Tutorial #4: Gaussian GLM - additive

Guy F. Sutton

Centre for Biological Control, Rhodes University

4. Gaussian GLM

In the previous tutorials, we learned how to use a gaussian GLM when our predictor variable was either a single categorical (`site`) or continuous (`body_length`) variable. In the next tutorial, we will look at a gaussian GLM when we use both predictor variables in the same model. The coding of the model is exactly the same - the only aspect that changes is the interpretation of the results we get from `R`.

We will still be considering the same hypothetical study where we measure the body mass of 20 insects (`body_mass`), their body lengths (`body_length`), at each of two sites (`site`). Ultimately, we could be interested in asking whether (`body_mass`) is correlated with (`body_length`), and whether this relationship differs between sites (`site`).

```
head(data)
```

```
##    body_mass body_length site
## 1 18.7754002    15.28778    A
## 2 25.5245663    19.07157    A
## 3 23.4864950    14.49438    A
## 4 23.5963224    26.04058    A
## 5 28.9805369    11.87531    A
## 6  0.7743048    20.52689    A
```

4.1. Gaussian GLM - single continuous predictor

Let's consider a simple model where we look at whether the body mass of an insect (`body_mass`) is correlated with insect `body_length` AND if insect `body_mass` differed between `site`. Here, we have a single continuous predictor variable (`body_length`) and a single categorical predictor variable (`site`).

H_0: Larger insects do not weigh more than smaller insects, and there is no difference in insect body mass between sites.

H_1: Larger insects weight more than smaller insects, and there is a difference in insect body mass between sites.

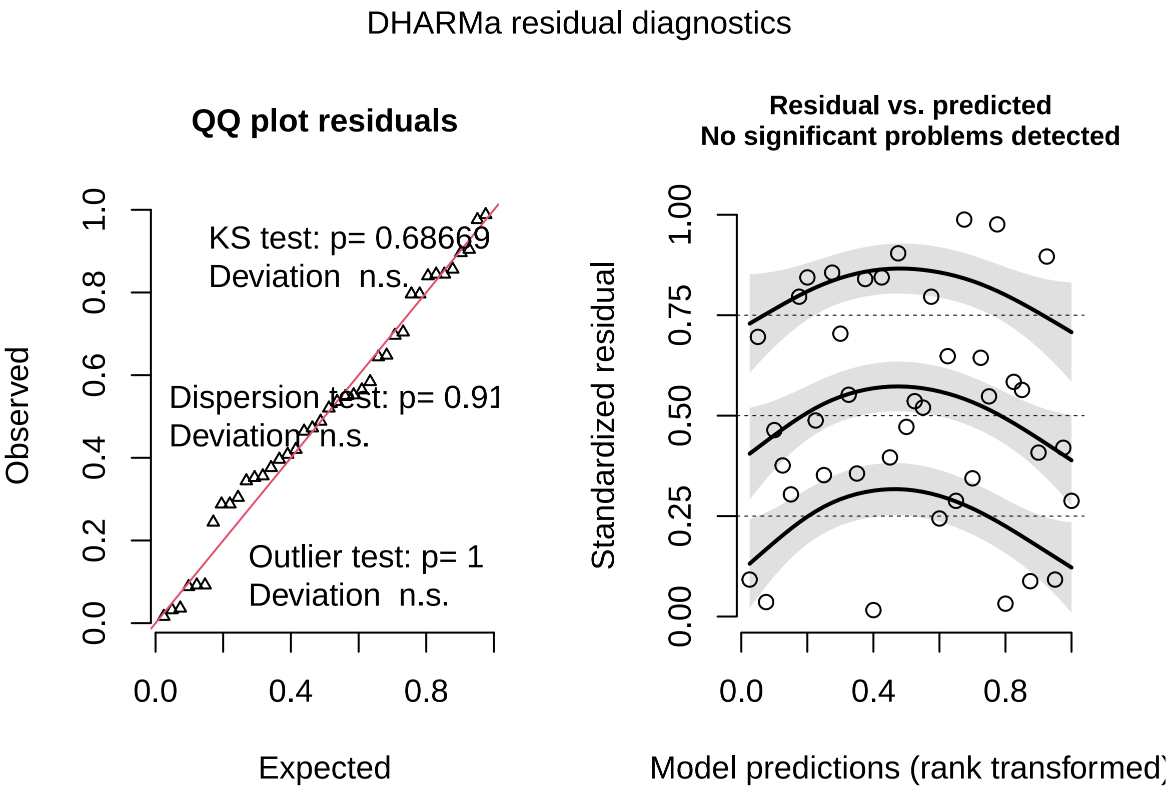
Fitting the model

```
mod1 <- glm(body_mass ~ body_length + site,
            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 seemed appropriate.

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 relatively similar.

- Interpreting this plot we can happily see that the bold black lines for the two boxplots are approximately at y = 0.50 and the range of the boxplots is almost equal.

Take home: The GLM that we fit seems like a good fit to the data. The inferences that we draw and results we obtain should be reasonable.

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 have 2 or more predictor variables (here: `body_length` and `site`), we need to calculate p-values using type II 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",
           # Specify that we want type II SOS
           type = "II")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: body_mass
##             LR Chisq Df Pr(>Chisq)
## body_length   0.0110  1     0.9165
## site          7.6113  1     0.0058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the LRT tells us that there is no evidence for a statistical correlation between insect `body_length` and `body_mass` ($\chi^2 = 0.01$, d.f. $= 1$, P $= 0.917$), but insect `body_mass` is significantly different between `sites` ($\chi^2 = 7.61$, d.f. $= 1$, P < 0.001).

The interpretation of the parameter estimates from `summary` becomes quite tricky when we have more than one predictor variable in our model, and is beyond the scope of this course. Please feel free to drop me an email if you need to explore this option.