Predicting Flight Response of Geese

Author: Leigh West

Evaluating Suitability of a Logistic Regression Model

We are interested in evaluating the impact of increased helicopter traffic in an estuary on flocks of a particular species of geese. Logistic regression is the appropriate regression analysis to perform when the response variable is dichotomous. In this particular data set the flight response of the geese is recorded as either low (0) or high (1), and accordingly, logistic regression is the appropriate method to evaluate this binary data. Moreover, a linear model is not appropriate because a linear model relies on the assumption that the data is normally distributed. As this data is binary, a Bernoulli distribution is likely to be suitable.

Model Selection with Backward Stepwise Regression

Variable screening was performed in R using backward stepwise regression, commencing with the interaction model and the lower model set to the null model. Per Table 1, the interaction model produced an Akaike's Information Criteria (AIC) of 350. Removing the interaction term would have increased the AIC to 366 (and increased the residual deviance), and hence backward stepwise regression did not progress beyond the initial step.

Table 1: Backward Stepwise Regression (First and Only Step)

```
First Step

Start: AIC=350
RESPONSE ~ ALTITUDE * LATERAL
                    Df   Deviance   AIC
<none>                        342   350
- ALTITUDE:LATERAL  1         360   366
```

Per the analysis of deviance of the final stepwise model (Table 2), altitude (p-value = 0.01 < 0.05), lateral (p-value ≈ 0) and the interaction between altitude and lateral (p-value = 3.2×10^{-5} < 0.05) predictor variables are considered useful as they are able to explain a significant amount of the deviance. In addition, once the main effects and interaction term are fitted to the model, a residual deviance of 342 remains, corresponding to a p-value on the chi-square distribution of 1 > 0.05 (see Table 2). Hence we can infer that the residual deviance is not significant and that fitting the altitude and lateral predictor variables and their interaction accounts for the majority of the variation. It appears that the model is adequate.

Table 2: Analysis of Deviance Table (Final Stepwise Model)

Analysis of Deviance Table

Model: binomial, link: logit

Response: RESPONSE

	Df	Deviance	Res Df	Res Dev	Pr(>Chi)
NULL	1		463	619	
ALTITUDE	1	6.7	462	612	0.0096
LATERAL	1	252.5	461	360	<2e-16
ALTITUDE:LATERAL	1	17.3	460	342	3.2e-05

Determining the Final Model

Table 3: Table of Regression Coefficients (Final Stepwise Model)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.82504	0.49922	7.66	1.8e-14
ALTITUDE	-0.20301	0.10558	-1.92	0.055
LATERAL	-0.39094	0.04839	-8.08	6.5e-16
ALTITUDE:LATERAL	0.04014	0.00949	4.23	2.3e-05

In the summary data in Table 3 (above), we see that the coefficient of the lateral predictor is -0.39, which is statistically different from zero (p-value ≈ 0) and is therefore considered a useful predictor. The altitude predictor variable is not statistically significant ($p = 0.055 > 0.05$), however the interaction between the altitude and lateral predictor variables is ($p = 2.3 \times 10^{-5} < 0.05$). Therefore, the lateral main effect term should remain in the model. Note also, that per the stepwise regression (Table 3), the model's AIC is lowest when both main effects and the interaction term are included. Considering this, the final model determined as follows:

$$\log_e \frac{\pi}{1-\pi} = 3.83 - 0.20 * Altitude - 0.39 * Lateral + 0.04 * Altitude{:}Lateral$$

where π is the proportion of geese whose flight response was "high", and our regression coefficients are on the logit scale.

Determining the Probability of a High Flight Response

The predicted probability of a high flight response of the geese for a helicopter's given altitude and lateral distance from the flock is provided in Table 4 (below). Note that our model has a negative coefficient for both the altitude and lateral distance predictor values, with the coefficient of the lateral distance predictor almost twice as large as that of the altitude predictor. Accordingly, we see a progressive but relatively shallow decline in the probability of a high flight response as altitude increases and lateral distance remains constant at 0 metres. Note also that by observing the response at a lateral distance of 0 we're also able to compare the effects of altitude independent of the interaction term.

Table 4: Probability Matrix of High Flight Response of Geese
(A = altitude; L = lateral distance: in hundreds of metres)

	A3	A6	A9	A12
L0	0.961432	0.9313	0.881	0.800
L10	0.624965	0.7513	0.846	0.909
L20	0.100232	0.4024	0.803	0.961
L30	0.007392	0.1305	0.752	0.984
L40	0.000498	0.0324	0.692	0.993

However, the effect of the interaction term is evident when we consider the probability of a high flight response as lateral distance increases at an altitude of 300 metres compared to 1,200 metres. Recall that the coefficient of lateral distance is negative and approximately twice that of altitude, and hence, we see the probability of a high flight response reduce to approximately 0 for large lateral distance values (i.e. 4,000 metres) when altitude is held constant at 300 metres. However, the odds of a high flight response actually increase as lateral distance increases when altitude is fixed at 1,200 metres. This highlights that the interaction term results in the effect of altitude differing if the helicopter is laterally closer vs farther.

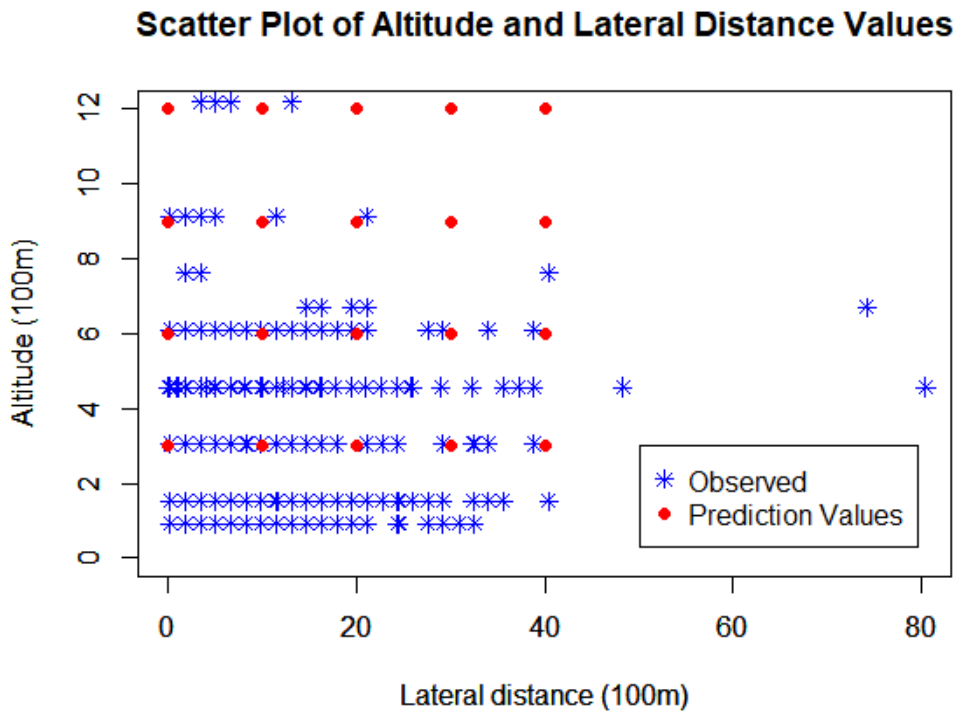

Figure 1: A scatter plot of observed helicopter altitudes and lateral distances (blue) and the altitude and lateral distance predictor values passed in to the model.

Figure 1 (above) is a scatter plot diagram of the observed altitudes and lateral distances of the helicopter (in blue), with the pairs of various, arbitrary altitudes and lateral distances passed into our model (in red). The figure shows that that the majority of values used to produce the prediction matrix in Table 4 are within the range of our observed values. However, the altitude and lateral distance pairs (900, 4,000), (1,200, 3,000) and (1,200, 4,000) appear to be outside of our experimental region and are therefore constitute extrapolation.

Predicting using values outside of our experimental region may lead to errors of prediction which are much larger than expected, and therefore should not be relied upon.