# GLMs for proportion data: Binomial errors

**Practical 3**

David Orme

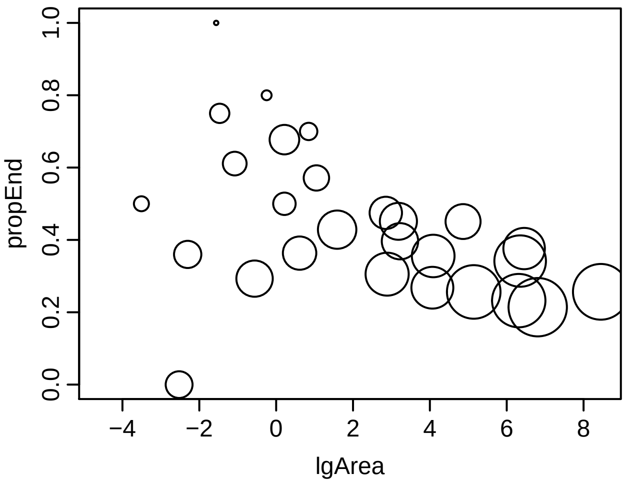
We introduced generalised linear models (GLM) using count data and the Poisson error structure. We're now going to look at using a different error structure that is important for fitting proportional data or binary data. One important thing to note is makes use of proportion data where we have the number of successes out of a number of trials (binomial data). This tells us a lot about the variance of the expected proportion and the weight we assign to different data points.

## Endemicity on the Galapagos Islands

We'll introduce the model using a simple reanalaysis of the species richness data from the Galapagos. As well as the number of plant species, we also have a record of the number of endemic species on each island. This gives a binomial estimate of the proportion of endemics on each island. The question is – does the proportion of endemic species vary with island area?

```
odonata <- read.delim('gala.txt')
```

```
str(gala)

## 'data.frame': 30 obs. of  7 variables:
##  $ Species  : int  58 31 3 25 2 18 24 10 8 2 ...
##  $ Endemics : int  23 21 3 9 1 11 0 7 4 2 ...
##  $ Area     : num  25.09 1.24 0.21 0.1 0.05 ...
##  $ Elevation: int  346 109 114 46 77 119 93 168 71 112 ...
##  $ Nearest  : num  0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
##  $ Scruz    : num  0.6 26.3 58.7 47.4 1.9 ...
##  $ Adjacent : num  1.84 572.33 0.78 0.18 903.82 ...

gala$propEnd <- gala$Endemic / gala$Species
gala$lgArea    <- log(gala$Area)
plot(propEnd ~ lgArea , data=gala, cex=log(Species/2))
```



The figure above uses the plot character size (controlled using cex for character expansion) to show the number of species behind each estimated proportion. It looks like you get fewer endemics on larger islands, but is this significant?

For binomial data, we need to change the error family and link function to family=**binomial**(link=logit). Again, the logit link is the default, so we could omit it and just use family=binomial. For binomial models, we also have to let the model know what the number of species (the *binomial denominator*) are as well as the proportion endemic. There are two ways of doing this in R. One is to create a response variable as a matrix with two columns showing the number of successes (endemic species) and number of failures (in this case, non-endemic species) so that the row sums give the total number of species.

```
resp <- with(gala, cbind(Endemics, Species-Endemics))
galaMod <- glm(resp ~ lgArea, data=gala, family=binomial(link=logit))
```

Another, which I think is easier, is to give the proportion as the response variable and the total number of species using the weights option.

```
galaMod <- glm(propEnd ~ lgArea, weights=Species,
               data=gala, family=binomial(link=logit) )
```

As for the Poisson GLM, the binomial GLM has an analysis of deviance table and uses a $\chi^2$ test on the change in deviance rather than an $F$ test.

```
anova(galaMod, test='Chisq')

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: propEnd
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                      29        154
## lgArea  1     44.1        28        110  3.2e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

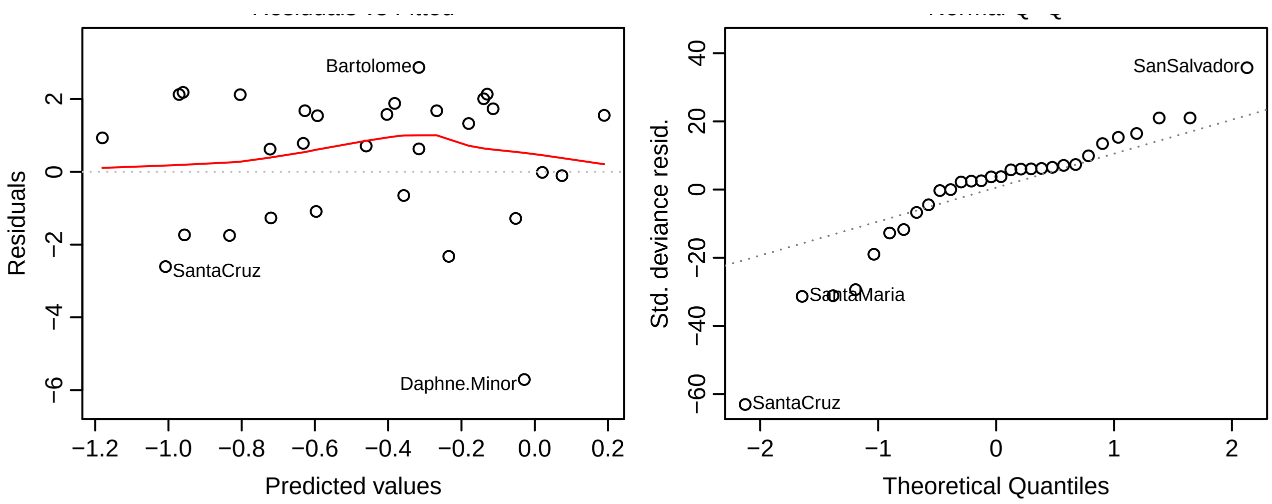
```
1-pchisq(44.053,1)

## [1] 3.2e-11
```

And again – just as in Practical 3 on poisson GLMs for count data, we can look at the model coefficients and deviance explained.

```
summary(galaMod)

##
## Call:
## glm(formula = propEnd ~ lgArea, family = binomial(link = logit),
##     data = gala, weights = Species)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.709  -0.979   0.858   1.720   2.870
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.2936     0.0883   -3.32  0.00089 ***
## lgArea       -0.1049     0.0158   -6.65  2.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.39  on 29  degrees of freedom
## Residual deviance: 110.33  on 28  degrees of freedom
## AIC: 223
##
## Number of Fisher Scoring iterations: 4

(galaMod$null.deviance - galaMod$deviance)/galaMod$null.deviance

## [1] 0.285
```

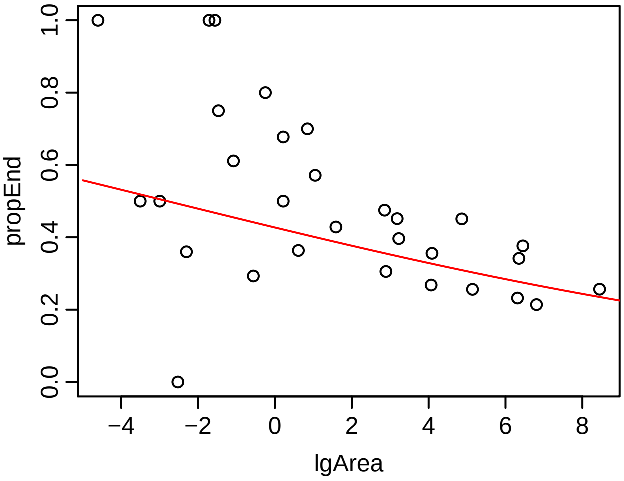
The diagnostic plots for this model aren't wonderful – we won't worry about this now.

```
par(mfrow=c(1,2))
plot(galaMod, which=c(1,2))
```



As for the poisson data in the last practical, we can work out the predicted values on the proportion scale the easy way:

```
# predict for a neat sequence of log area values
pred <- expand.grid(lgArea = seq(-5, 9, by=0.1))
pred$fit <- predict(galaMod, newdata=pred, type='response')

# plot the logged data and the model lines
plot(propEnd ~ lgArea, data=gala)
lines(fit ~ lgArea, data=pred, col='red')
```



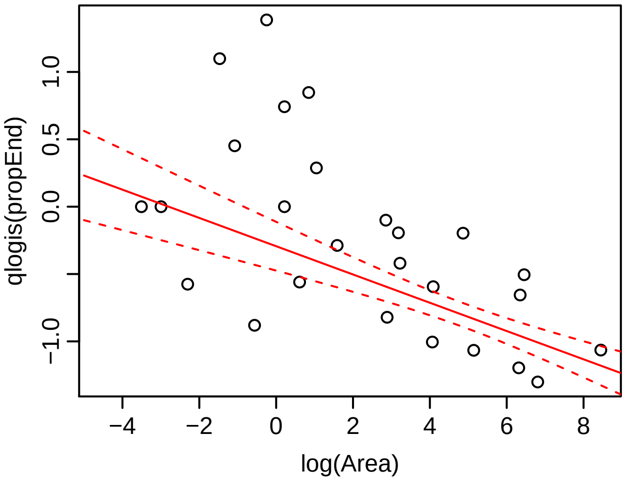
To get confidence limits, we need to use the inverse link function again. For the logit link, the two functions are:

$$\text{logit link} = \log\left(\frac{p}{1-p}\right), \qquad \text{logit link inverse} = \frac{e^x}{1+e^x}$$

There are handy functions to do these conversions: **qlogis**() converts proprotions to logit values and **plogis**() converts logit predictions back to proportions.

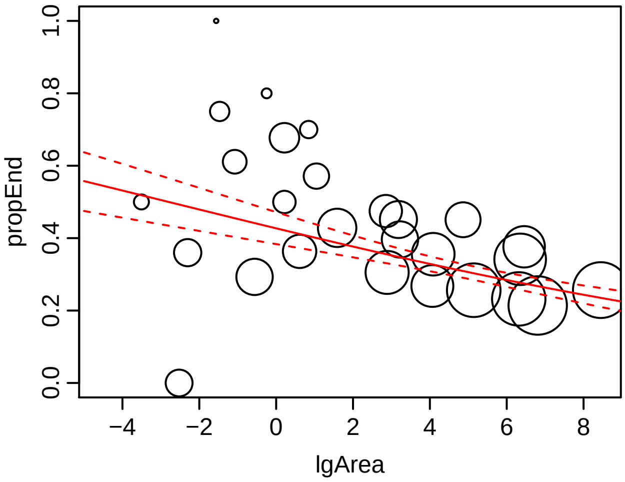
```
# predict for a neat sequence of log area values
pred <- expand.grid(lgArea = seq(-5, 9, by=0.1))
predMod <- predict(galaMod, newdata=pred, se.fit=TRUE)
```

```
# get the fit and confidence limits
pred$fit <- predMod$fit
pred$se.fit <- predMod$sefit
pred$confint <- predMod$se.fit * qt(0.975, df=galaMod$df.residual)

# plot the logit transformed proportion data and the model lines
plot(qlogis(propEnd) ~ log(Area), data=gala)
lines(fit ~ lgArea, data=pred, col='red')
lines(fit + confint  ~ lgArea, data=pred, col='red', lty=2)
lines(fit - confint  ~ lgArea, data=pred, col='red', lty=2)
```



We can now back transform them on to the data. This means we need to know the inverse link function for our model, which for **log**() is **exp**().

```
# plot the proportion data
plot(propEnd ~ lgArea , data=gala, cex=log(Species/2))
# add the link inverse transformed lines
lines(plogis(fit) ~ lgArea, data=pred, col='red')
lines(plogis(fit + confint)  ~ lgArea, data=pred, col='red', lty=2)
lines(plogis(fit - confint)  ~ lgArea, data=pred, col='red', lty=2)
```



## Predicting threat in Galliformes

We're now going to look at a model with binary data – going back to the galliform data set, we're going to test if life history predicts whether species are threatened or not.

```
galliformes <- read.csv('galliformesData.csv')
```

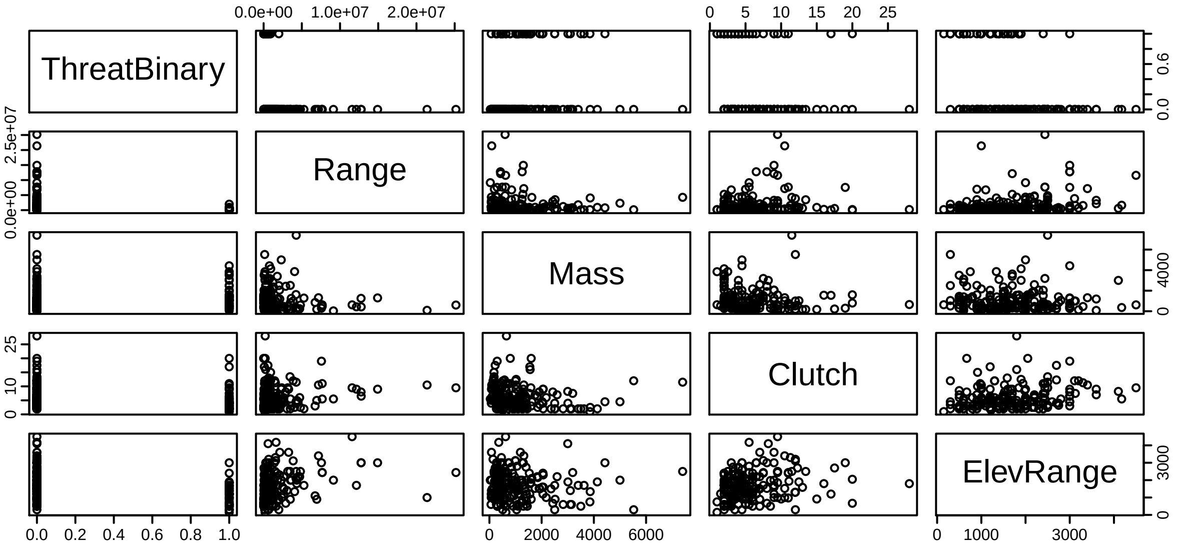
```
str(galliformes)

## 'data.frame': 268 obs. of  8 variables:
##  $ Family   : Factor w/ 6 levels "Cracidae","Megapodiidae",..: 1 2 2 5 4 4 5 5 5 5 ...
##  $ CName    : Factor w/ 268 levels "Aceh Pheasant",..: 251 249 41 71 258 19 10 66 193 203 ...
##  $ SName    : Factor w/ 268 levels "Aburria aburri",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Status04 : Factor w/ 5 levels "1 (LC)","2 (NT)",..: 2 1 3 3 3 1 1 1 1 1 ...
##  $ Range    : int  778240 205731 2571 202631 315160 658866 1542970 11577357 446540 345484 ...
##  $ Mass     : num  1423 1600 1600 1418 815 ...
##  $ Clutch   : num  3.25 20 NA 2.5 12 NA 11.3 9.5 11 12.3 ...
##  $ ElevRange: int  2000 2050 1000 1200 NA NA 3300 4500 2400 2700 ...

# drop incomplete rows in the data
galliformes <- na.omit(galliformes)
```

The first thing to do is to convert the threat status column from IUCN categories to a simple threatened (1) or not threatened (0) numeric variable.

```
galliformes$ThreatBinary <- ifelse(galliformes$Status04 %in% c("1 (LC)", "2 (NT)"), 0, 1)
```

The life history variables we'll use are body mass, geographic range, clutch size and elevational range.  If we check those, they all show strong right skew – the points are all clumped over to the left:

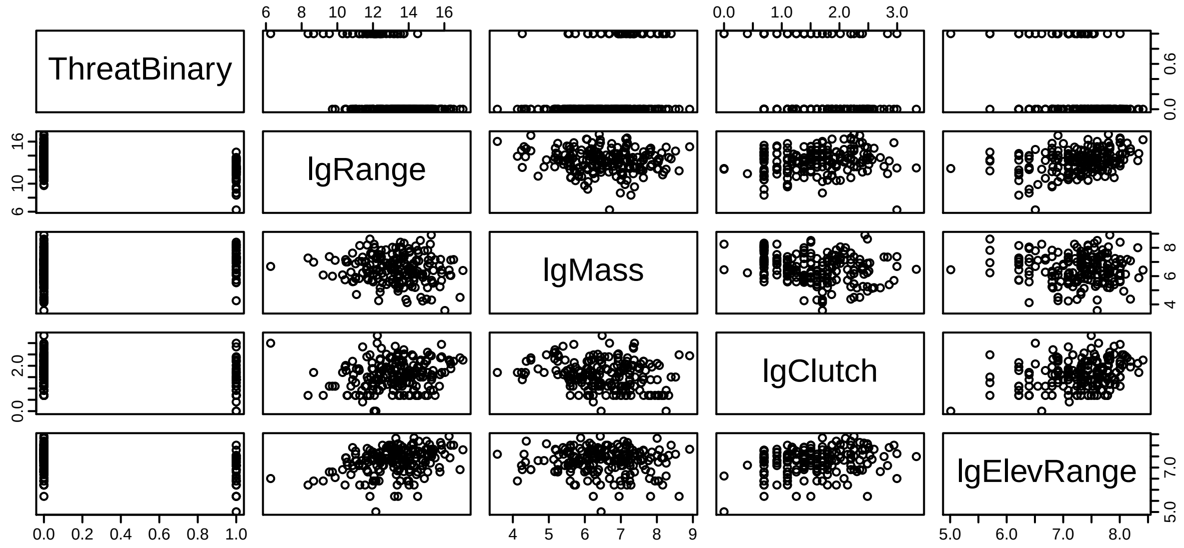
```
pairs(ThreatBinary ~ Range + Mass + Clutch + ElevRange, data=galliformes)
```



So, log transformation[1] to the rescue once again.

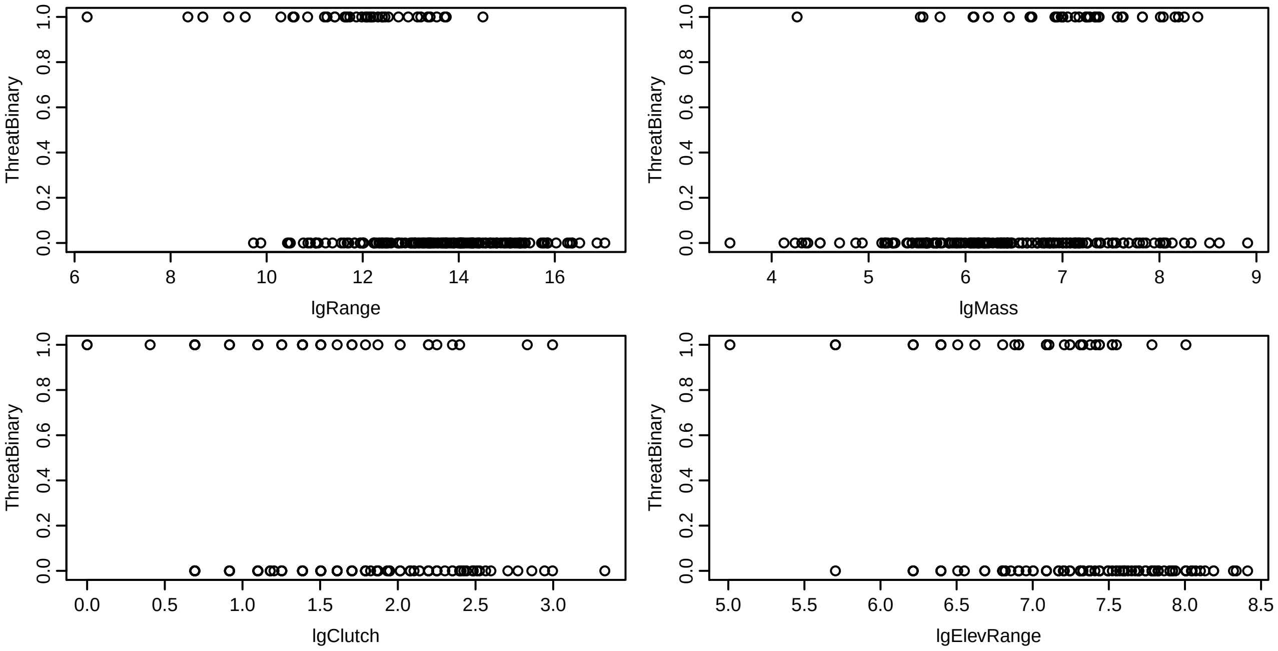
[1] That seems to happen a lot in these examples!

```
galliformes$lgMass <-  log(galliformes$Mass)
galliformes$lgRange <-  log(galliformes$Range)
galliformes$lgClutch <-  log(galliformes$Clutch)
galliformes$lgElevRange <-  log(galliformes$ElevRange)

pairs(ThreatBinary ~ lgRange +lgMass + lgClutch + lgElevRange, data=galliformes)
```



Now we can look at some plots of the relationships to be modelled. Binary plots are hard to read but you can see that for most of those variables, the ones and zeros are concentrated in different places along the $x$ axis: as the variables change, the probability of being threatened changes.

```
par(mfrow=c(2,2))
plot(ThreatBinary ~ lgRange, data=galliformes)
plot(ThreatBinary ~ lgMass, data=galliformes)
plot(ThreatBinary ~ lgClutch, data=galliformes)
plot(ThreatBinary ~ lgElevRange, data=galliformes)
```



Now the hard bit – and this is a trickier example: try and find the best model to explain threat status, starting with the model below, including all four variables and all two-way interactions. Feel free to use **step**() and **drop1**()! An important point to note is that – for *binary data only* – it isn't necessary to provide weights to the binomial glm. Each data point represents a single case and has the same weight.

```
galliMod <- glm(ThreatBinary ~ (lgRange +lgMass + lgClutch + lgElevRange)^2,
                data=galliformes, family=binomial(link=logit))
```

# Report statistics

## Endemicity on the Galapagos

### Methods

I fitted a Binomial GLM with a logit link to predict the proportion of endemic plants on islands of the Galapagos as a function of island area in km$^2$. The proportion of endemics was weighted within the binomial model by the number of species on each island. I tested model significance using $\chi^2$ tests and assessed model suitability using diagnostic plots of deviance residuals.

### Results

Log island area is highly explanatory of plant endemicity ($\chi^2_1 = 44.053, p << 0.0001$), with a lower proportion of endemics found on larger islands. The model explains $28.5\%$ of the deviance in endemicity. The model is shown in Figure 1 and the model coefficients are shown in Table 1

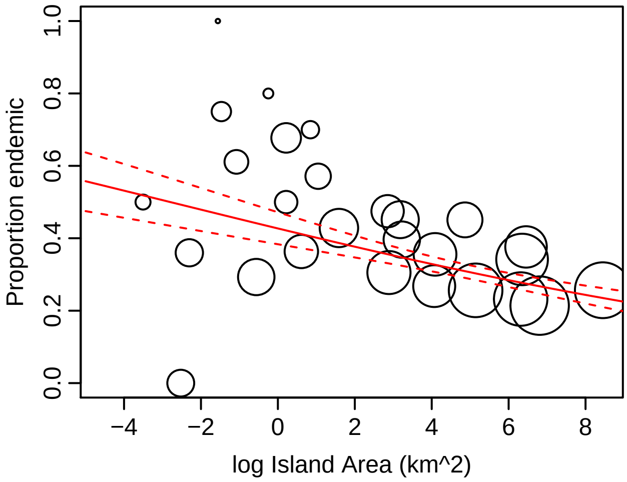


**Figure 1:** *Plant endemicity on the islands of the Galapagos as a function of island area. Predictions of a binomial generalised linear model (solid line) are shown with 95% confidence limits (dashed lines).*

Table 1: Model coefficients, standard errors and signficance tests for a generalised linear model with binomial errors of plant endemicity as a function of island area in the Galapagos.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.2936 | 0.0883 | -3.32 | 0.0009 |
| lgArea | -0.1049 | 0.0158 | -6.65 | <0.0001 |

## Predicting threat in Galliformes

### Methods

I fitted a Binomial GLM with a logit link predicting whether the threat status of galliform species (threatened: VU, EN, CR; not threatened: LC, NT) is predicted from their life history. I started with a maximal model including body mass (g), geographic range (km$^2$), clutch size and elevational range (metres) and all two-way interactions. All four explanatory variables were log transformed before analysis. I initially simplified the model using an AIC based step fit and then used $\chi^2$ tests in analysis of deviance to test marginally significant terms and further simplify the model.

## Results

The final model included only main effects for log geographic range ($\chi^2_1 = 40.7, p < 0.0001$), log body mass ($\chi^2_1 = 13.3, p = 0.0003$) and log elevational range ($\chi^2_1 = 9.4, p = 0.002$): species are more likely to be threatened at small geographic and elevational range and at larger body size (Figure 2, Table 2).

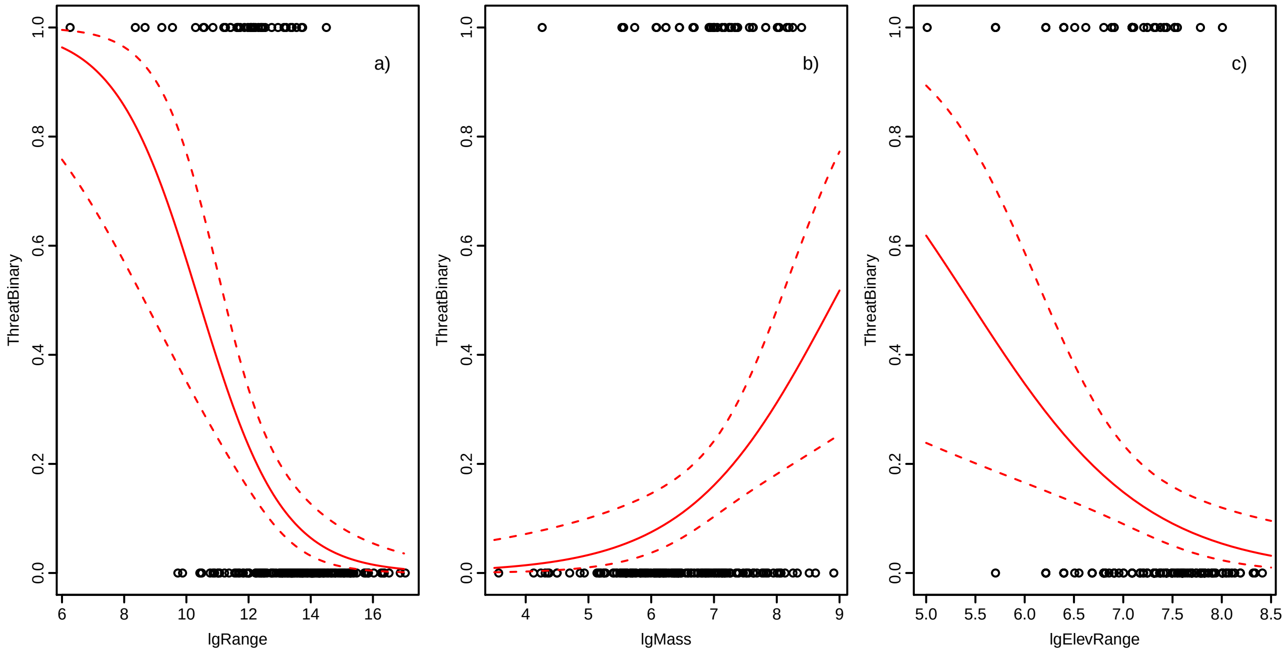


**Figure 2:** *Probability of species threat as a function of a) geographic range size, b) body mass and c) elevational range in galliformes. Predictions and 95% confidence intervals from a binomial model (red) for each variable are shown at the mean log values for the other variables in the model.*

Table 2: Model coefficients, standard errors and signficance tests for a generalised linear model with binomial errors of threat status in galliform birds.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 10.2783 | 3.4248 | 3.00 | 0.0027 |
| lgRange | -0.7438 | 0.1663 | -4.47 | <0.0001 |
| lgMass | 0.8618 | 0.2664 | 3.24 | 0.0012 |
| lgElevRange | -1.1156 | 0.3729 | -2.99 | 0.0028 |