Tutorial #9: Binary and Proportion models - Logistic GLM

Guy F. Sutton

Centre for Biological Control, Rhodes University

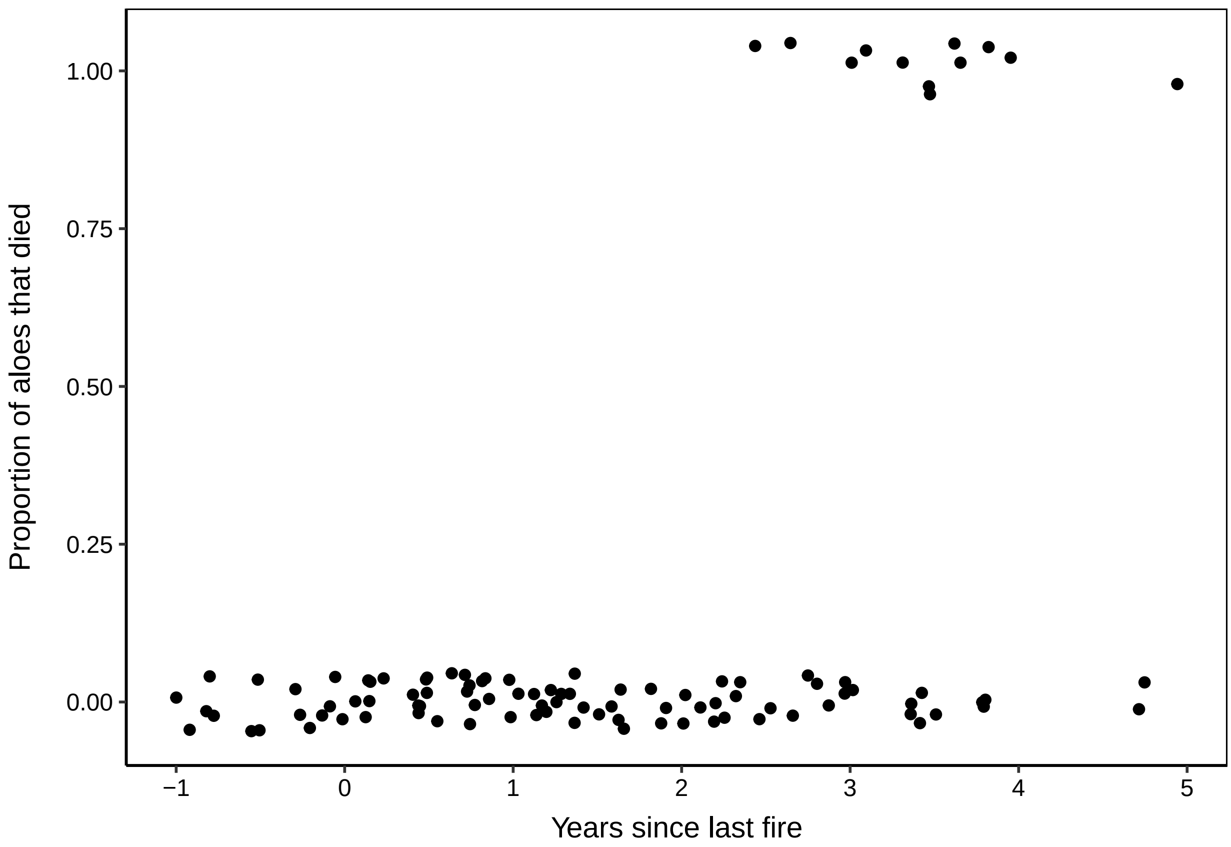
9.1. Analysis of binary or proportion data

In the previous tutorials, we have covered the analysis of normally distributed data (Gaussian GLM) and the analysis of count/abundance data (Poisson and Negative Binomial GLM). However, many ecological studies collect binary data (e.g. Yes/No, Dead/Alive, Present/Absent) or proportion data (e.g. Proportion plant cover, Proportion of insects that died in different treatments). These data usually do not meet the assumptions being made by any of the statistical analyses we have covered so far.

To analyse these types of datasets, a natural starting point is a logistic GLM.

Let's consider a hypothetical study in Kruger National Park where ecologists have established transects along which they count the proportion of an endangered aloe that are dead versus alive (`aloe_alive`) and the number of years since a fire last burned through the transect (`time_since_fire`). The ecologists are interested in knowing whether fire increases the mortality of the endangered aloe.

```
##           x1 y years_since_fire aloe_alive
## 1 -0.1224600 1                0          0
## 2  0.5524566 1                1          0
## 3  0.3486495 1                1          0
## 4  0.3596322 0                1          0
## 5  0.8980537 1                2          0
## 6 -1.9225695 0                4          0
```



Fitting the logistic GLM

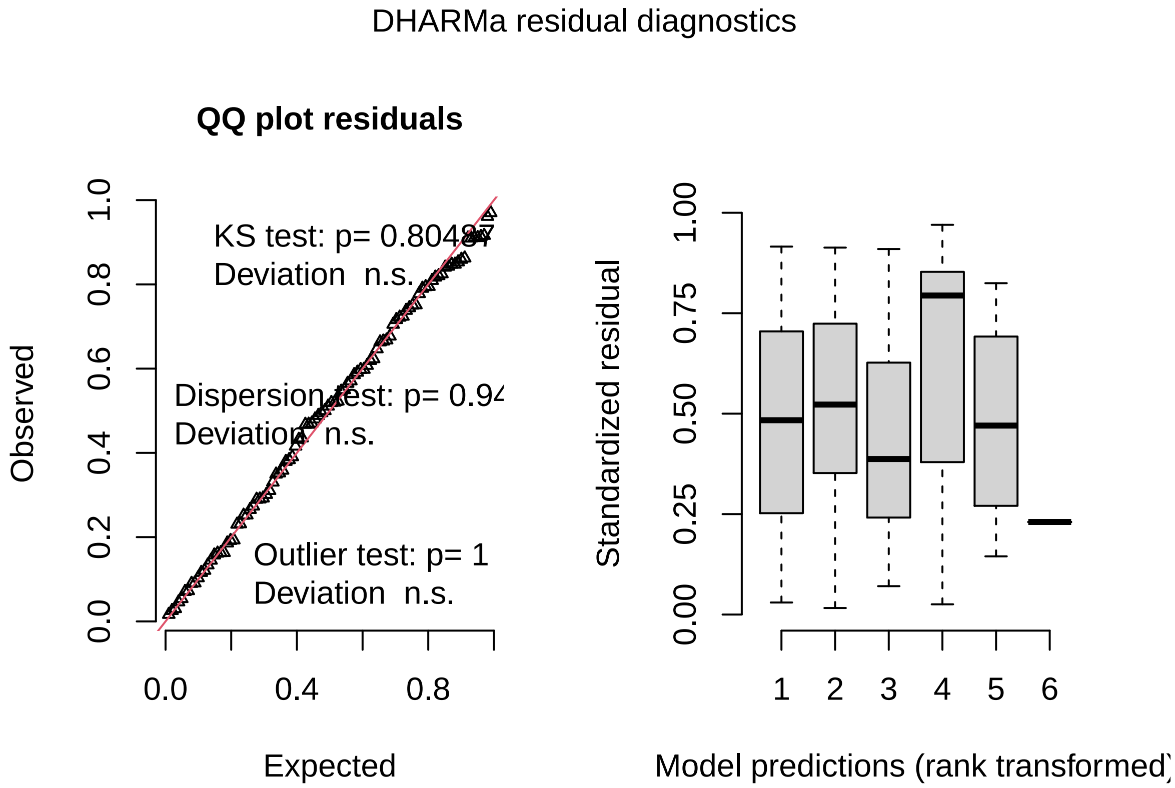
We can go back to the `glm` function, and alter the `family` argument to tell `R` that we want a logistic regression.

```
mod1 <- glm(aloe_alive ~ years_since_fire,
            data = data,
            family = poisson(link = "log"))
```

Evaluate model fit

Evaluating the fit of a logistic GLM (actually any GLM's) is exactly the same as for the Gaussian and Poisson GLM's that we have already covered.

```
mod1_diagnostics <- DHARMa::simulateResiduals(mod1)
plot(mod1_diagnostics)
```



LHS plot (QQplot): This plot tells us whether our data conforms to a binomial data distribution as assumed by a logistic GLM. If our GLM is a good fit to the data, the white triangles will fall approximately on the red 1:1 line and the KS test P-value will be greater than 0.05. The Kolmorogorov-Smirnoff test (KS test) evaluates whether our data follow the distribution we specified in the `family` argument in the GLM call above (remember: we said `family = binomial`).

- Interpreting this plot tells us that the binomial data distribution was a good fit. Note that the points fall approximately along the 1:1 red line and the KS test result was not significant - which tells us that our model approximates a binomial distribution well.

RHS plot (Fitted vs Residuals): This plot tells us whether there are issues with homogeneity of variances. GLM's, like ANOVA and linear regression, assume that the variance in the different groups, or across the range of x values, are approximately equal. We want to see the bold black lines for the different groups fall approximately at y = 0.50 and the range of the boxplots should be similar.

- Interpreting this plot we can see that the bold line falls approximately on the y = 0.50 line, and the range of the boxplots is okay (the x = 6 box is a bit worrisome though). If this was a model for a paper/thesis, I would scrutinize this further. Let's proceed anyway.

Results

Now we get to the bit we are interested in: assessing statistical significance and calculating p-values. To do this, we will use a Likelihood Ratio Test (LRT). When we only have 1 predictor variable (here: `years_since_fire`), we need to calculate p-values using type I sum-of-squares (SOS). SOS's are just different ways that we ask `R` to calculate p-values.

PLEASE DO NOT USE SUMMARY() - THIS WILL PRODUCE THE WRONG P-VALUES WHEN YOU HAVE MORE THAN 1 PREDICTOR VARIABLE

```r
# Perform LRT  with type I sum-of-squares
car::Anova(mod1,
           # Specify we want a Likelihood Ratio Test
           test = "LR")
```

Parameter significance

```
## Analysis of Deviance Table (Type II tests)
##
## Response: aloe_alive
##                  LR Chisq Df Pr(>Chisq)
## years_since_fire   18.342  1  1.846e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the LRT tells us that the number of years since the last fire (`years_since_fire`) was a statistically significant predictor of the proportion of aloes that died (`aloe_alive`) ($\chi^2 = 18.34$, d.f. $= 1$, $P < 0.001$).

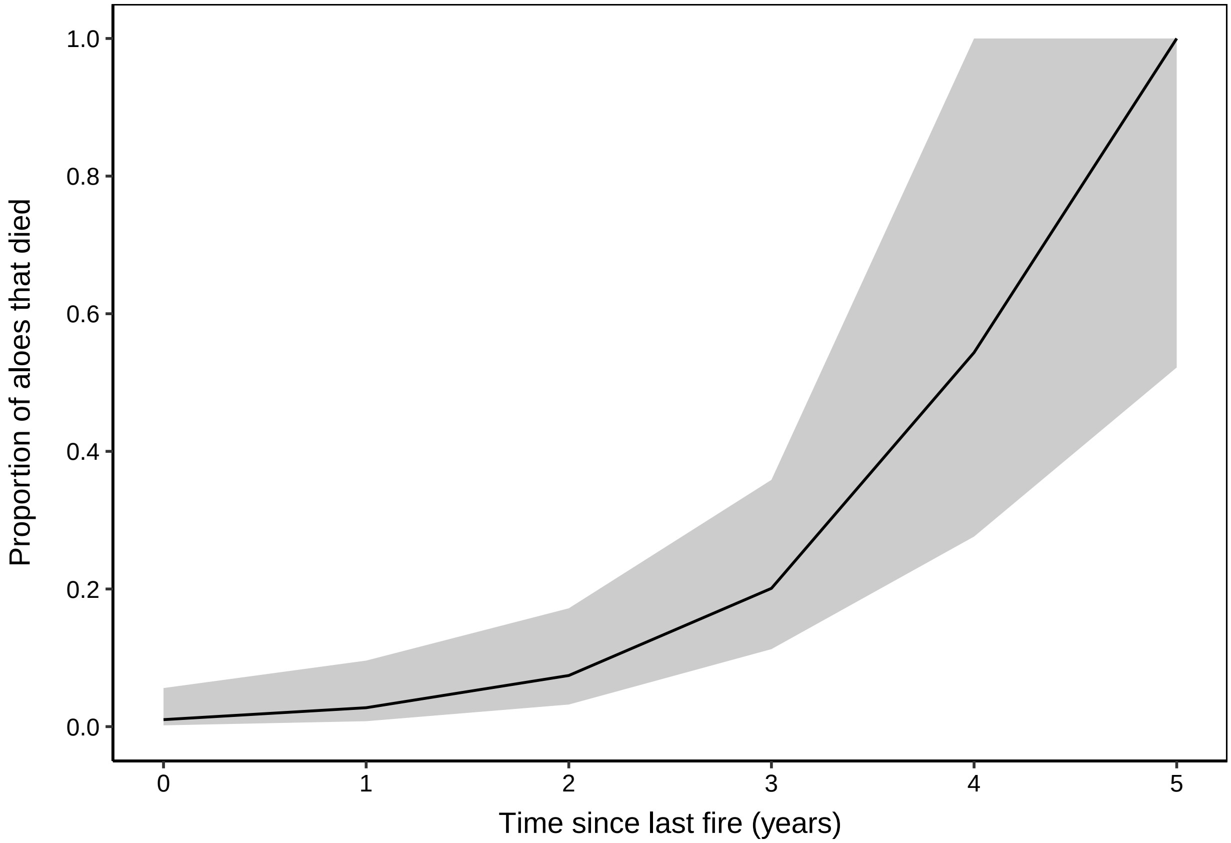
But, did aloe mortality increase or decrease with the time since last fire? We could easily plot the data, or we can specifically calculate the direction and magnitude of the effect.

```r
library(ggeffects)
# Extracted model predictions
preds <- ggpredict(mod1, terms = c("years_since_fire"))
preds <- as.data.frame(preds) %>%
  dplyr::mutate(conf.high = dplyr::case_when(
    conf.high > 1 ~ 1,
    TRUE ~ conf.high)) %>%
  dplyr::mutate(predicted = dplyr::case_when(
    predicted > 1 ~ 1,
    TRUE ~ predicted))
head(preds)
```

Making a plot

```
##   x  predicted std.error    conf.low  conf.high group
## 1 0 0.01014524 0.8716081 0.001838011 0.05599849     1
## 2 1 0.02744856 0.6383646 0.007854935 0.09591721     1
## 3 2 0.07426376 0.4282379 0.032081911 0.17190705     1
## 4 3 0.20092517 0.2955258 0.112585466 0.35858026     1
## 5 4 0.54361539 0.3454749 0.276199565 1.00000000     1
## 6 5 1.00000000 0.5286207 0.521898893 1.00000000     1
```

```
ggplot(data = preds, aes(x = x,
                         y = predicted)) +
  geom_ribbon(aes(ymin = conf.low,
                  ymax = conf.high),
              fill = "grey80") +
  geom_line(aes(y = predicted)) +
  scale_y_continuous(breaks = seq(0, 1, 0.2),
                     limits = c(0, 1)) +
  labs(x = "Time since last fire (years)",
       y = "Proportion of aloes that died")
```


Proportion of aloes that died
Time since last fire (years)