# Tutorial #3: Gaussian GLM - continuous predictor X

Guy F. Sutton

Centre for Biological Control, Rhodes University

## 3. Gaussian GLM

In the previous tutorial, we learned how to use a gaussian GLM when our predictor variable is categorical (remember: we looked at whether insect body mass differed between sites). In the next tutorial, we will look at a gaussian GLM when our predictor variable is continous. 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
```

### 3.1. 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`. Here, we have a single continuous predictor variable (`body_length`).

**$H_0$**: Larger insects do not weigh more than smaller insects.

**$H_1$**: Larger insects weight more than smaller insects.

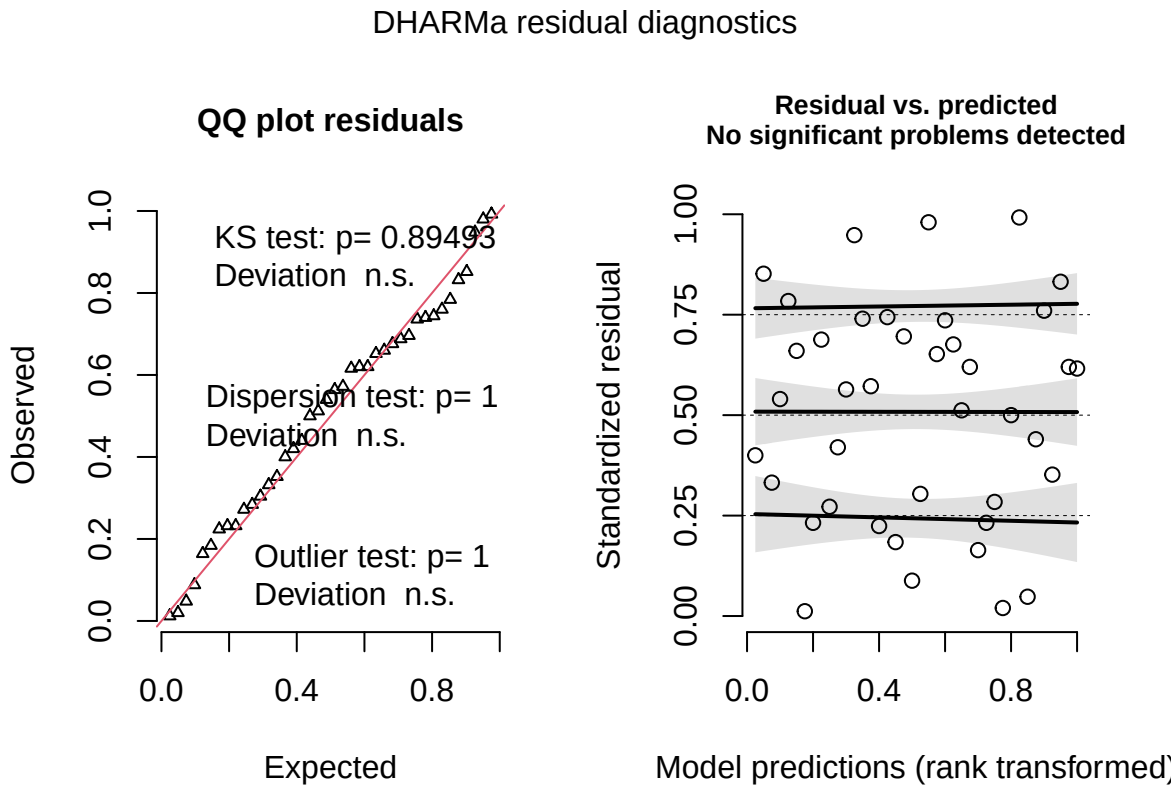
#### Fitting the model

```
mod1 <- glm(body_mass ~ body_length,
            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 only have 1 predictor variable (here: `body_length`), 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**

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

```
## Analysis of Deviance Table (Type II tests)
##
## Response: body_mass
##             LR Chisq Df Pr(>Chisq)
## body_length  0.47183  1     0.4921
```

So, the LRT tells us that there is no evidence for a statistical correlation between insect `body_length` and `body_mass` ($\chi^2 = 0.47$, d.f. $= 1$, P $= 0.49$).

In the next section, we will look at how much insect body mass changes with insect body length. From the LRT above, we should find very little evidence for a relationship below. This step and how we interpret the output below is the major difference from tutorial 2.

```
# Extract parameter estimate
summary(mod1)
```

```
##
## Call:
## glm(formula = body_mass ~ body_length, family = gaussian(link = "identity"),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -28.3150   -7.2688    0.8704    6.9707   26.0653
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.1705     6.4551   5.139  8.6e-06 ***
## body_length  -0.2707     0.3940  -0.687    0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 143.4142)
##
##     Null deviance: 5517.4  on 39  degrees of freedom
## Residual deviance: 5449.7  on 38  degrees of freedom
## AIC: 316.09
##
## Number of Fisher Scoring iterations: 2
```

The estimate for `(Intercept)` is the value of our response variable (`body_mass`) when (`body_length`) = 0. Obviously, this interpretation is nonsensical because you cannot have `body_length` = 0.

The estimate for `body_mass` tells us how much our response variable (`body_mass`) changes for a 1 unit increase in our predictor variable (`body_length`). For example, the estimate here tells us that for every 1 cm increase in insect body length, body mass decreases by 0.27g. If the estimate was 0.27, then for every 1 cm increase in insect body length, body mass would increase by 0.27g.