Moth Coloration and Natural Selection

Izabela Litwin

Introduction

The goal of this study is to understand whether the proportion removed differs between dark morph moths and light morph moths and, more importantly, whether this difference depends on the distance from Liverpool. If the relative proportion of dark morph removals increases with increasing distance from Liverpool, that would be evidence in support of survival of the fittest, via appropriate camouflage.

In this study, I present stable, interpretable, theoretically valid models to address these research questions. Since this is a planned experiment, the focus of this report is inference, not prediction.

The final findings suggest that there is enough statistical evidence to claim that the proportion removed differs between dark morph moths and light morph moths. The odds of removal for the dark morph moth, relative to the odds of removal for the light morph moth, increase with increasing distance from Liverpool.

Exploratory Data Analysis

Part 1: Univariate EDA

The data analyzed in this report contains 56 observations and 6 variables. The variables include the name of the area selected by Bishop, their distances from Liverpool in miles, type of moth glued to trunk (light or dark), exposure of the side of tree where moths were placed, number of moths placed and the number of moths removed. Additionally, I added a variable called `proportion` that represents the proportion of moths removed (out of moths placed). Tshis variable can naturally vary from 0 to 1.

There are 7 locations and 8 observations under each location. The distance varies from 0 to 51.2 miles from Liverpool. There are 28 dark morph moths and 28 light morph moths. There are 14 observations on each side of a tree where moths were placed. The number of morphs placed varies from 13 to 23; the number of moths removed varies from 0 to 16.

The response variable is the proportion of moths removed (out of moths placed). Based on the summary of variables and the histograms (Table 1 and Figure 1), the proportion of moths removed varies from 0 to 76.92%. On average, 31.35% of moths were removed. The distribution is right-skewed with some outliers. The distributions of side, morph and location have the same number of observations in each category.

Table 1: Summary of Continuous Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
DIST	56	27.229	17.156	0	7.2	41.5	51
PLACED	56	17.286	3.686	13	14	21	23
REMOVED	56	5.464	3.098	0	3	7	16
PROPORTION	56	0.313	0.162	0.000	0.211	0.385	0.769

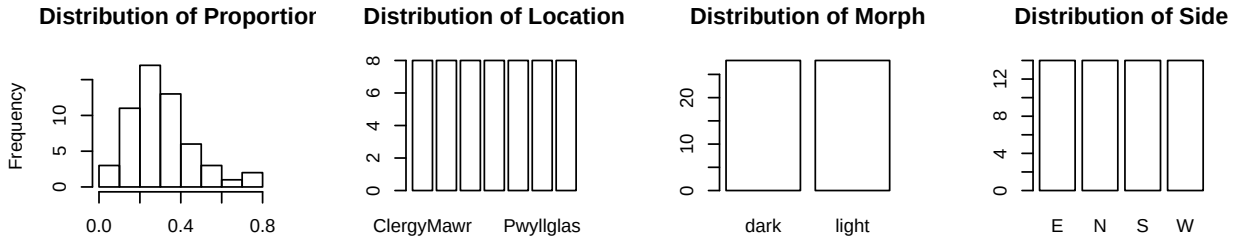

Figure 1: Univariate EDA

Part 2: Multivariate EDA

Given the correlation matrix (Table 2), we can see that the number of moths placed is correlated with the distance from Liverpool (0.48), the number of moths removed is correlated with the number of moths placed (0.47) and the number of moths removed is correlated with the distance from Liverpool (0.30). This creates a potential interaction effect.

Table 2: Covariance Matrix (Continuous variables)

	DIST	PLACED	REMOVED	PROPORTION
DIST	1.00	0.48	0.30	0.11
PLACED	0.48	1.00	0.47	0.08
REMOVED	0.30	0.47	1.00	0.90
PROPORTION	0.11	0.08	0.90	1.00

Based on the stacked bar plots and side-by-side scatterplots (Figure 2), we can tell that the response variable (proportion) is correlated with the location. In particular, the higher proportion of moths removed was in Hawarden. Also, it looks like more of the moths removed were dark moths. Also, on average, those moths were removed more from the South than other directions. Also, the proportion of moths removed is very correlated with the number of moths removed which is logical given that the distribution of the number of moths placed is very concentrated and has very small variations.

Plots analyzing the relationships between all available variables and the response are displayed in Figure 2.

Additionally, in order to analyze the relationship between the controlled factors and the response, I plotted the proportion of moths removed against distance from Liverpool for both types of moths. Figure 3 suggests that the proportion of moths removed for dark morphs is larger than for light morphs as the distance increases.

Modeling & Diagnostics

Part 3: Statistical Models

To answer the research questions, I constructed two additive models.

Distance from Liverpool in miles is treated as a continuous variable. I decided not to discretize it because of the continuous nature of `distance`. The type of moth (`morph`) is a categorical variable with "light" and "dark" categories. This variable is treated as a factor in the model. Additionally, there is an interaction term between `distance` and `morph` that will help answer the research question.

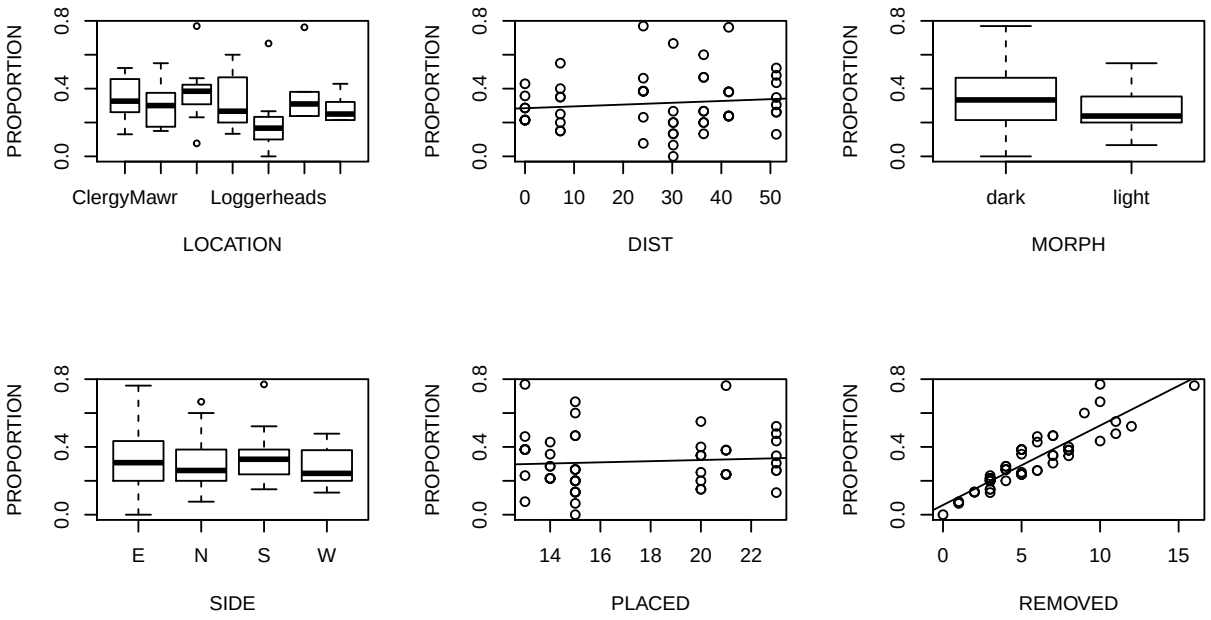

Figure 2: Multivariate EDA

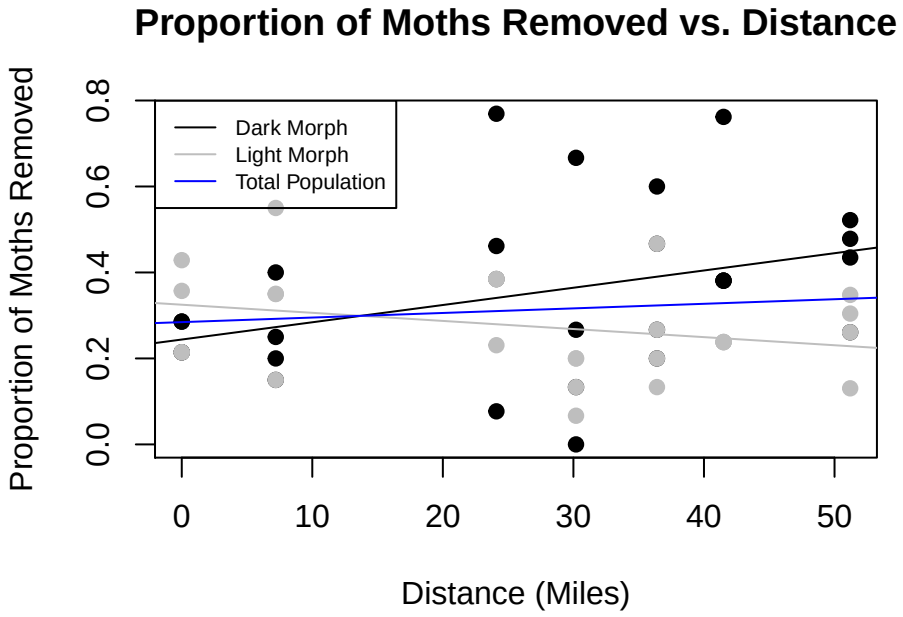

Figure 3: Relationship between Proportion of Moths Removed, Morph and Distance

I decided not to include `location` variable as it conveys the same information as `distance` does. I did not include `side` variable because it is not needed to answer the research questions. Based on the correlation matrix (Table 2), we know that `placed` and `removed` are potential confounders as they are correlated with both `distance` and `proportion`. Variables `placed` and `removed` were used to compute the proportion of moths removed. Therefore, I removed those two features from the model as well.

Since the dependent variable is a proportion bounded by 0 and 1, it was logit-transformed.

Model 1: Parametric Model - GLM

The first model I constructed is a logistic regression model:

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1 - proportion_i}) = \beta_0 + \beta_{distance}\text{distance}_i \\ + \beta_{morph}\text{morph}_i \\ + \beta_{distance:morph}\text{distance} : \text{morph}_i\end{aligned}$$

The number of moths placed was used as binomial prior weights in the GLM model in order to account for the fact that different locations had different total numbers of placed moths. I used `placed` to re-weight the data to correct for the discrepancy.

The summary of this model is included below in Table 3.

I made the following assumptions while fitting this model:

- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome
- There is no influential values (extreme values or outliers) in the continuous predictors
- There is no high intercorrelations (i.e. multicollinearity) among the predictors.

Model 2: Non-Parametric Model - GAM

The second model I constructed is a general additive model from a binomial family:

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1 - proportion_i}) = \beta_0 + \text{s}(\text{distance}, \text{by} = \text{morph}, \text{k} = 6) \\ + \beta_{morph}\text{morph}_i \\ + \text{s}(\text{distance}_\text{i}, \text{k} = 6)\end{aligned}$$

I used the `mgcv` package to fit this model in R. GAMs take each predictor variable in the model and separate it into sections (delimited by "knots"), and then fit polynomial functions to each section separately, with the constraint that there are no kinks at the knots (second derivatives of the separate functions are equal at the knots). The goal is to minimize the residual deviance (goodness of fit) while maximizing parsimony (lowest possible degrees of freedom).

I fit an additive model of the proportion of moths removed on morph, distance, and interaction of morph and distance, with smoothing splines for distance, interaction term and a step function for morph. Smooth terms as distance represented using penalized regression splines with df. Because there are only a few different values of `distance`, I set the number of basis functions to k=6 which sets the upper limit on the degrees of freedom for a smooth using s. The exact choice of k is not generally critical: it was chosen to be large enough that we are reasonably sure of having enough degrees of freedom to represent the underlying "truth"

	Model 1	Model 2
(Intercept)	-0.72^{***}	-0.97^{***}
	(0.19)	(0.10)
DIST	-0.01	
	(0.01)	
MORPHdark	-0.41	0.35^{*}
	(0.27)	(0.14)
DIST:MORPHdark	0.03^{***}	
	(0.01)	
EDF: s(DIST):MORPHlight		1.00^{***}
		(1.00)
EDF: s(DIST):MORPHdark		0.00
		(0.00)
EDF: s(DIST)		1.00^{**}
		(1.00)
AIC	268.02	268.02
BIC	276.13	276.13
Log Likelihood	-130.01	-130.01
Deviance	93.27	93.27
Num. obs.	56	56
Deviance explained		0.19
Dispersion		1.00
R^2		0.16
GCV score		0.81
Num. smooth terms		3

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Statistical models

reasonably well, but small enough to maintain reasonable computational efficiency. For the interaction term, I used "by" which creates the "factor smooth: smoothing class, where a smooth function of `distance` is created for each level of `morph`.

Again, the number of moths placed was used as binomial prior weights in the GAM model in order to account for the fact that different locations had different total numbers of placed moths. I used `placed` to re-weight the data to correct for the discrepancy.

The summary of this model is included below in Table 3.

Part 4: A Non-Parametric Test

The bootstrap allows us to measure the uncertainty in the (cross-validation) mean-squared-error calculations to help decide whether one of the models is actually better than the other. To compare those two models, I bootstrapped 5-fold cross-validation. I created $B = 500$ bootstrap samples each consisting of $n = 56$ observations from the data set selected at random with replacement. I used "resampling cases" form of the bootstrap (nonparametric bootstrap). For each of the bootstrap samples, I randomly divided the n observations into 5 disjoint sets of equal size. Treating each of the 5 folds as test data and the other 4 as training data, I calculated prediction error for each model and called the average of the prediction errors for the GLM model MSE1b. The average of GAM model is called MSE2b. I drew a histogram of MSE2b - MSE1b to see if there's visual evidence whether one model is better.

The proportion of test errors of MSE2b - MSE1b being larger than 0 is equal to 94%. In Figure 4, we see that the GLM model (logistic regression) is almost uniformly better. Therefore, the final model is Model 1:

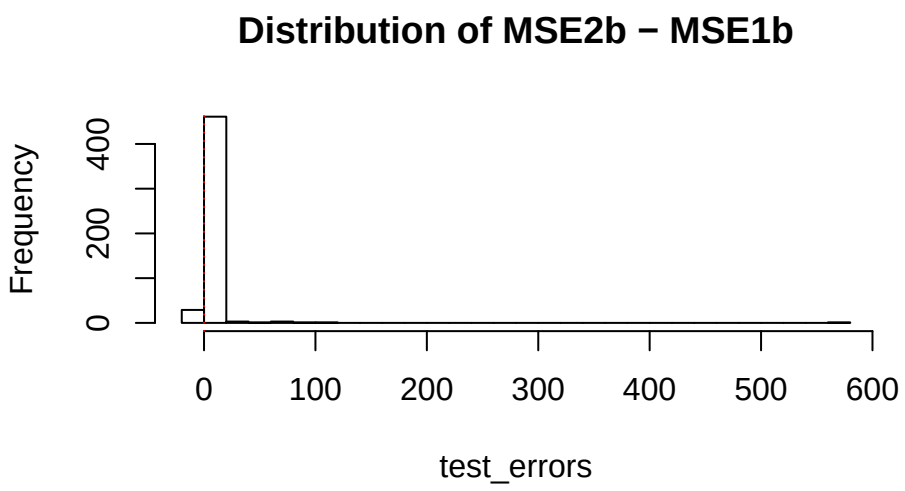

Figure 4: Distribution of the Test Statistics

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1-proportion_i}) = -0.72 &- 0.01\text{distance}_i \\ &- 0.41\text{morph}_i \\ &+ 0.03\text{distance} : \text{morph}_i\end{aligned}$$

Part 5: Model Diagnostics

To assess the model fit, I performed the likelihood ratio (deviance) goodness of fit test.

This likelihood ratio test statistic is approximately distributed as a chi-square deviate with n-q degrees of freedom, where q is the number of covariates. If D is larger than expected (i.e., the p-value is small), this means that the model with the covariates included is not sufficient to explain the data. The p-value from this global goodness-of-fit is 0.000387. It implies that for this data, the model does not appear to fit very well.

There are several variables such as `distance` and `morph` with statistically insignificant coefficients. Only the interaction term looks like it is significantly impacting the fit. Of course, some variables are correlated with others (refer to the correlation table in EDA). Even though neither of `distance` and `morph` has a significant coefficient, they could each be making the other's effect.

Additionally, to improve the accuracy of my model, I tried to make sure that the model assumptions hold true for the data.

1) Linearity Assumption

I checked the linear relationship between continuous predictor variables and the logit of the outcome by visually inspecting the scatter plot between each predictor and the logit values.

Based on Figure 5, the variable `distance` is not linear and might need some transformations. If the scatter plot shows non-linearity, we need other methods to build the model such as including 2 or 3-power terms, fractional polynomials and spline function.

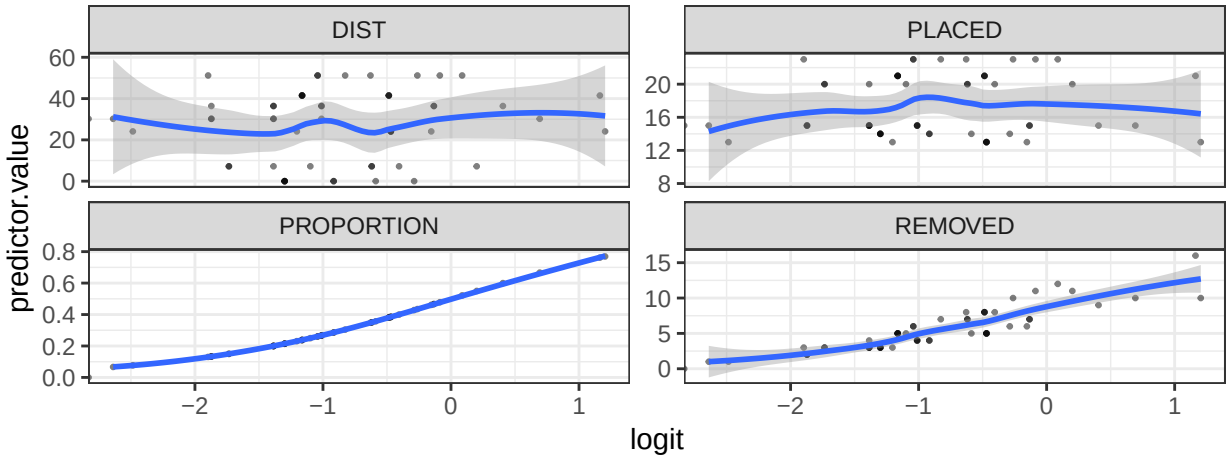

Figure 5: Linearity Assumption - Scatter Plots

2) Outliers and Influential Values

There are two main outliers that have the highest `proportion` in the data set:

- Observation 47 with a proportion of moths removed of 0.7619. Placed in Pwyllglas, 41.5 miles from Liverpool. Dark morph.
- Observation 22 with a proportion of moths removed of 0.7692. Placed in Hawarden, 24.1 miles from Liverpool. Dark morph.

I also checked the data for influential values which are extreme individual data points that can alter the quality of the logistic regression model. The most extreme values in the data can be examined by visualizing the Cook's distance values. In Figure 6 I label the top 3 largest values.

It appears like not all outliers are influential observations. To check whether the data contains potential influential observations, the standardized residual error can also be inspected. Data points with absolute standardized residuals above 3 represent possible outliers and may deserve closer attention.

I computed the standardized residuals and the Cook's distance using the R function augment() in `broom` package.

I filtered the potential influential data points with the Cook's distance. The data for those observations, according to the Cook's distance, is displayed in Table 4 below.

Table 4: Influential Points According to Cook's Distance

PROPORTION	DIST	MORPH	.fitted	.se.fit	.sigma	.cooksd	index
0.5500000	7.2	light	-0.7845943	0.1566877	1.313099	0.1719233	11
0.7619048	41.5	dark	-0.3611427	0.1136470	1.266501	0.2012019	47
0.2608696	51.2	dark	-0.1816708	0.1503585	1.321420	0.1480530	53

I filtered the potential influential data points with the absolute value of the standardized residuals being above 3. The data for those observations, according to the standardized residuals, is displayed in Table 5 below.

Overall, based on the Cook's distance and the absolute value of the standardized residuals, it appears like observation 47 with the `proportion` of 0.7619 is the most influential point. It is the outlier that I identified upfront. With this outlier, potential solutions include removing that record or using non-parametric methods.

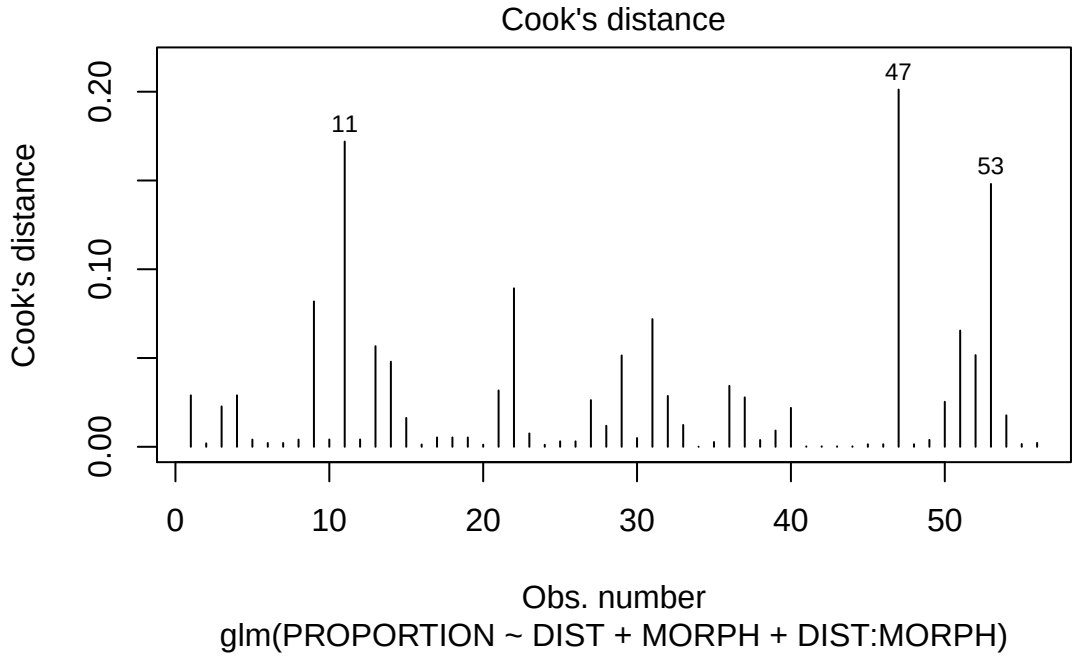

Figure 6: Cook's Distance Values

Table 5: Influential Points According to Standardized Residuals

PROPORTION	DIST	MORPH	.fitted	.se.fit	.sigma	.cooksd	index
0.7692308	24.1	dark	-0.6830820	0.1027513	1.273420	0.0892669	22
0.0000000	30.2	dark	-0.5702183	0.0958522	1.247626	0.0719213	31
0.7619048	41.5	dark	-0.3611427	0.1136470	1.266501	0.2012019	47

3) Multicollinearity

I also checked the model for multicollinearity which corresponds to a situation where the data contain highly correlated predictor variables.

Table 6: Multicollinearity: Variance Inflation Factors

	VIF
DIST	2.066409
MORPH	3.833723
DIST:MORPH	5.181157

As a rule of thumb, a variance-inflation value that exceeds 5 (or 10) indicates a problematic amount of collinearity. Based on Table 6, in the data set there is collinearity: `distance:morph` has a VIF value of 5.18. To solve this problem and produce a better model fit, this variable could potentially be removed if it is not needed to answer the research question.

Results

Part 6: Proportion Removed for Dark Morph Moths and Light Morph Moths

In order to verify if the proportion of removed moths is different between dark and light moths, I fit the following 2 models:

Model 1) The logistic regression model with three predictors (Model 1 from Part 3: Statistical Models)

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1-proportion_i}) = \beta_0 &+ \beta_{distance}\text{distance}_i \\ &+ \beta_{morph}\text{morph}_i \\ &+ \beta_{distance:morph}\text{distance} : \text{morph}_i\end{aligned}$$

Model 2) The logistic regression model with one predictor

$$\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1-proportion_i}) = \beta_0 + \beta_{distance}\text{distance}_i$$

Then, I tested the following hypotheses:

$$H_0 : \beta_{morph} = \beta_{distance:morph} = 0$$
$$H_A : \beta_{morph} \neq \beta_{distance:morph} \neq 0$$

To do that, I performed a deviance test (chi-square) presented in Table 7. The drop in deviance statistic was 20.396 on 2 df. That yields the p-value of 0.00004. The deviance test suggests that Model 1 makes a significant improvement over Model 2. Therefore, we reject H_0. There is strong evidence suggesting that the proportion of removed moths is different between dark and light morph moths.

Table 7: ANOVA: Deviance Test 1

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
54	113.667	NA	NA	NA
52	93.271	2	20.396	0

The assumptions I made while conducting this test are mostly those of a logistic regression model, i.e., I assumed that there is a linear relationship between the logit of the outcome and each predictor variables, there are no influential values in the continuous predictors and that there are no high intercorrelations (i.e., multicollinearity) among the predictors.

Part 7: Proportion Removed for Dark Morph Moths and Light Morph Moths vs. Distance

To verify whether a potential difference between dark and light moths depends on the distance from Liverpool, I fit two models:

Model 1) The logistic regression model with three predictors (Model 1 from Part 3: Statistical Models)

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1-proportion_i}) = \beta_0 &+ \beta_{distance}\text{distance}_i \\ &+ \beta_{morph}\text{morph}_i \\ &+ \beta_{distance:morph}\text{distance} : \text{morph}_i\end{aligned}$$

Model 2) The logistic regression model with two predictors

$$\begin{aligned}\text{logit}(\text{proportion}_\text{i}) = log(\frac{proportion_i}{1-proportion_i}) = \beta_0 &+ \beta_{distance}\text{distance}_i \\ &+ \beta_{morph}\text{morph}_i\end{aligned}$$

Then, I tested the following hypotheses:

$$H_0 : \beta_{distance:morph} = 0$$
$$H_A : \beta_{distance:morph} \neq 0$$

To do that, I performed a deviance test (chi-square) presented in Table 8. The drop in deviance statistic was 11.931 on 1 df. That yields the p-value of 0.00055. The deviance test suggests that Model 1 makes a significant improvement over Model 2. Therefore, there is strong evidence of an interaction. The odds of removal for the dark morph, relative to the odds of removal for the light morph, increase with increasing distance from Liverpool. In other words, the relative proportion of dark morph removals increases with increasing distance from Liverpool.

Table 8: ANOVA: Deviance Test 2

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
53	105.202	NA	NA	NA
52	93.271	1	11.931	0.001

The effect of the interaction term is as follows: the odds of removing dark morph moths at the distance of 1 mile from Liverpool is 0.6815 ($exp(\beta_{morph} + \beta_{distance:morph})$) times higher than the odds of removing light morph moths.

The assumptions I made while conducting this test are mostly those of a logistic regression model, i.e., I assumed that there is a linear relationship between the logit of the outcome and each predictor variables, there are no influential values in the continuous predictors and that there are no high intercorrelations (i.e., multicollinearity) among the predictors.

Conclusions/Discussion

There is enough statistical evidence to claim that the proportion removed differs between dark morph moths and light morph moths. The relative proportion of dark morph removals increases with increasing distance from Liverpool. Therefore, there is evidence in support of survival of the fittest, via appropriate camouflage.

I cannot make any causal statements because this is not a fully randomized experiment as the moths were not completely randomly assigned to trees, positions on trees and distances. Also, I cannot make any causal

statements because there is a possibility of confounding variables (weather conditions, presence of people in the area, etc. that may affect the presence of birds or predators).

In the future, it would be beneficial to gather more data to perform this analysis and possible, conduct a randomized study. Furthermore, it would be helpful to account for potential confounding variables such as weather, the human presence or any other factors that may affect the presence of birds or predators which remove the moths from the trees.