Logistic Regression Models

Cheng Peng

Contents

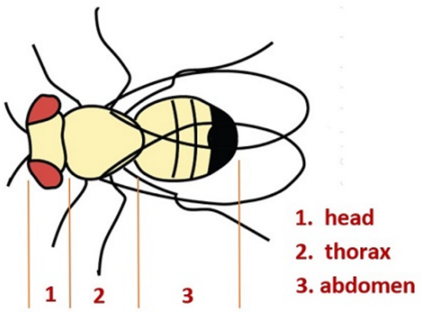
1 Introduction

Linear regression models are used to assess the association between the continuous response variable and other predictor variables. If the response variable is a binary categorical variable, the linear regression model is not appropriate. We need a new model, the logistic regression model, to assess the association between the binary response variable and other predictor variables.

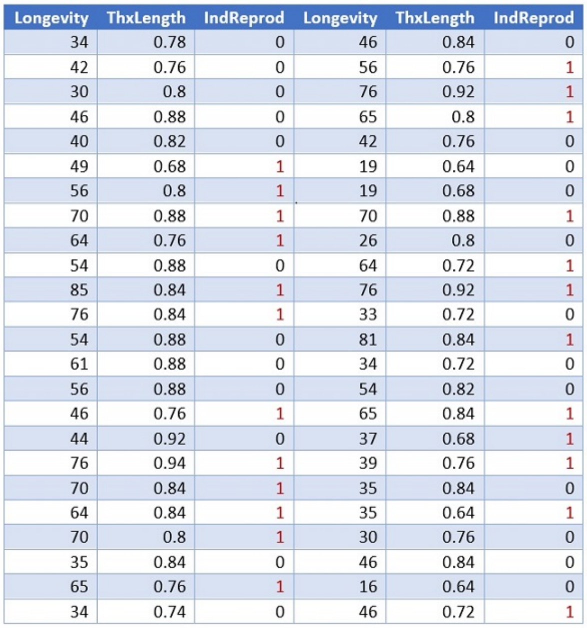
This module focuses on the regression model with a binary response.

2 Motivational Example and Practical Question

Example : Longevity in male fruit flies is positively associated with adult size. However, reproduction has a high physiological cost that could impact longevity.



The original study looks at the association between longevity and adult size in male fruit flies kept under one of two conditions. One group is kept with sexually active females over the male's life span. The other group is cared for in the same way but kept with females who are not sexually active.



Longevity	ThxLength	IndReprod	Longevity	ThxLength	IndReprod
34	0.78	0	46	0.84	0
42	0.76	0	56	0.76	1
30	0.8	0	76	0.92	1
46	0.88	0	65	0.8	1
40	0.82	0	42	0.76	0
49	0.68	1	19	0.64	0
56	0.8	1	19	0.68	0
70	0.88	1	70	0.88	1
64	0.76	1	26	0.8	0
54	0.88	0	64	0.72	1
85	0.84	1	76	0.92	1
76	0.84	1	33	0.72	0
54	0.88	0	81	0.84	1
61	0.88	0	34	0.72	0
56	0.88	0	54	0.82	0
46	0.76	1	65	0.84	1
44	0.92	0	37	0.68	1
76	0.94	1	39	0.76	1
70	0.84	1	35	0.84	0
64	0.84	1	35	0.64	1
70	0.8	1	30	0.76	0
35	0.84	0	46	0.84	0
65	0.76	1	16	0.64	0
34	0.74	0	46	0.72	1

Figure 1: Fruit Flies Data Table

The above table gives the longevity in days for the male fruit flies given an opportunity to reproduce (IndReprod = 0) and for those deprived of the opportunity (IndReprod = 1).

The data was collected from a case-control study design. The original study association analysis used the multiple linear regression model in which Longevity was a response variable and Thorax and IndReprod were used as predictor variables. In this example, we build a logistic regression to assess the association between longevity and reduction. Due to the case-control study design, the resulting logistic regression cannot be used as a predictive model.

Since the response variable is binary (i.e., it can only take on two distinct values, 0 and 1 in this example), the linear regression line is a bad choice since (1) the response variable is not continuous, (2) the fitted regression line can take on any values between negative infinity and positive infinity. The response variable takes on only either 0 or 1 (see the dark red straight line).

A meaningful approach to assess the association between the binary response variable and the predictor variable by looking at how the predictor variables impact the probability of observing the **bigger** value of the response variable. **If the response is a character variable, the one that has a higher alphabetical**

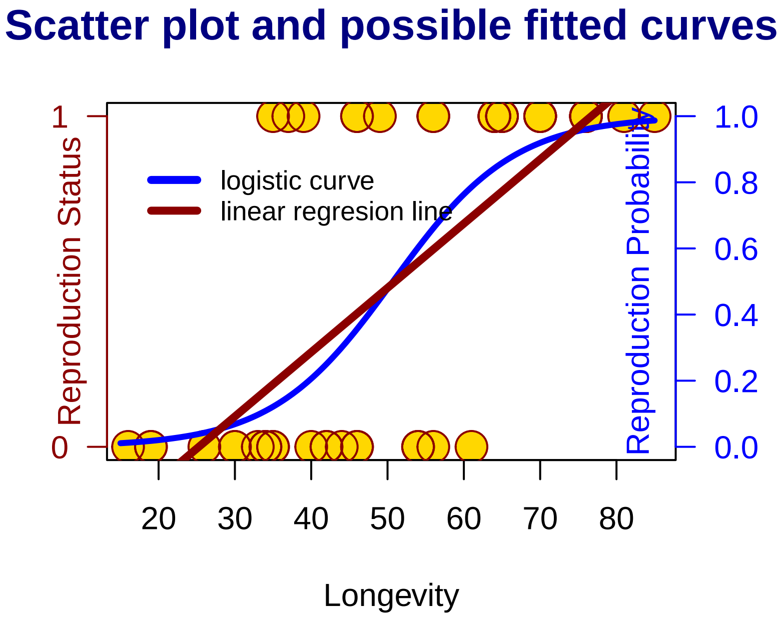


Figure 2: The scatter plots of a binary response v.s. a continuous predictor variable

order for the character response is called a "bigger" value. The above **S** curve describes the relationship between the values of the predictor variable(s) and the probability of observing the **bigger** value of the response variable.

3 Logistic Regression Models and Applications

Logistic regression is a member of the generalized linear regression (GLM) family which includes the linear regression models. The model-building process is the same as that in linear regression modeling.

3.1 The Structure of the Logistic Regression Model

The logistic regression models the actual probability function of observing the **bigger** value of the response variable. In the above example, the **bigger** value of the response variable is **Y = IndReprod**. Therefore, the simple logistic regression model for giving the predictor variable `Longevity` is given by

$$P(Y=1)=\frac{\exp(\beta_0+\beta_1\text{Longevity})}{1+\exp(\beta_0+\beta_1\text{Longevity})}$$

If β_1 (also called slope parameter) is equal to zero, the longevity and the reproduction status are NOT associated. Otherwise, there is an association between the response and the predictor variables. The sign of the slope parameter reflects the positive or negative association.

3.2 Assumptions and Diagnostics

There are assumptions for the binary logistic regression model.

- The response variable must be binary (taking on only two distinct values).
- Predictor variables are assumed to be uncorrelated.

- The functional form of the predictor variables is correctly specified.

The model diagnostics for logistic regression are much more complicated than the linear regression models. We will not discuss this topic in this course.

3.3 Coefficient Estimation and Interpretation

The estimation of the logistic regression coefficients is not as intuitive as we saw in the linear regression model (regression lines and regression plane or surface). An advanced mathematical method needs to be used to estimate the regression coefficient. The R function **glm()** can be used to find the estimate of regression coefficients.

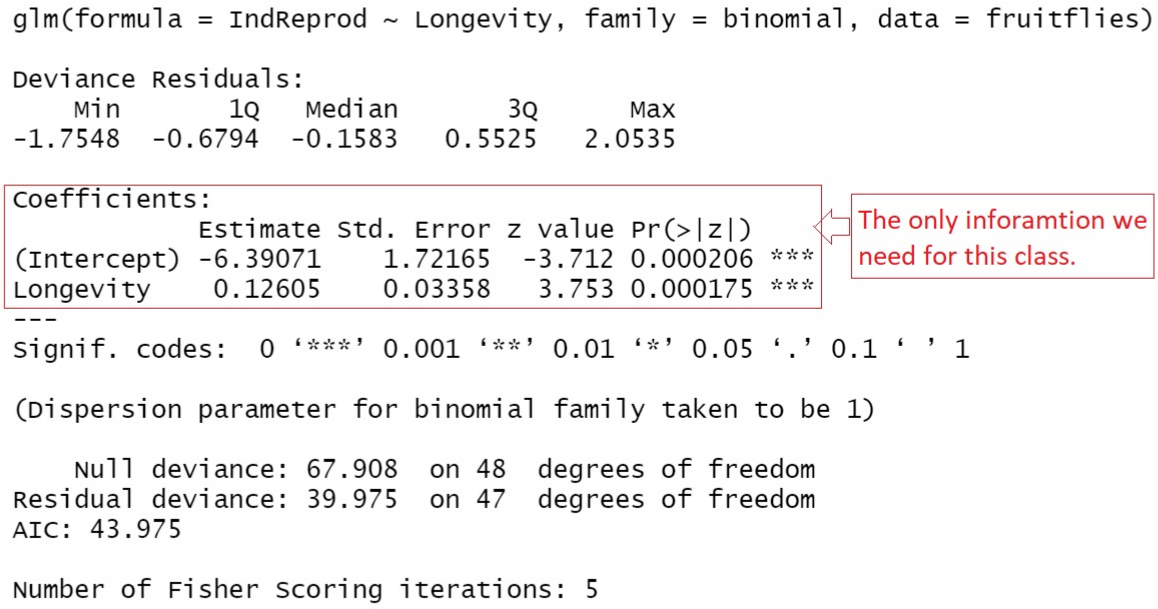
The interpretation of the coefficients of the logistic regression model is also not straightforward. In the motivational example, the value of β_1 is the change of log odds of observing reproduction status to be 1. As usual, we will not make an inference from the intercept. In case-control data, the intercept is inestimable.

The output of **glm()** contains information similar to what has been seen in the output of **lm()** in the linear regression model.

3.4 Use of glm() and Annotations

We use the motivational example to illustrate the setup of **glm()** and the interpretation of the output.

```
## Fit the logistic regression
mymodel =glm(IndReprod ~ Longevity,   # model formula, the response variable is on the left side.
         family=binomial,             # this must be specified since the response is binary!
         data=fruitflies)             # data frame - the variables in the model MUST identical
                                      # to the ones in the data frame!!!
```



```
glm(formula = IndReprod ~ Longevity, family = binomial, data = fruitflies)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7548  -0.6794  -0.1583   0.5525   2.0535

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.39071    1.72165  -3.712 0.000206 ***
Longevity    0.12605    0.03358   3.753 0.000175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.908  on 48  degrees of freedom
Residual deviance: 39.975  on 47  degrees of freedom
AIC: 43.975

Number of Fisher Scoring iterations: 5
```

The only inforamtion we need for this class.

Figure 3: The complete output of glm()

The output has four major pieces of information: model formula, summary of the deviance residuals, significant test results of the predictor variables, and goodness-of-fit measures. In this course, we only focus on the significance tests.

3.5 Applications of Logistic Regression Models

Like other regression models, logistic regression models have two basic types of applications: association analysis and prediction analysis. Since the response variable in logistic regression is binary (i.e., it has two possible distinct values). Depending on applications, the pair of the two possible distinct values could be `success` vs `failure`, `disease` vs `disease-free`, `acceptance` vs `rejection`, etc. For convenience, we numerically encode the value of the interest as **1** and **0** for the other value. For example, if we are interested in studying the factors associated with the disease (status), then the numerical encoding is

$$Y = \begin{cases} 1, & \text{if diseased} \\ 0, & \text{if disease-free} \end{cases}.$$

The logistic regression model is explicitly given by

$$P(Y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}.$$

3.5.1 Association Analysis

The association analysis focuses on the interpretation of the regression coefficients that have information about whether predictor variables are associated with the response variable through the probability of the **bigger** value of the response variable. The fundamental idea of logistic association analysis is to see whether a predictor variable is associated with the probability $P(Y = \text{bigger value})$.

To be more specific, if the coefficient of a predictor variable is significant (i.e., the p-value < 0.05 at a 5% level of significance), we say the predictor variable is significantly associated with the probability. This is the primary interest in classical statistical inference. In the above example, the p-value associated with **Longevity** is $0.000175 < 0.05$ which implies that **Longevity** and the probability of **being deprived of the opportunity**.

3.5.2 Predictive Analysis

Since the logistic regression builds the relationship between the probability of observing the **bigger** value of the response and the predictor variable, the predicted value of the logistic regression is the probability of observing the **bigger** value of the response. The practical question is to predict **actual value** of the response variable. For example, in the above example, if we are given the longevity of a fruit fly, we want to predict whether the fruit fly is **being deprived of the opportunity**. That is, we want to predict the value of the response but not the probability of observing the **bigger** value of the response. This is the primary type of question in most data science projects (i.e., machine learning projects).

The issue is that predicting the **value of the response variable** requires a cut-off probability to assign a value to the response variable. The prediction process of a logistic regression model is depicted in the following figure.

4 Case Studies

We present two examples in this section.

4.1 The simple logistic regression model

Suzuki et al. (2006) measured sand grain size on 28 beaches in Japan and observed the presence or absence of the burrowing wolf spider Lycosa ishikariana on each beach. Sand grain size is a measurement variable, and spider presence or absence is a nominal variable. Spider presence or absence is the dependent variable;

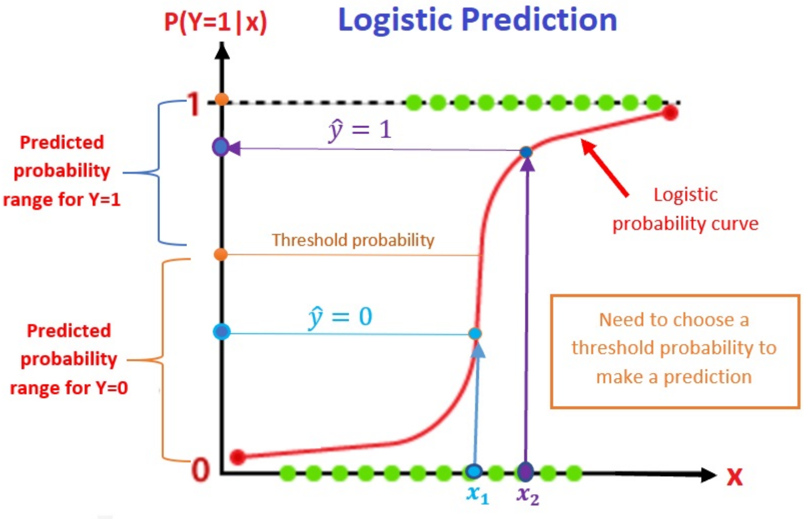
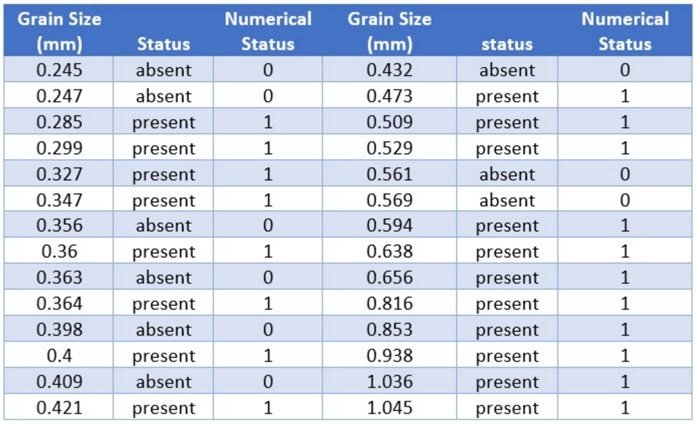

P(Y=1|x)
Logistic Prediction
1
Predicted probability range for Y=1
$\hat{y} = 1$
Logistic probability curve
Threshold probability
Need to choose a threshold probability to make a prediction
$\hat{y} = 0$
Predicted probability range for Y=0
0
x_1
x_2
X

Figure 4: Prediction process of the logistic regression models

if there is a relationship between the two variables, it would be sand grain size affecting spiders, not the presence of spiders affecting the sand.



Grain Size (mm)	Status	Numerical Status	Grain Size (mm)	status	Numerical Status
0.245	absent	0	0.432	absent	0
0.247	absent	0	0.473	present	1
0.285	present	1	0.509	present	1
0.299	present	1	0.529	present	1
0.327	present	1	0.561	absent	0
0.347	present	1	0.569	absent	0
0.356	absent	0	0.594	present	1
0.36	present	1	0.638	present	1
0.363	absent	0	0.656	present	1
0.364	present	1	0.816	present	1
0.398	absent	0	0.853	present	1
0.4	present	1	0.938	present	1
0.409	absent	0	1.036	present	1
0.421	present	1	1.045	present	1

Figure 5: Spider Data Table

```
grainsize=c(0.245, 0.247, 0.285, 0.299, 0.327, 0.347, 0.356, 0.360, 0.363, 0.364,
            0.398, 0.400, 0.409, 0.421, 0.432, 0.473, 0.509, 0.529, 0.561, 0.569,
            0.594, 0.638, 0.656, 0.816, 0.853, 0.938, 1.036, 1.045)
status=c(0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1,
         1, 1, 1, 1)
spider = as.data.frame(cbind(grainsize = grainsize, status = status))
```

Fitting a Simple Logistic Regression Model

Since there is only one predictor variable in this study, simply choose the simple linear regression model for this data set.

```
spider.model = glm(status ~ grainsize,
                   family = binomial,
                   data = spider)
```

```
significant.tests = summary(spider.model)$coef
kable(significant.tests, caption = "Summary of the significant tests of
      the logistic regression model")
```

Table 1: Summary of the significant tests of the logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.647625	1.354191	-1.216686	0.2237237
grainsize	5.121553	3.006174	1.703678	0.0884413

Association Analysis

The above significant tests indicate that the grain size does not achieve significance (p-value = 0.08844) at level 0.05. Note that the p-value is calculated based on the sample, it is also a random variable. Moreover, the sample size in this study is relatively small. We will claim the association between the two variables. As the grain size increases by one unit, the log odds of observing the wolf spider burrowing increase by 5.121553. In other words, the grain size and the presence of spiders are positively associated.

Prediction Analysis

As an example, we choose two new grain sizes 0.33 and 0.57, and want to predict the presence of the wide spiders on the beaches with the given grain sizes. We used the R function **predict()** in linear regression, we used the same function to make predictions in the logistic regression model.

```
spider.model = glm(status ~ grainsize,
                   family = binomial,
                   data = spider)
##
mynewdata = data.frame(grainsize=c(0.275, 0.57))
pred.prob = predict(spider.model, newdata = mynewdata,
        type = "response")
## threshold probability
cut.off.prob = 0.5
pred.response = ifelse(pred.prob > cut.off.prob, 1, 0)  # This predicts the response
pred.table = cbind(new.grain.size = c(0.275, 0.57), pred.response = pred.response)
kable(pred.table, caption = "Predicted Value of response variable
      with the given cut-off probability")
```

Table 2: Predicted Value of response variable with the given cut-off probability

new.grain.size	pred.response
0.275	0
0.570	1

4.2 Multiple Logistic Regression Model

In this case study, we used a published study on bird introduction in New Zealand. The objective is to predict the success of avian introduction to New Zealand. Detailed information about the study can be found in the following article. https://pengdsci.github.io/STA551/w04/Correlates-of-introduction-success-in-exotic.pdf. The data is included in the article. A text format data file was created and can be downloaded or read directly from the following URL: https://pengdsci.github.io/STA551/w06/img/birds-data.txt.

The response variable: Status - status of success Predictor variables:

- length - female body length (mm)
- mass = female body mass (g)
- range = geographic range (% area of Australia)
- migr = migration score: 1= sedentary, 2 = sedentary and migratory, 3 = migratory
- insect = the number of months in a year with insects in the diet
- diet = diet score: 1 = herbivorous; 2 = omnivorous, 3 = carnivorous.
- clutch = clutch size
- broods = number of broods per season
- wood = use as woodland scored as frequent(1) or infrequent(2)
- upland = use of the upland as frequent(1) or infrequent(2)
- water = use of water scored as frequent(1) or infrequent(2)
- release = minimum number of release events
- indiv = minimum of the number of individuals introduced.

We next read the data from the given URL directly to R. Since there are some records with missing values. We drop those records with at least one missing value.

There are several categorical variables coded in numerical form. Among them, **migr** and **diet** have three categories and the rest of the categorical variables have two categories. In practice, a **categorical variable with more than two categories must be specified as factor variables** so R can define dummy variables to capture the difference across the difference.

We conducted an exploratory analysis on **nigr** and **diet** and found a flat discrepancy across the effect. We simply treat them as discrete numerical variables using the numerical coding as the values of the variables.

```
NZbirds=read.table("https://pengdsci.github.io/STA551/w06/img/birds-data.txt", header=TRUE)
birds = na.omit(NZbirds)
```

- **Build an Initial Model**

We first build a logistic regression model that contains all predictor variables in the data set. This model is usually called the full model. Note that the response variable is the success status (1 = success, 0 = failure). Species is a kind of ID, it should not be included in the model.

```
initial.model = glm(status ~ length + mass + range + migr + insect + diet + clutch + broods +
                      wood + upland + water + release + indiv, family = binomial, data = birds)
coefficient.table = summary(initial.model)$coef
kable(coefficient.table, caption = "Significance tests of logistic regression model")
```

Table 3: Significance tests of logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.3380099	5.7167615	-1.1086714	0.2675720
length	-0.0028150	0.0053169	-0.5294328	0.5965052

	Estimate	Std. Error	z value	Pr(>|z|)
mass	0.0026679	0.0016740	1.5937709	0.1109874
range	-0.1316071	0.3502344	-0.3757688	0.7070888
migr	-2.0435447	1.1158239	-1.8314222	0.0670376
insect	0.1479915	0.2123645	0.6968752	0.4858809
diet	2.0285053	1.8832013	1.0771580	0.2814097
clutch	0.0793804	0.2683046	0.2958594	0.7673375
broods	0.0217705	0.9283268	0.0234513	0.9812903
wood	2.4902098	1.6416010	1.5169397	0.1292819
upland	-4.7134743	2.8648273	-1.6452909	0.0999098
water	0.2349442	2.6701934	0.0879877	0.9298864
release	-0.0129156	0.1932112	-0.0668472	0.9467033
indiv	0.0159265	0.0083240	1.9133152	0.0557077

The p-values in the above significance test table are all bigger than 0.5. We next search for the best model by dropping some of the insignificant predictor variables. Since there are so many different ways to drop variables, next we use an automatic variable procedure to search the final model.

- **Automatic Variable Selection**

R has an automatic variable selection function **step()** for searching the final model. We will start from the initial model and drop insignificant variables using AIC as an inclusion/exclusion criterion.

In practice, sometimes, there may be some practically important predictor variables. Practitioners want to include these practically important variables in the model regardless of their statistical significance. Therefore we can fit the smallest model that includes only those practically important variables. The final model should be **between** the smallest model, which we will call a **reduced model**, and the initial model, which we will call a **full model**. For illustration, we assume **insect** and **range** are practically important, we want to include these two variables in the final model regardless of their statistical significance.

In summary, we define two models: the full model and the reduced model. The final best model will be the model between the full and reduced models. The summary table of significant tests is given below.

```
full.model = initial.model  # the *biggest model* that includes all predictor variables
reduced.model = glm(status ~ range + insect , family = binomial, data = birds)
final.model =  step(full.model,
                    scope=list(lower=formula(reduced.model),upper=formula(full.model)),
                    data = birds,
                    direction = "backward",
                    trace = 0)   # trace = 0: suppress the detailed selection process
final.model.coef = summary(final.model)$coef
kable(final.model.coef , caption = "Summary table of significant tests")
```

Table 4: Summary table of significant tests

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.4376339	2.1379440	-1.6079158	0.1078536
mass	0.0019393	0.0007326	2.6470676	0.0081193
range	0.0000141	0.3098265	0.0000456	0.9999636
migr	-2.0239952	0.9603017	-2.1076659	0.0350599
insect	0.2704685	0.1425911	1.8968119	0.0578528
wood	1.9489529	1.3174130	1.4793789	0.1390391
upland	-4.7306337	2.0981447	-2.2546746	0.0241538
indiv	0.0138120	0.0040579	3.4037151	0.0006648

- **Interpretation - Association Analysis**

The summary table contains the two practically important variables **range** and **insect**. **range** does not achieve statistical significance (p-value ≈ 1) and **insect** is slightly higher than the significance level of 0.005. Both variables are seemingly positively associated with the response variable.

The following interpretation of the individual predictor variable assumes other life-history variables and the introduction effort variable.

migr and **upland** are negatively associated with the response variable. The odds of success in introducing migratory birds are lower than the sedentary birds. Similarly, birds using upland infrequently have lower odds of being successfully introduced than those using upland frequently.

insect is significant (p-value =0.058). The **odds of success** increase as the number of months of having insects in the diet increases.

mass and **indiv** are positively associated with the response variable. The odds of success increase and the body mass increases Similarly, the odds of success increase as the number of minimum birds of the species increases.

wood does not achieve statistical significance but seems to be positively associated with the response variable.

- **Predictive Analysis**

As an illustration, we use the final model to predict the status of successful introduction based on the new values of the predictor variables associated with two species. See the numerical feature given in the code chunk.

```
mynewdata = data.frame(mass=c(560, 921),
                       range = c(0.75, 1.2),
                       migr = c(2,1),
                       insect = c(6, 12),
                       wood = c(1,1),
                       upland = c(0,1),
                       indiv = c(123, 571))
pred.success.prob = predict(final.model, newdata = mynewdata, type="response")
##
## threshold probability
cut.off.prob = 0.5
pred.response = ifelse(pred.success.prob > cut.off.prob, 1, 0)  # This predicts the response
## Add the new predicted response to Mynewdata
mynewdata$Pred.Response = pred.response
##
kable(mynewdata, caption = "Predicted Value of response variable
      with the given cut-off probability")
```

Table 5: Predicted Value of response variable with the given cut-off probability

mass	range	migr	insect	wood	upland	indiv	Pred.Response
560	0.75	2	6	1	0	123	0
921	1.20	1	12	1	1	571	1

The predicted status of the successful introduction of the two species is attached to the two new records.