Tutorial #6: A Bad Model

Guy F. Sutton

Centre for Biological Control, Rhodes University

6. What does a poor model fit look like?

In the previous tutorials, all our models have fit the data we are modelling quite well. In this next tutorial, we will see an example of a GLM that is a really poor fit to the data. It is really important that you only assess the results from a model that fits the data well - otherwise the results you obtain and inferences you draw will be misleading, or just plain wrong.

Let's consider a study where we sample the number of birds along transects at 10 different sites, with five sites located in the `grassland` biome and the other five sites located in the `thicket` biome. Our study wants to know whether bird species richness `sp_rich` differs between biomes `biome`.

```
head(data)
```

```
##   sp_rich biome
## 1       3     A
## 2       5     A
## 3       0     A
## 4       6     A
## 5       5     A
## 6       6     A
```

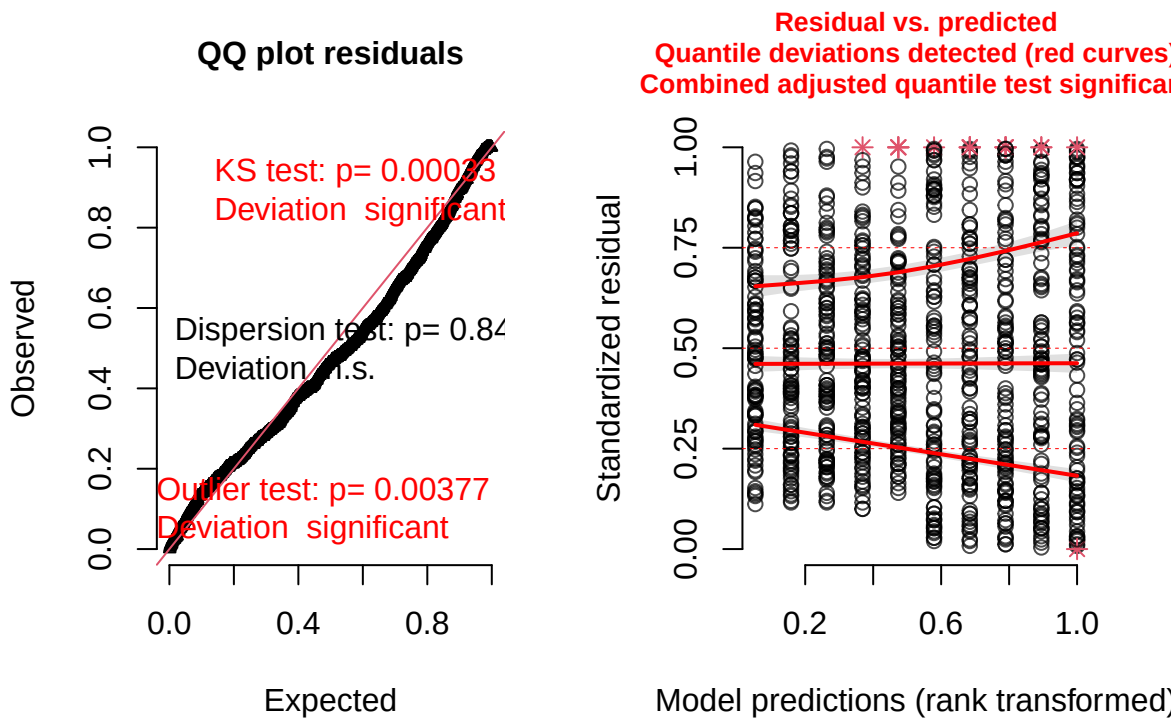
Fitting the model

```
mod1 <- glm(sp_rich ~ biome,
            data = data,
            family = gaussian(link = "identity"))
```

Evaluate model fit

Before we look at the results from our model, we must first check whether the GLM that we fit was an appropriate choice for our data.

```
mod1_diagnostis <- DHARMa::simulateResiduals(mod1)
plot(mod1_diagnostis)
```



LHS plot (QQplot): This plot tells us whether our data conforms to a normal (or gaussian) data distribution. 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 = gaussian").

- Interpreting this plot tells us that the gaussian data distribution was not a good fit. Note that the points fall below the 1:1 red line and the KS test result was significant - which tells us that our model does not approximate a normal distribution.

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 [0.25, 0.50, 0.75]. If there are issues, the bold lines will fall away from the y = [0.25, 0.50, 0.75], and all the lines will be red (not black).

- Interpreting this plot we can see that the bold lines do not fall on the y = [0.25, 0.50, 0.75] lines, as would be expected. The title of the plot tells us (and the red colouration) that there is a significant deviation from equal variances

Take home: The GLM that we fit was not a good fit to the data. We should not proceed with looking and the results - they will be misleading or wrong. We need to use a different data distribution to model our data.