MATH501 Project: Modelling The Number Of Feeds For Bird Chicks

Student ID - 35955285

19th November 2021

Introduction

The number of individual feeds of chicks is important in their development into adult birds. This study aims to model the number of individual feeds observed for each chick from a particular species of bird, over a certain time period, to determine which factors affect the amount of food chicks from this species receive.

Alongside the response variable, $\mathrm{Y_i}$ - the integer number of feeds observed for a particular chick, the data used in this study contains 7 covariates. These include: $\mathrm{O_i}$, the time (in decimal hours) over which that chick was observed; $\mathrm{S_i}$, sex of the chick; $\mathrm{A_i}$, most common age of the non-breeding group members of the unit - either `Adult` or `Yearling`; $\mathrm{R_i}$, pedigree-based relatedness of the brood (0.5=first-order, 0.25=second order, 0=more distant); $\mathrm{Z_i}$, age of the chick in days; $\mathrm{B_i}$, integer number size of the brood (the number of chicks in the clutch to which the individual chick belongs); $\mathrm{U_i}$, the number of yearlings and adults in a breeding unit.

The Methods, Results and ensuing Discussion have been packaged into one intertwined section to demonstrate the model building process in chronological order. This section is broken down into further subsections, each relating to a stage of the model building process.

Methods, Results and Discussion

Exploratory Data Analysis. We first examine the data to get an overview. There are 97 observations, none of which have any missing or unrecorded entries, and the 7 covariates outlined in the introduction. We have count data, so we expect to construct a Poisson logistic regression model.

By producing plots of each covariate against the response, and then with the covairates against each other, we learn that there is a significant outlier at observation 27 - with the observation time ($\mathrm{O_i}$) of 10000 over 140 times greater than second largest value. We therefore remove this observation. There is a strong positive correlation between the number of feeds (Y_i) and observation time period (O_i). This is because each chick has been monitored over a different length of time and so our number of feeds observed needs to be considered in context as a rate. Therefore we consider observation time period (O_i) as an offset in our model so that our resulting model is not skewed by having a disparity in observation period between birds.

Model Formulation. We initially fit a Poisson logistic regression model with all main effects, since we are using count data and need a discrete distribution such that all entries are greater than 0. We include an offset for Observation Time (O_i) and try 2 different additive offsets in the main effects model. The first offset is just O_i and yields a BIC of 4461.6 (3.d.p) compared to the offset of $\log(O_i)$ which yields a BIC of 740.206 (3.d.p). As a result, we continue with $\log(O_i)$ as our additive offset because it has a lower BIC and so produces the better model. Our initial main effects model is:

$$Y_i \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i \log(O_i)) \quad \text{where the link function is} \quad \eta_i = \log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

with $\mathbf{x}_i$ the vector of covariates and $\boldsymbol{\beta}$ the vector of model coefficients.

For this model, we then check the leverages to identify any influential observations. We have omitted the plot to save space, but these are all reasonable since they are all less than the threshold value of $\frac{3\times(\text{No. Parameters} = 8)}{(\text{Observations} = 96)} = 0.25$ (3sf).

We are interested in finding the most parsimonious model, the model that explains the data best with the fewest number of parameters. Therefore we use a backwards fitting model

selection process, omitting one covariate from our model at a time and testing the model against the original (null model) via a Likelihood Ratio Test. All of the p-values are less than the 0.05 significance level, so all components of the model are significant.

Model Interactions. Plotting the residuals for this model against the covariates shows that the residuals are more variable when the most common age of the non-breeding group members (A_i) is `Yearling` compared to when it is `Adult`. This implies there may be some interaction involving this covariate and so we now use a forward fitting process, working from the current null model, to find any significant two-way interaction terms involving A_i. After adding each interaction in turn, we test the model against the original using a Likelihood Ratio test and choose the model with the lowest p-value at each iteration. We then repeat this, adding interactions until no model gets a p-value less than 0.05 and so all significant components are added. From this, there are 3 interaction terms we add. Table 1 shows the interaction term added at each iteration of this forward selection procedure (Run P1). However, calculating the the leverages again for this new model, we find one extreme value greater than $\frac{3\times(\text{No. Parameters} = 12)}{(\text{Observations} = 96)} = 0.375$ (3sf) (namely observation 61), so we remove this observation and restart again with the newly cleaned data. We repeat this process a twice with the results shown in Table 1. We find a significant leverage at observation 79 on Run 2 and Run P3 produces 2 significant interaction terms and no significant leverages.

Model Diagnostics. To begin testing the appropriateness of this model, we conduct a model deviance test. The model deviance is 156 (3.s.f), which is greater than the critical value at the 5% significance level (108 3.s.f) by a large margin and so our model fit is not very good. To try and improve this, we run the model fitting procedure again using a Negative-Binomial response model in place of Poisson, where the Negative-Binomial distribution is a generalized Poisson distribution including a Gamma noise variable. This model fitting procedure emulates the procedure used in the previous 2 sections and the results are displayed in Table 1. The final model produced contains just one interaction, between the most common age of the non-breeding group members and the age of the chick in days. The corresponding leverages are shown in Figure 1 and are all clearly less than the significant threshold of 0.287.

We then repeat the deviance test with our new model. We get a model deviance of 100 (3.s.f) which is less than the critical value at the 5% significant level - 108 (3.s.f). Therefore, our new model is a much better fit than its previous Poisson counterpart. This is further reinforced by a plot of the residuals against the fitted model (Figure 1), which shows no clear pattern and implies a good model fit.

Run	Iteration 1	p-value (3.s.f)	Iteration 2	p-value (3.s.f)	Iteration 3	p-value (3.s.f)
P1	$A_i \times Z_i$	8.37×10^{-16}	$A_i \times R_i$	4.15×10^{-8}	$A_i \times B_i$	0.00995
P2	$A_i \times Z_i$	3.24×10^{-17}	$A_i \times R_i$	8.25×10^{-4}	$A_i \times B_i$	0.0118
P3	$A_i \times Z_i$	6.5×10^{-17}	$A_i \times B_i$	0.0174		
NB1	$A_i \times Z_i$	10^{-8}	$A_i \times R_i$	3.37×10^{-5}	$A_i \times B_i$	0.0277
NB2	$A_i \times Z_i$	5.56×10^{-10}	$A_i \times R_i$	0.00505	$A_i \times B_i$	0.0354
NB3	$A_i \times Z_i$	2.15×10^{-10}				

Table 1: Significant p-values for each forward selection run through (both for Poisson and Negative Binomial Model), with interaction added at each iteration.

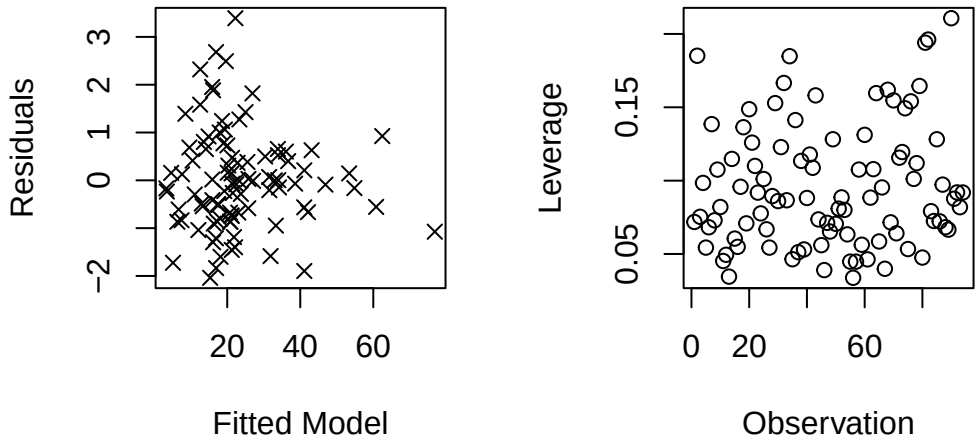

Figure 1: Left: Plot of Residuals Against Fitted Model. Right: Plot of Leverage By Observation.

Analysis and Conclusion

The final fitted model for our data $\mathrm{Y_i}$ follows a Negative-Binomial distribution: $Y_i \overset{\text{indep}}{\sim} \text{Negative-Binomial}\,(\mu_i, \alpha = 42.0)$. Where the mean component is given by:

$$\mu_i = \exp(log(\mathrm{O}_i)+\beta_1+\beta_2\cdot 1_{\mathrm{S=male}}+\beta_3\cdot 1_{\mathrm{A=yearling}}+\beta_4\cdot 1_{\mathrm{R=0.25}}+\beta_5\cdot 1_{\mathrm{R=0.5}}+\beta_6\cdot Z_i+\beta_7\cdot B_i+\beta_8\cdot U_i+\beta_9\cdot A_i\cdot Z_i)$$

and the coefficients are (to 3.s.f) as follows: $\beta_1 = -0.157$, $\beta_2 = 0.177$, $\beta_3 = -1.21$, $\beta_4 = 0.440$, $\beta_5 = 0.378$, $\beta_6 = 0.0184$, $\beta_7 = 0.267$, $\beta_8 = -0.0690$, $\beta_9 = 0.0744$.

Interpreting these values, we deduce that a female adult chick with more distant pedigree receives $\exp(-0.157) = 0.854$ (3.s.f) per hour. The additional amount of feed per hour that a male chick can expect is 1.19 (3.s.f).For each additional chick in the brood, the chick revives 1.31 more feeds per hour and for each additional bird in the caring unit the chick receives 1.07 extra feeds per hour.

We are interested in predicting the number of feeds for a female chick observed over 1 hour, where $R_i = 0.5$, $Z_i = 10$, $B_i = 2$, $U_i = 10$ and A_i is either `yearling` (Chick A) or `adult` (Chick B). Using the equation above, we calculate μ_i for each respective chick getting μ_A=0.88 (3.d.p) and μ_B=1.283 (3.d.p). A 95% confidence interval for the prediction of Chick A is given by (0.688, 1.072) and for Chick B (1.051, 1.515). This indicates that a female chick with the listed characteristics is expected to get less food when when in a unit with a higher number of yearlings than adults, and therefore more competition.

In conclusion, we have built a robust model to predict the number of feeds for chicks per hour, given a set of conditions. We have assumed that the data are independently distributed, meaning that if any chicks from this study were part of the same breeding unit then our model may be incorrect. A Negative Binomial logistic regression was found to be the best fitting model.