Mini Project 401

36668259

Findings

The data set has 6 covariates (x_1,x_2,..., x_6) and a time variable and the response y.

Using a Poisson GLM is intuitive here because the data lacks binary outcomes, ruling out binomial regression. Additionally, the the time variable (t) in the data is the time taken for an event to occur (hours). These properties align with the assumptions of a Poisson model. While a Negative Binomial (NB) model could also be considered, its use would depend on evidence of over dispersion or over fitting in the Poisson model [1].

Poisson GLM

In Poisson regression, y represents event counts within varying time windows, requiring an offset to ensure accurate relationships between covariates and the response. An offset was used to account for variability in the time periods over which observations were taken. This makes sure the differences in y are not due to longer observations but reflections from the covariates. Three GLM models were created (all with offset of $\ln(t)$ (because the log link poisson regression is in use)): "with27" (including an outlier), "no27" (excluding the outlier), and "no.offset.no27" (excluding the outlier and offset). An extreme outlier ($y = 10,000$) in row 27 inflated the AIC to ~25,100. Removing it improved the model fit, reducing the AIC to ~726, and adding the offset further enhanced the fit, as shown by lower AIC and BIC scores.

```
c(BIC(with27),AIC(with27),BIC(no.offset.no27),AIC(no.offset.no27),
  BIC(no27),AIC(no27))
```

```
## [1] 37250.9544 37230.3567   749.2078   726.1287   740.2060   719.6912
```

Both AIC and BIC measure "badness of fit", with lower values indicating better models. AIC uses a lighter penalty and focuses on predictive accuracy, while BIC applies a stricter penalty based on sample size, favoring simpler models, especially with large datasets [2]. Their equations are:

$$\text{AIC} = 2p_* - 2\ell(\hat{\theta})$$

$$\text{BIC} = p_* \ln(n) - 2\ell(\hat{\theta})$$

$p_* =$ No. of parameters in the model, $\ell(\hat{\theta}) =$ the log-likelihood of the model and $n =$ the sample size. As stated above both AIC and BIC will penalise the model for complexity and this is addressed in AIC as $2p_*$ and in BIC as $p_* ln(n)$. Both log-likelihoods in each criterion measure how well the model fits. Therefore, the higher the log-likelihood the better the fit.

Cook's Distance (Fig 2) identifies influential points, with values above 1 [3] indicating high influence and values below $\frac{4}{n}$ [4] indicating minimal impact. In the Poisson GLM, an outlier caused extreme Cook's Distance values (above 10,000), but removing it reduced the range to 0 to 0.4. This explains the lower AIC and BIC scores post-removal. However, the fit remains sub optimal, as many points still exceed the $\frac{4}{n}$ threshold. To address this, each covariate will be examined for influence using backward fitting, followed by further analysis through residual diagnostics.

Backwards Fitting and Interactions

```
## [1] 1.231105e-04 8.093231e-07 7.291241e-18 2.145549e-28 3.980790e-31
## [6] 1.596946e-12
```

Backwards fitting is a process that starts with a model containing all covariates and removes the least significant covariates. After completing backwards fitting, all covariates are found to be statistically significant. Each p-value is small (well below 0.05), meaning that removing any covariate worsens the model fit. So, each variable meaningfully contributes to explaining y.

```
## [1] 1.231105e-04 8.093231e-07 7.291241e-18 2.145549e-28 3.980790e-31
## [6] 1.596946e-12
```

```
cat("Final 2 Forward Fitting p-vals :",
    c( anova(glm_x4_x5_x3_x6, glm_x4_x5_x3_x6_x1)$P[2],
       anova(glm_x4_x5_x3_x6, glm_x4_x5_x3_x6_x2)$P[2] ))
```

```
## Final 2 Forward Fitting p-vals : 0.002868024 1.716353e-05
```

All p-values are below 0.05, so all covariates are significant. This means no covariates will be removed from the "best model". Both backward and forward fitting confirm the same results, showing the model is robust, stable, and that each covariate has a meaningful relationship with the outcome. So, the "best model" has been identified, and x_1 interactions will now be tested.

```
ix1x3 <- glm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + offset(logt) + x1*x3,
             family = poisson(link = "log"), data = data_remove)

c(anova(no27, ix1x2)$P[2],anova(no27, ix1x3)$P[2],anova(no27, ix1x4)$P[2],
  anova(no27, ix1x5)$P[2],anova(no27, ix1x6)$P[2])
```

```
## [1] 0.09752534 0.38021034 0.48690646 0.24433489 0.86905491
```

The model ix1x3 tests the interaction between x_1 and x_3, with a p-value (~0.380) indicating no statistical significance at the 0.05 level. This result is consistent across all x_1 interactions, suggesting they do not improve the model or explain the variance in the response variable. Therefore, these terms donot improve the final model and so they will not be introduced within the "best model".

```
ix2x3 <- glm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + offset(logt) + x2*x3,
             family = poisson(link = "log"), data = data_remove)

c(anova(no27, ix2x3)$P[2])
```

```
## [1] 1.132588e-08
```

This report focuses on x_1 interactions however, upon looking into x_2 there are some interactions which are significant (p-val $<$ 0.05). But, for x_1 there aren't any significant interactions.

Residuals and Model Diagnostics

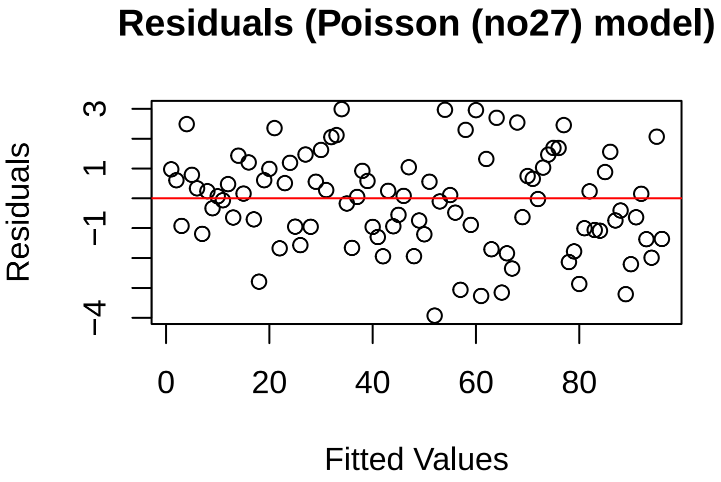

Figure 1: Residuals vs. fitted values, minor dispersion evident

Checking the "best model" with the NULL model (or "intercept" model, also without the outlier) is a good way to check if the model has improved with the additional covariates and as stated above the model definitely improved which is a baseline approach to show how useful the covariates are. The best model significantly improved from the intercept model to the model with the AIC score dropping from ~1223 to ~720.

```r
cat("Res Dev =",signif(summary(no27)$deviance,3),"with",summary(no27)$df.residual,"dof")
```

```
## Res Dev = 244 with 88 dof
```

A test one can compute from the best model is to see whether there is any over dispersion using the dispersion parameter [5] where the Residual Deviance is divided by the Degrees of Freedom to obtain the dispersion parameter. In this case it is ~2.77 which is greater than 1. This suggests there is some sort of over dispersion. Further to this, plotting the observed vs fitted (Fig 3) values showed some scatter around $y = x$ line. This means the poisson is a good fit showing that predictions align with the observed values. Even within this plot, there seemed to be over dispersion at around fourty to sixty fitted values.

Figure 1 shows that the model fits the fitted values well as there is some scatter around 0. There is some over dispersion as some points reach +3 and -4 however, generally the model seems to fit well. Furthermore, the residuals appear randomly scattered around zero, which suggests that the model does not show obvious signs of a pattern or trend, supporting the assumption that the Poisson model may be appropriate.

After plotting the leverages (Fig 4), observations 7, 29, 34, 61, 69, and especially 92 have high leverage, with 92 being the most impactful. High leverage points like 92 deviate from the average. Removing such points improves AIC and BIC scores. Finally, when plotting the 6 covariates as leverages (Fig 5) they show the same plot as the full leverage plot, suggesting that all six covariates contribute significantly to the model.

```r
c(BIC(no27no92), AIC(no27no92))
```

```
## [1] 734.1998 713.7688
```

Predictions (using no27)

Predicted values : $x_1 = $ m, $x_3 = 0.5$, $x_4 = 12$, $x_5 = 4$, $x_6 = 8$ with, $x_2 =$ "yearling" or "adult"

When using the log link function in a GLM, predictions represent the natural logarithm of the expected count. This is written as $\ln(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$. To interpret these predictions as expected counts, they must be exponentiated. This is why the predicted values are transformed. Using the predict.glm [6] function in R with the appropriate inputs fits the GLM and returns expected counts along with standard errors. Confidence intervals were calculated.

```r
# Display the prediction and confidence intervals
pred_a <- exp(preds_a$fit) ; CIlo <- exp(CIloeta) ; CIhi <- exp(CIhieta)

cat("The prediction for Adult :",c(pred_a, CIlo, CIhi), "\n")
```

```
## The prediction for Adult : 3.038041 2.816096 3.277477
```

```r
pred_y <- exp(preds_y$fit) ; CIlo_y <- exp(CIloeta_y) ;CIhi_y <- exp(CIhieta_y)

cat("The prediction for Yearling :", c(pred_y, CIlo_y, CIhi_y))
```

```
## The prediction for Yearling : 2.407602 2.217914 2.613512
```

Future Observations

To conlude, the poisson model was used to model the relationship between the response variable and the predictors. The model did create some good insights and fit relatively well, however some diagnostics and tests showed that there was some over dispersion. To address this, there could be some tests done between the poisson model and the NB model because NB regression includes an extra term [5] and this term accounts for over dispersion as seen below (lower AIC and BIC score).

```
## [1] 691.9836 668.9045
```

Finally, extra interactions could have been computed to create the absolute "best model".

Appendix

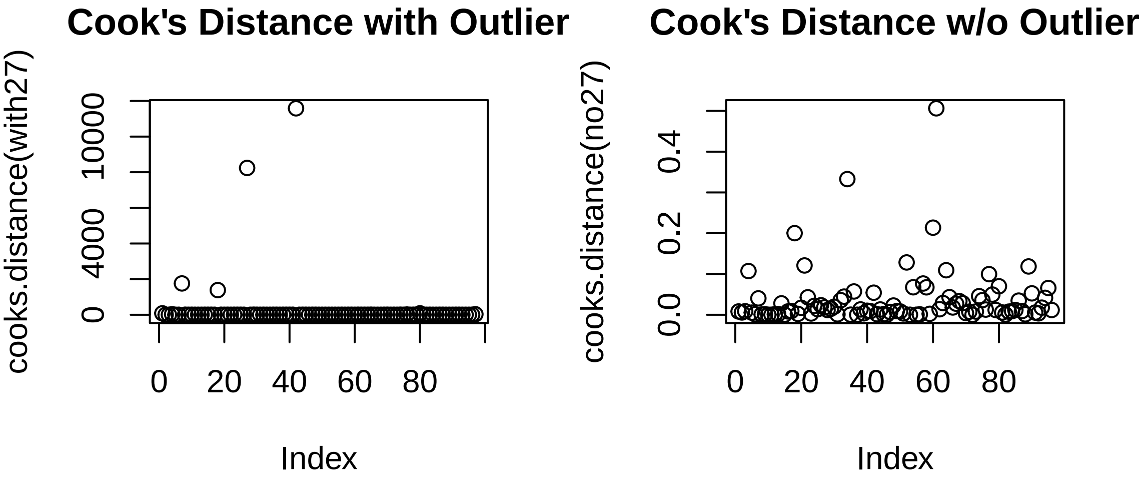

Figure 2: Cooks Distance with outlier vs no outlier.

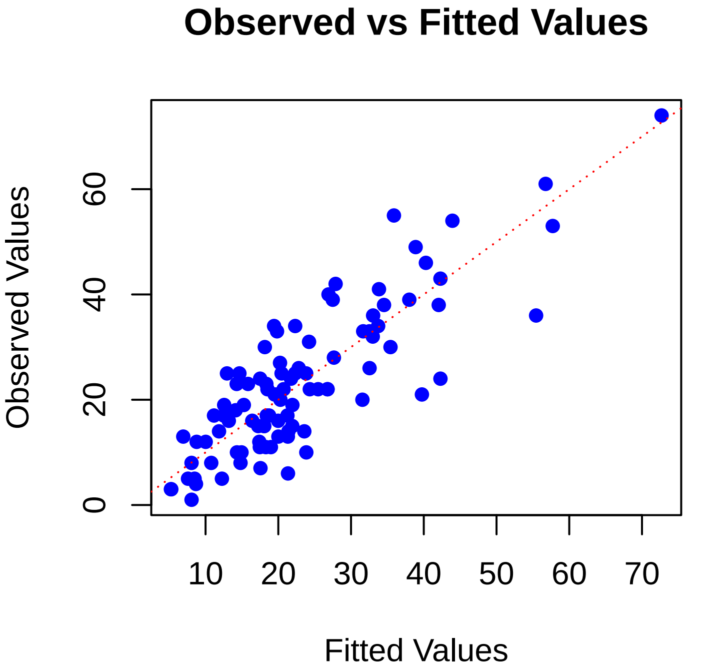

Figure 3: Fitted values vs. observed values for the Poisson model, showing a good fit

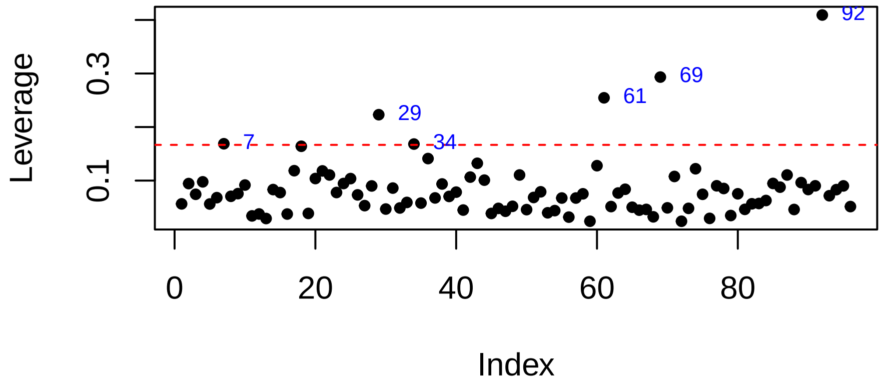

Figure 4: Leverage plot to show which observations have the most influence

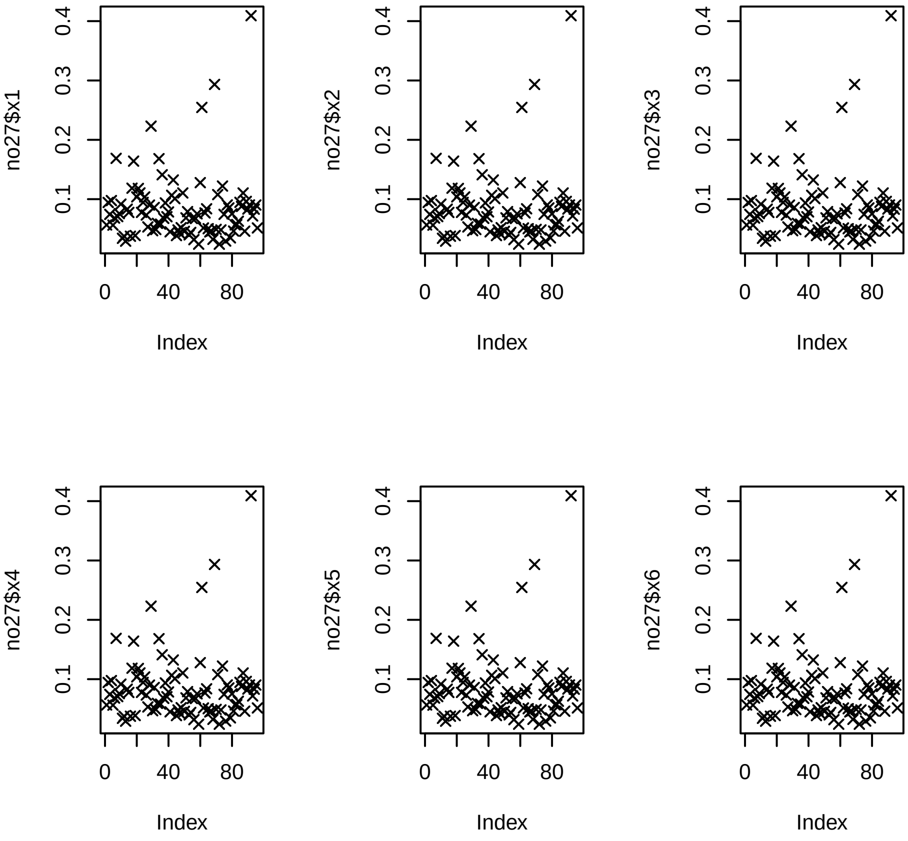

Figure 5: Leverage plot to show which covariates have the most influence on leverages

References

1. Roback P, Legler J. Chapter 4 poisson regression. https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html; 2021.

2. Zajic A. What is akaike information criterion (AIC)? https://builtin.com/data-science/what-is-aic; 2022.

3. Glen S, Leonardo A. https://www.statisticshowto.com/cooks-distance/;

4. Bobbitt Z. How to identify influential data points using cook's distance. https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/; 2019.

5. Ford C. Getting started with negative binomial regression modeling. https://library.virginia.edu/data/articles/getting-started-with-negative-binomial-regression-modeling; 2016.

6. Bobbitt Z. How to use the predict function with glm in R (with examples). https://www.statology.org/r-glm-predict/; 2021.