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status, data = df_cleaned, init.theta = 0.9022135578,
    link = log)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.01893    0.15799  12.778  < 2e-16 ***
animal_typeDOG              0.20473    0.07799   2.625 0.008659 **
animal_typeWILDLIFE         0.37388    0.57373   0.652 0.514620
intake_typeOWNER SURRENDER -0.93785    0.13783  -6.804 1.01e-11 ***
intake_typeSTRAY           -0.57653    0.13135  -4.389 1.14e-05 ***
```

```
chip_statusSCAN NO CHIP     0.14810    0.08121   1.824 0.068190 .
chip_statusUNABLE TO SCAN  -0.66838    0.18756  -3.564 0.000366 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9022) family taken to be 1)

    Null deviance: 1728.7  on 1404  degrees of freedom
Residual deviance: 1636.9  on 1398  degrees of freedom
AIC: 7642.4

Number of Fisher Scoring iterations: 1

              Theta:  0.9022
          Std. Err.:  0.0454

 2 x log-likelihood:  -7626.4060
```

The results show that after removing the outliers, the refitted model has an AIC of 7642.4, which is 695.199 lower than the previous model's AIC, indicating an improvement in model fit.

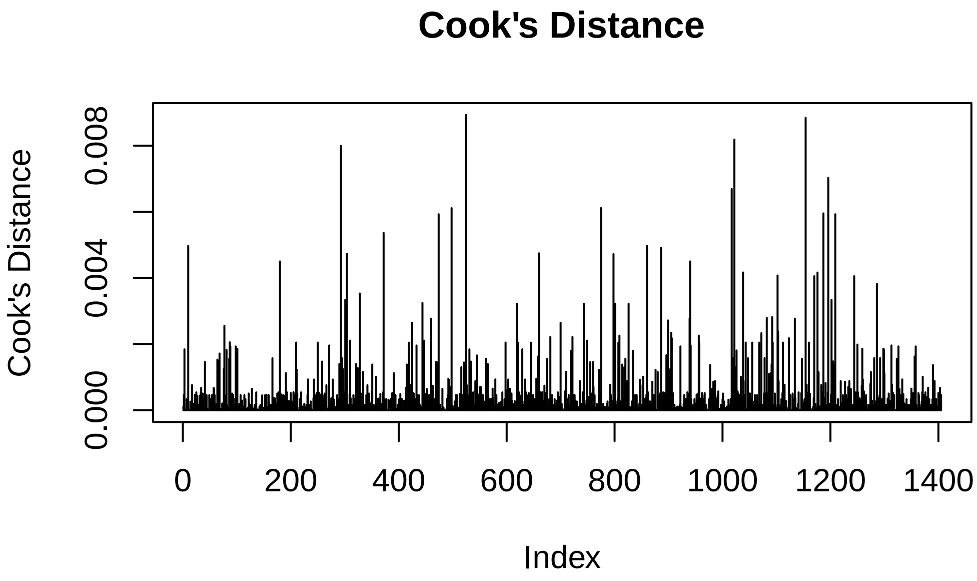
coefficient

```r
df_cleaned$residuals <- residuals(final_model, type = "pearson")
df_cleaned$fitted_values <- fitted(final_model)
ggplot(df_cleaned, aes(x = fitted_values, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Fitted Values vs. Pearson Residuals",
       x = "Fitted Values",
       y = "Pearson Residuals")
```



The Pearson residuals are distributed around 0 without any distinct U-shaped or V-shaped patterns, indicating that the model does not suffer from severe systematic bias and has a good overall fit.

```
plot(cooks.distance(final_model), type="h",
     main="Cook's Distance", ylab="Cook's Distance")
abline(h = 1, col = "red", lty = 2)
```



Cook’s Distance values are generally low, indicating that no single data point has an excessively large influence on the model, ensuring stable model fitting.

```
deviance(final_model) / df.residual(final_model)
```

```
[1] 1.17089
```

The calculated value of 1.17089 is slightly greater than 1, indicating a mild degree of overdispersion in the data. However, overall, the model fits well and can still accurately describe the data distribution.