Basic GLMs in R

Clarke van Steenderen

2024-05-20

R tutorial 5: Running Generalised Linear Models (GLMs)

A GLM is a more versatile version of a linear regression model. It accepts a wider variety of response variable types, such as count and binary data. This is often useful when a data set does not meet the expected statistical assumptions required in the linear modelling process (i.e. Gaussian expectations). When applying linear models, one assumes that the error terms have a normal distribution (i.e. there is a constant relationship between mean and variance). GLMs, however, can be applied when error structures are non-normal.

Data types are numerous, where you will typically have one or more of the following:

- Continuous -> e.g. measurements of height, mass, length, temperature
- Count -> e.g. number of species
- Proportion -> e.g. the proportion of a population that shares a morphological feature
- Binary -> e.g. dead or alive, male or female, healthy or sick
- Categorical/factors -> e.g. different groups, such as a control and drug treatments

Here is a general guide to choosing which statistical test may be best suited for your data. Have a look at Dai Shizuka's site for additional examples. If you have:

- A continuous predictor AND continuous response variable -> use linear regression or a GLM with **Gaussian** distribution
- A continuous predictor AND binary response variable -> GLM with the **"binomial"** family (logistic regression)
- A continuous predictor AND counts as response variable -> GLM with **"Poisson"** or **"Negative binomial"** family (Poisson regression)
- A continuous predictor AND proportions as response variable -> GLM with **"binomial"** family
- A categorical predictor AND continuous response variable -> **ANOVA** (or t-test, if you are just comparing means), which can be also be run as a linear model or GLM. An ANOVA is essentially a LM or a GLM, just with a categorical predictor variable. An ANOVA compares means across groups, and looks at the variance between and within groups. If variance is greater between than within those groups, there is a significant difference between them.
- A categorical predictor AND counts as response variable -> **Chi-square** tests
- Multiple predictors, with some continuous and some categorical variables -> linear regression / GLM

A sum-of-squares test is applied when testing for the significance of your parameters. Generally, you would use a:

- Type I test when there is one predictor variable

- Type II test when there are two or more predictor variables AND NO interaction terms
- Type III test when there are two or more predictor variables AND an interaction term

The default **anova()** function in R runs a type I test, while the **car::Anova()** function can run type II and type III tests. If the interaction term in a type III test is not significant, re-run the test using type II to check the significance of the two predictor variables.

Let's set up our R session:

```
if (!require("pacman"))
  install.packages("pacman")
```

```
## Warning: package 'pacman' was built under R version 4.3.3
```

```
pacman::p_load(xlsx, janitor, ggplot2, Rmisc, dplyr,
               tidyverse, visreg, glmmTMB, gtsummary, effects, ggeffects,
               patchwork)

# Set plot theme
theme_set(theme_classic() +
theme(panel.border = element_rect(colour = "black", fill = NA),
axis.text = element_text(colour = "black"),
axis.title.x = element_text(margin = unit(c(2, 0, 0, 0), "mm")),
axis.title.y = element_text(margin = unit(c(0, 4, 0, 0), "mm")),
legend.position = "none"))
```

A basic GLM -> growth rates across groups

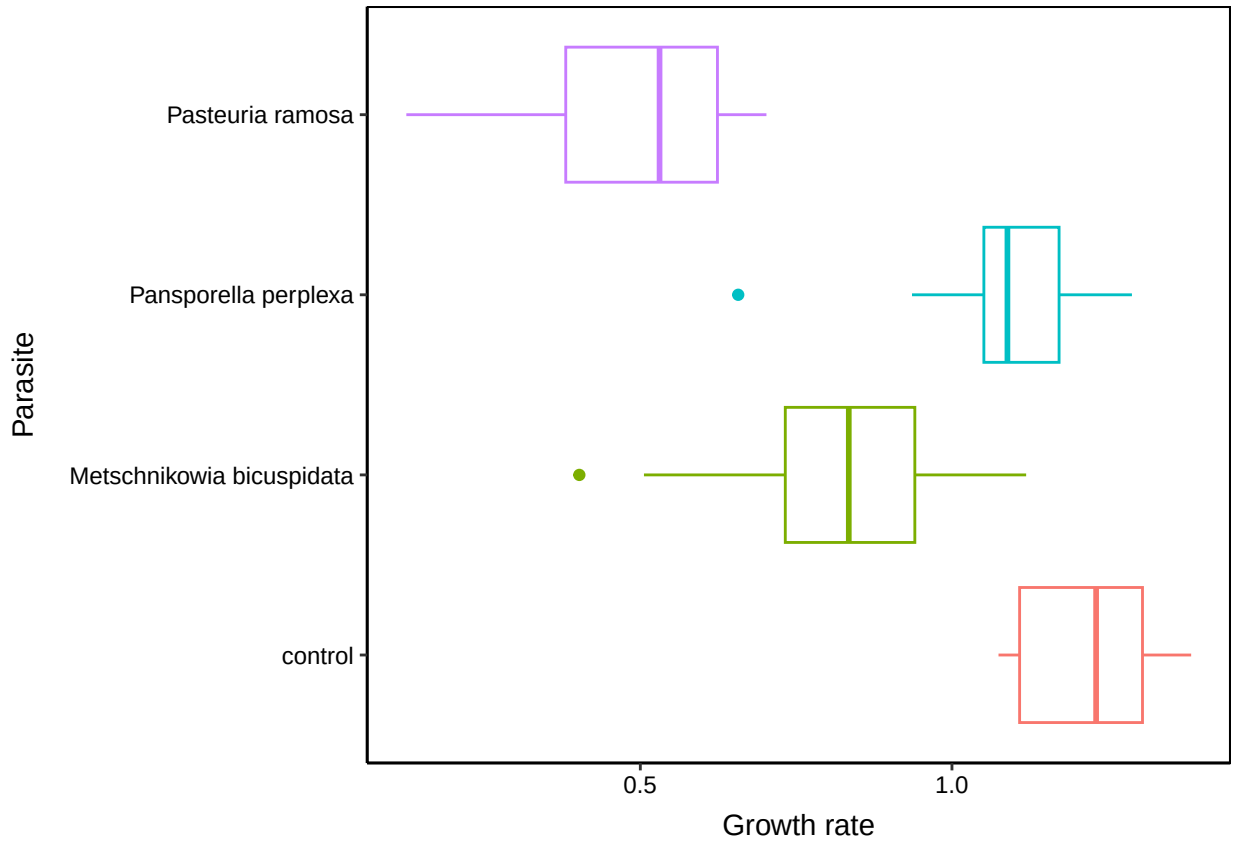
GAUSSIAN

Running this Gaussian GLM is essentially the same as a standard ANOVA or linear model. Here is a data set with growth rates (continuous variable) of different *Daphnia* species and a control (i.e. categories). This data comes from the book *Getting Started with R: An Introduction for Biologists* by Andrew Beckerman et al. 2017. We'll fit a Gaussian GLM here. Run a standard linear model and ANOVA for yourself as well, and compare the output.

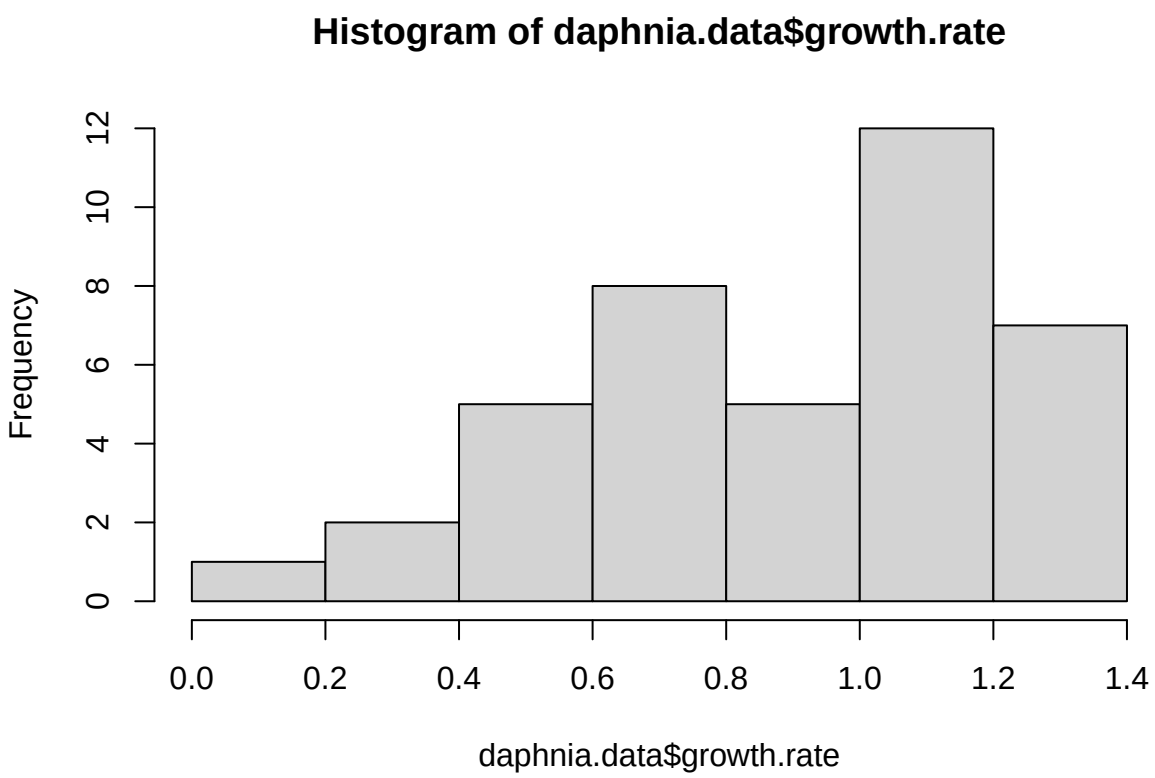
```
daphnia.data = read.csv("data/datasets_getting_started/Daphniagrowth.csv")
str(daphnia.data)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ parasite   : chr  "control" "control" "control" "control" ...
##  $ rep        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ growth.rate: num  1.07 1.27 1.32 1.08 1.2 ...
```

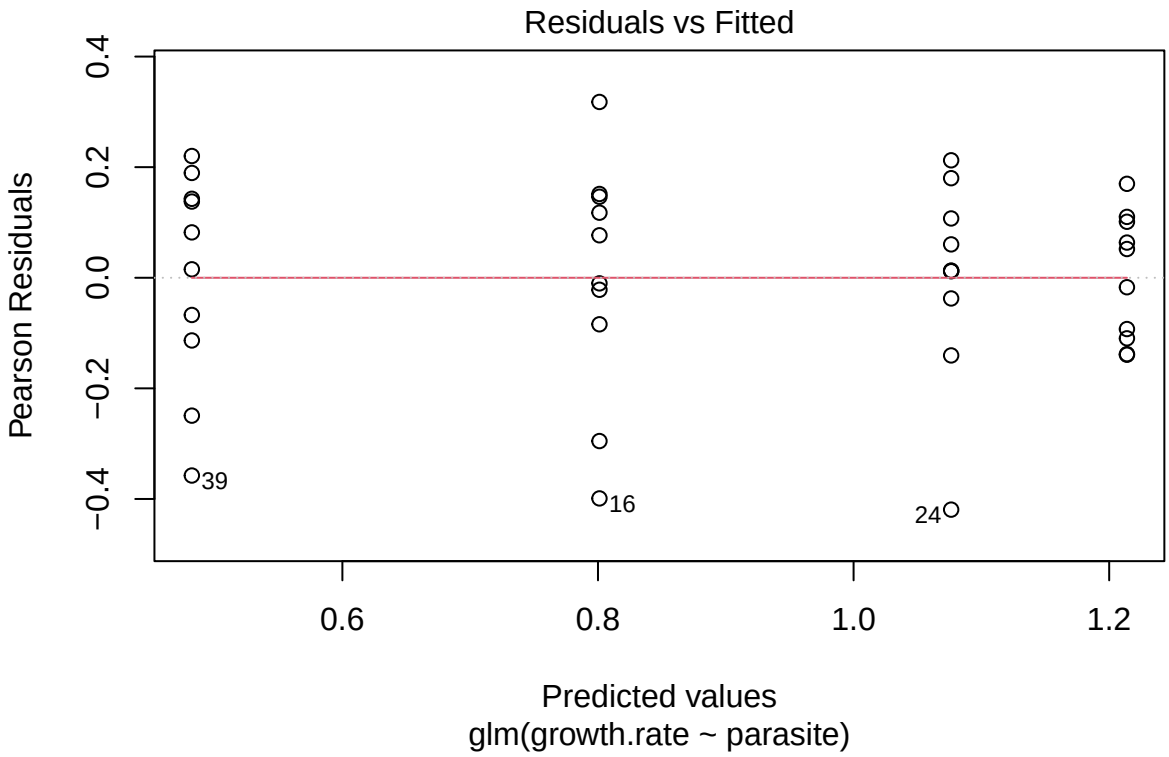
```
# Plot growth rate against parasite
ggplot2::ggplot(data = daphnia.data,
                aes(x = parasite, y = growth.rate, colour = parasite)) +
  geom_boxplot() +
  coord_flip() +
  xlab("Parasite") +
  ylab("Growth rate")
```



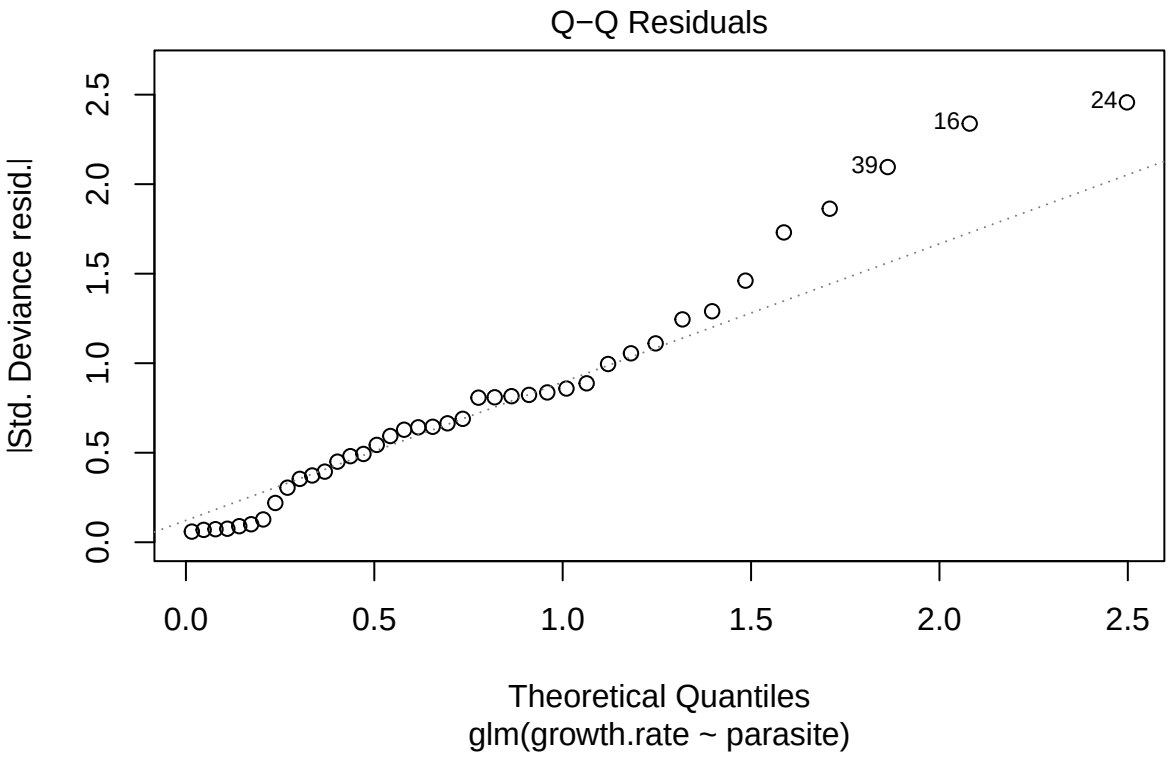

```
hist(daphnia.data$growth.rate)
```



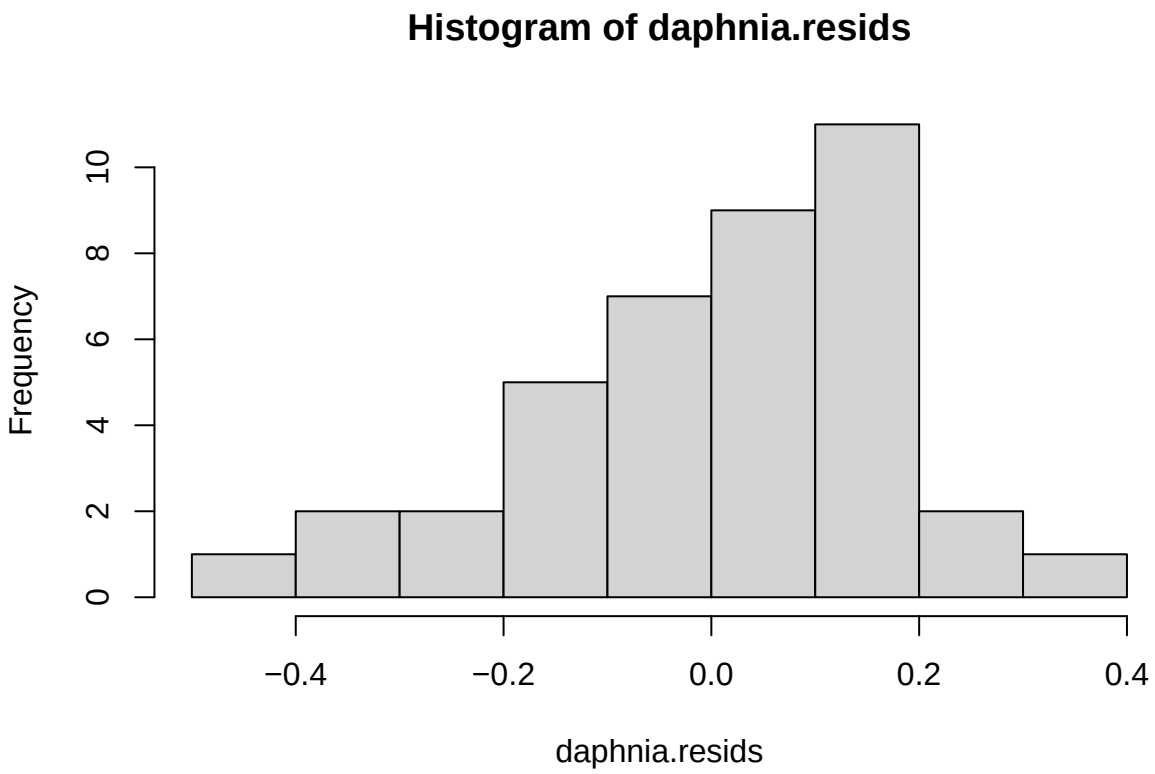

```r
# let's run a Gaussian GLM -> you could equally as easily run a linear model
daphnia.glm.gaus = glm(data = daphnia.data,
                       growth.rate ~ parasite,
                       family = gaussian)

# have a look at model diagnostics
plot(daphnia.glm.gaus, which = 1)
```



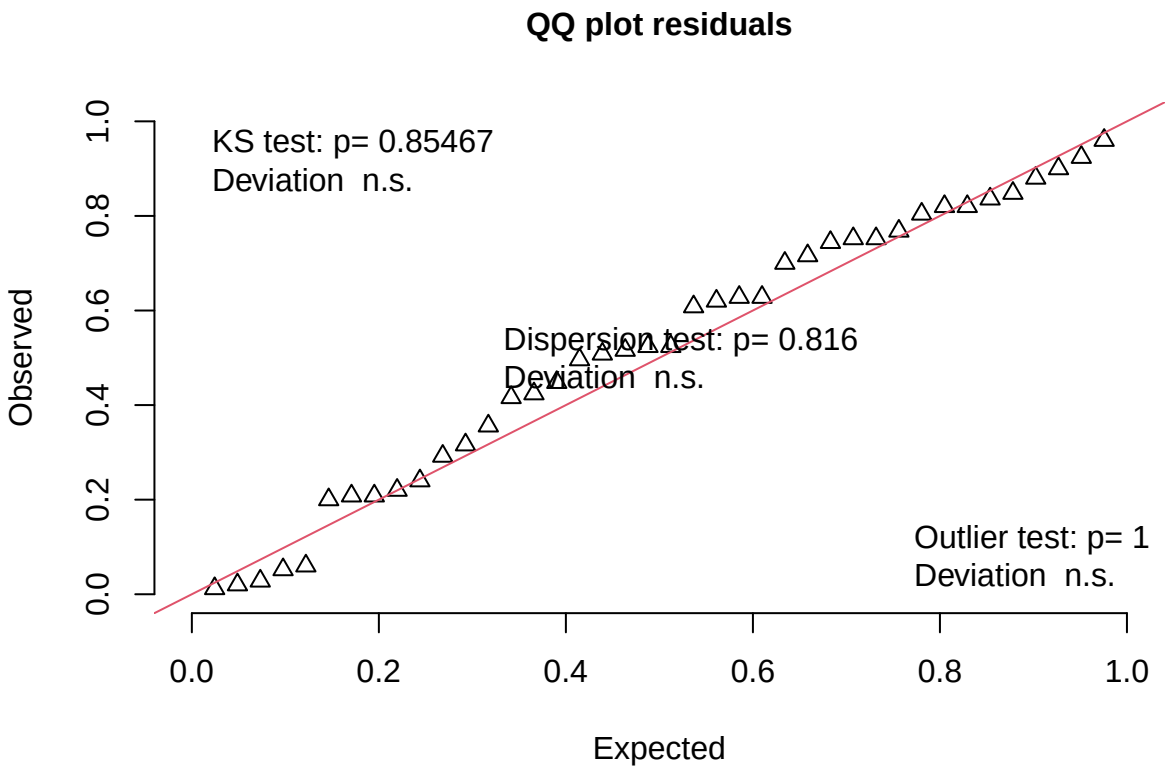

```
plot(daphnia.glm.gaus, which = 2)
```




```
daphnia.resids = resid(daphnia.glm.gaus)
hist(daphnia.resids)
```




```
DHARMa::plotQQunif(daphnia.glm.gaus)
```




```
#performance::check_model(daphnia.glm.gaus, check = "normality")

# we'll run a likelihood ratio test
# here we see that parasite does have an effect on growth rate
# Note that we are not running an ANOVA here, but rather producing a
# deviance table

# use anova for type I
anova(daphnia.glm.gaus, test = "LR")
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: growth.rate
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        39     4.3028
## parasite  3   3.1379        36     1.1649 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# get means
daphnia.means = Rmisc::summarySE(daphnia.data, measurevar = "growth.rate",
                                 groupvars = "parasite")

# summary of the model
# notice how the control group is not in the results table. This is because
# it is represented by the Intercept, and the other parasites are compared
# to that. The Intercept will be taken as the first level (i.e. alphabetically)
# of your groups.
summary(daphnia.glm.gaus)
```

```
##
## Call:
## glm(formula = growth.rate ~ parasite, family = gaussian, data = daphnia.data)
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.21391    0.05688  21.340  < 2e-16 ***
## parasiteMetschnikowia bicuspidata -0.41275    0.08045  -5.131 1.01e-05 ***
## parasitePansporella perplexa      -0.13755    0.08045  -1.710   0.0959 .
## parasitePasteuria ramosa          -0.73171    0.08045  -9.096 7.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.03235768)
##
##     Null deviance: 4.3028  on 39  degrees of freedom
## Residual deviance: 1.1649  on 36  degrees of freedom
## AIC: -17.935
##
## Number of Fisher Scoring iterations: 2
```

```r
# another neater way of viewing the summary results
gtsummary::tbl_regression(daphnia.glm.gaus)
```

Characteristic	Beta	95% CI	p-value
parasite			
control	—	—	
Metschnikowia bicuspidata	-0.41	-0.57, -0.26	<0.001
Pansporella perplexa	-0.14	-0.30, 0.02	0.10
Pasteuria ramosa	-0.73	-0.89, -0.57	<0.001

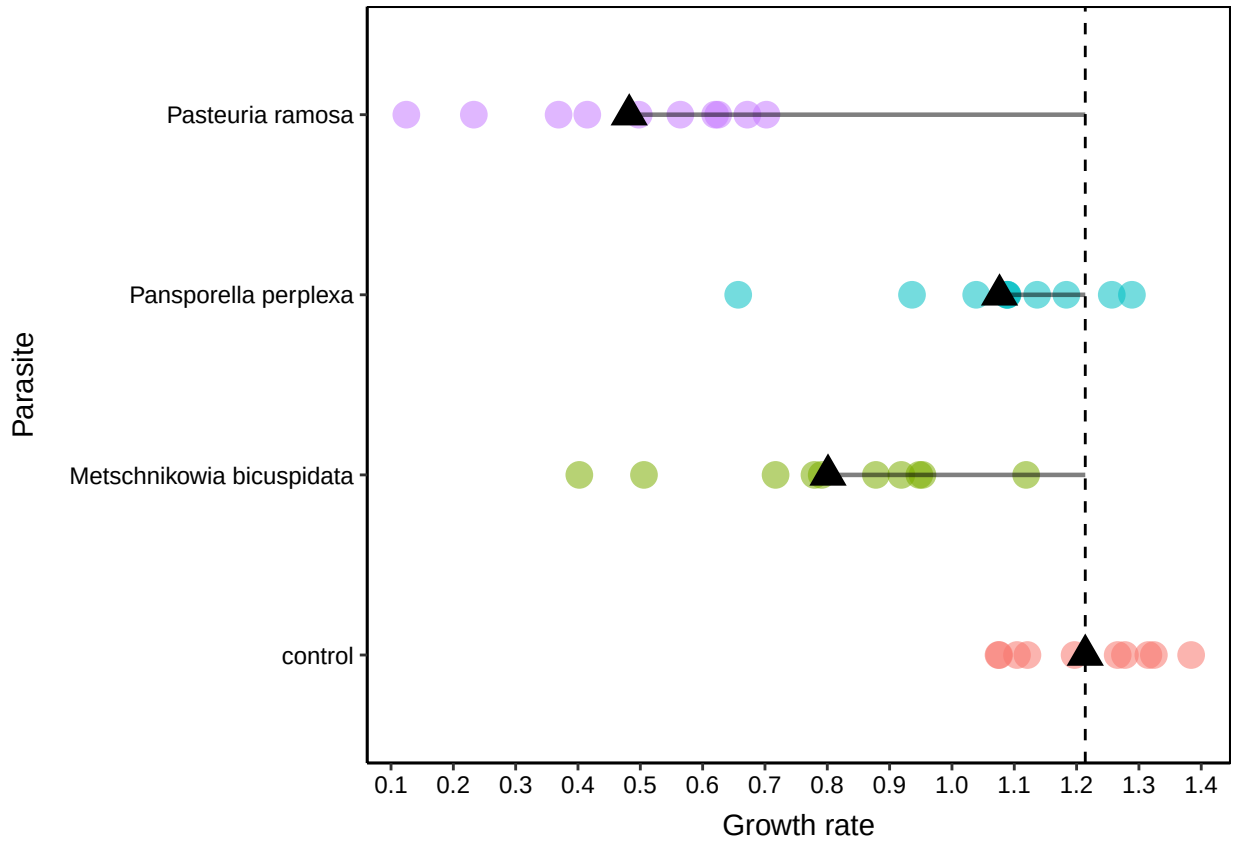
```r
# notice the four levels of parasites in the data: R orders them alphabetically
# this is why the control (first in alphabetical order) is the reference
# in the summary output
levels(factor(daphnia.data$parasite))
```

```
## [1] "control"                   "Metschnikowia bicuspidata"
## [3] "Pansporella perplexa"      "Pasteuria ramosa"
```

Let's plot the mean growth rate for each treatment. The dotted line is at **x = 1.21391**, which is the mean of the control. See the Estimate values in the **summary(daphnia.glm.gaus)** output. When there is a categorical predictor, the estimate values are the means. This changes when the predictor is a continuous variable though, so be careful.

The other negative values just indicate how far away each of the other parasites are away from the mean of the control treatment. For example, *M. bicuspidata* is -0.41275 away from the control -> 1.21391 - 0.41275 = 0.80116, and this difference is significant ($p < 0.01$). The horizontal black lines in the plot below show these distance differences from the control's mean value (dotted vertical line).

```
ggplot2::ggplot() +
  geom_point(data = daphnia.data, shape = 16,
             aes(x = parasite, y = growth.rate,
                 colour = parasite, size = 8, alpha = 0.5)) +
  geom_point(data = daphnia.means, aes(x = parasite, y = growth.rate, size = 8),
             shape = 17) +
  geom_hline(yintercept = 1.21391, linetype = "dashed") +
  geom_segment(aes(y=1.21391, x=4, yend=0.4822030, xend=4),
               linewidth = 0.8, alpha = 0.5) +
  geom_segment(aes(y=1.21391, x=3, yend=1.0763551, xend=3),
               linewidth = 0.8, alpha = 0.5) +
  geom_segment(aes(y=1.21391, x=2, yend=0.8011541, xend=2),
               linewidth = 0.8, alpha = 0.5) +
  scale_y_continuous(breaks=seq(0,2,by=0.1)) +
  coord_flip() +
  xlab("Parasite") +
  ylab("Growth rate")
```




```
# run some post hoc tests to see differences between groups
posthoc.daphnia = emmeans::emmeans(daphnia.glm.gaus, pairwise ~ parasite, adjust = "tukey")

posthoc.daphnia$contrasts %>%
  summary(infer = TRUE)
```

```
##  contrast                                          estimate     SE df lower.CL
##  control - Metschnikowia bicuspidata                  0.413 0.0804 36   0.1961
##  control - Pansporella perplexa                       0.138 0.0804 36  -0.0791
##  control - Pasteuria ramosa                           0.732 0.0804 36   0.5150
##  Metschnikowia bicuspidata - Pansporella perplexa    -0.275 0.0804 36  -0.4919
##  Metschnikowia bicuspidata - Pasteuria ramosa         0.319 0.0804 36   0.1023
##  Pansporella perplexa - Pasteuria ramosa              0.594 0.0804 36   0.3775
##  upper.CL t.ratio p.value
##    0.6294   5.131  0.0001
##    0.3542   1.710  0.3335
##    0.9484   9.096  <.0001
##   -0.0585  -3.421  0.0082
##    0.5356   3.965  0.0018
##    0.8108   7.386  <.0001
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 4 estimates
## P value adjustment: tukey method for comparing a family of 4 estimates
```

What can you conclude from these pairwise contrasts? Have a look at the plot again to check the comparison

between *P. perplexa* and the control. Look at the signs of the estimate values. For example, the contrast between the control and *M. bicuspidata* is +0.413, with a p-value < 0.001. This means that the control was significantly higher than the *M. bicuspidata* parasite. The *M. bicuspidata* - *P. perplexa* comparison was -0.275, which means that *M. bicuspidata* was significantly lower than *P. perplexa* ($p < 0.05$).

Let's run an ANOVA here to show that the output is the same as the GLM:

```
daphnia.aov = aov(growth.rate ~ parasite, data = daphnia.data)
summary(daphnia.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## parasite     3  3.138  1.0460   32.33 2.57e-10 ***
## Residuals   36  1.165  0.0324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And a linear model:

```
daphnia.lm = lm(growth.rate ~ parasite, data = daphnia.data)
summary(daphnia.lm)
```

```
##
## Call:
## lm(formula = growth.rate ~ parasite, data = daphnia.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41930 -0.09696  0.01408  0.12267  0.31790
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.21391    0.05688  21.340  < 2e-16 ***
## parasiteMetschnikowia bicuspidata -0.41275    0.08045  -5.131 1.01e-05 ***
## parasitePansporella perplexa      -0.13755    0.08045  -1.710   0.0959 .
## parasitePasteuria ramosa          -0.73171    0.08045  -9.096 7.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1799 on 36 degrees of freedom
## Multiple R-squared:  0.7293, Adjusted R-squared:  0.7067
## F-statistic: 32.33 on 3 and 36 DF,  p-value: 2.571e-10
```

```
anova(daphnia.lm, test = "F")
```

```
## Analysis of Variance Table
##
## Response: growth.rate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## parasite   3 3.1379 1.04597  32.325 2.571e-10 ***
## Residuals 36 1.1649 0.03236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLM using binomial data: dead or alive

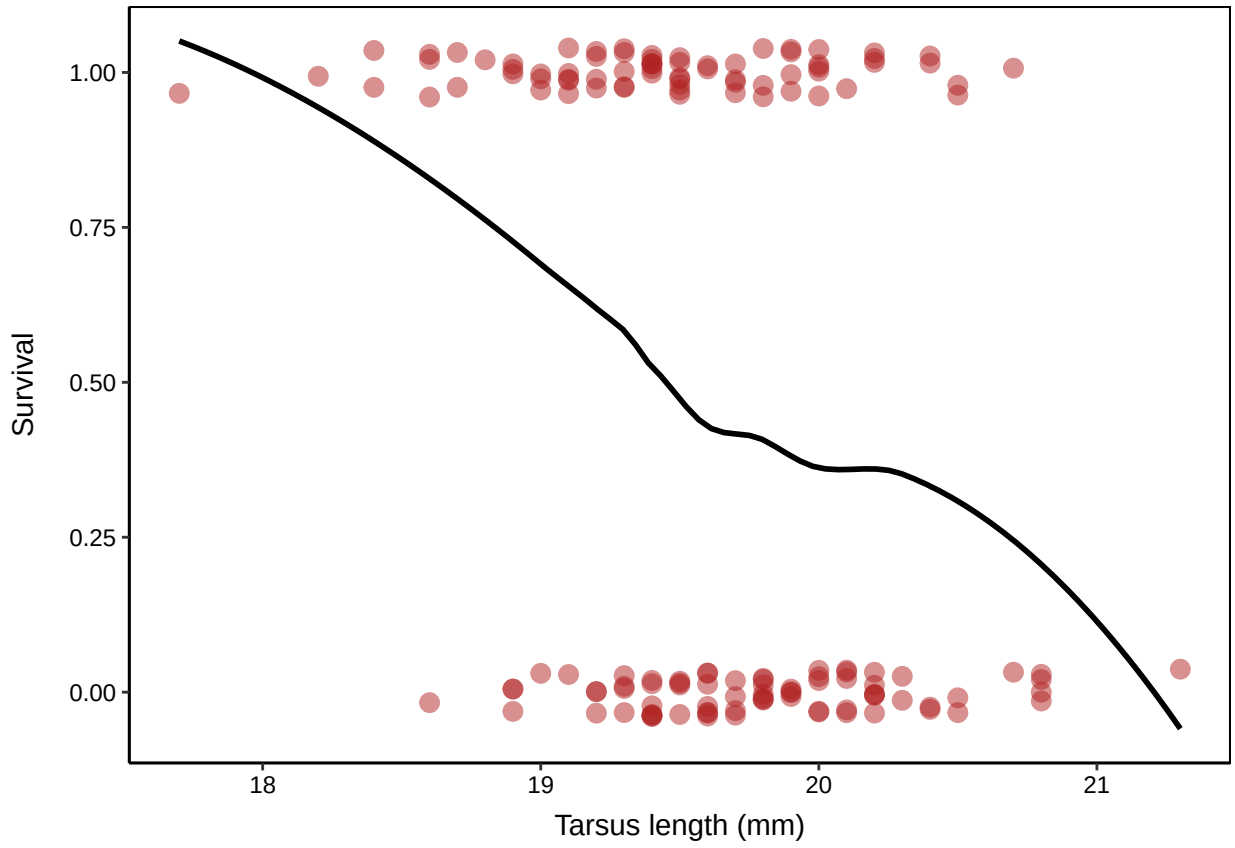
BINOMIAL

Let's have a look at a data set with morphological measurements (continuous) and survival statistics (binary: dead (0) or alive (1)) of North American song sparrows, as taken from Schluter and Smith, 1986. Let's start by fitting a linear model, just to see that it isn't appropriate for binary data!

```
sparrow.data = read.csv("data/songsparrow.csv")
head(sparrow.data)
```

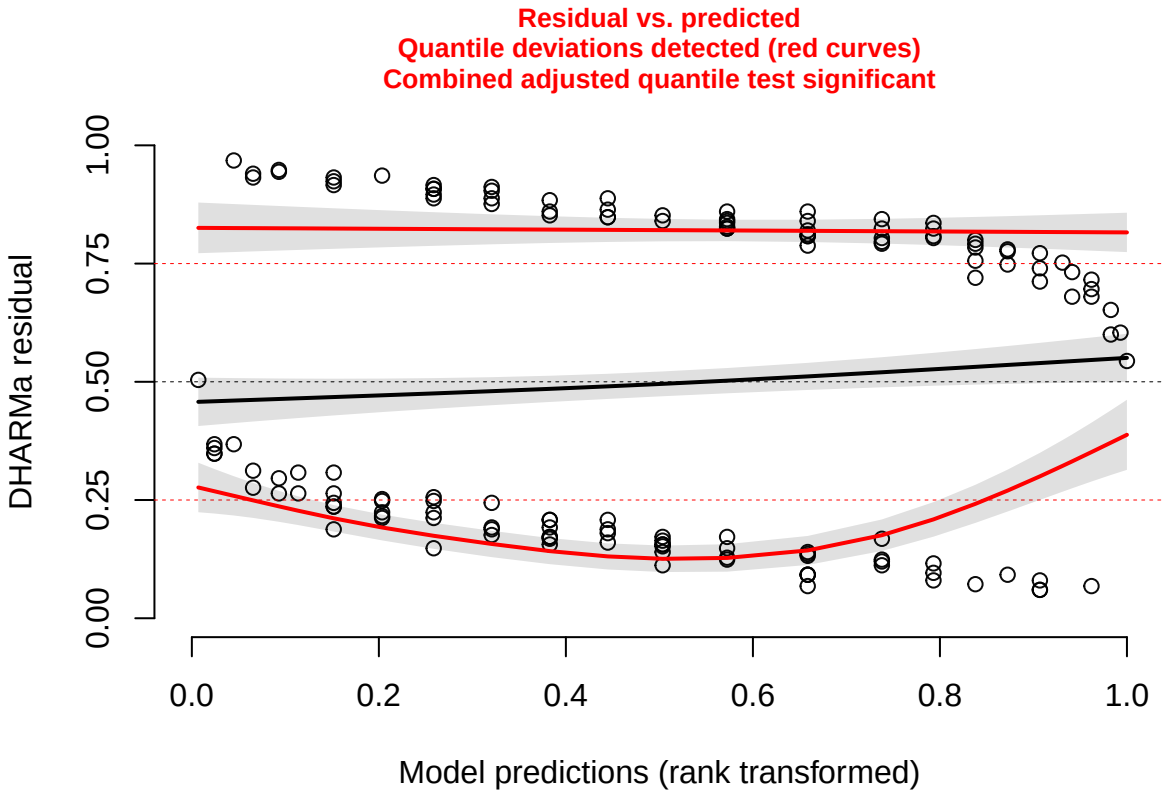
```
##   mass wing tarsus blength bdepth bwidth year sex survival
## 1 23.7 67.0   17.7     9.1    5.9    6.8 1978   f        1
## 2 23.1 65.0   19.5     9.5    5.9    7.0 1978   f        0
## 3 21.8 65.2   19.6     8.7    6.0    6.7 1978   f        0
## 4 21.7 66.0   18.2     8.4    6.2    6.8 1978   f        1
## 5 22.5 64.3   19.5     8.5    5.8    6.6 1978   f        1
## 6 22.9 65.8   19.6     8.9    5.8    6.6 1978   f        1
```

```
# plot tarsus length vs survival
# geom_jitter is an alternative to geom_point, but it presents the points
# in a clearer way, especially if there is a lot of overlap
ggplot2::ggplot(data = sparrow.data, aes(x = tarsus, y = survival)) +
  geom_jitter(color = "firebrick", size = 3,
              height = 0.04, width = 0,
              alpha = 0.5) +
  # this adds a trendline
  geom_smooth(method = "loess", linewidth = 1, col = "black", se = FALSE) +
  xlab("Tarsus length (mm)") +
  ylab("Survival")
```

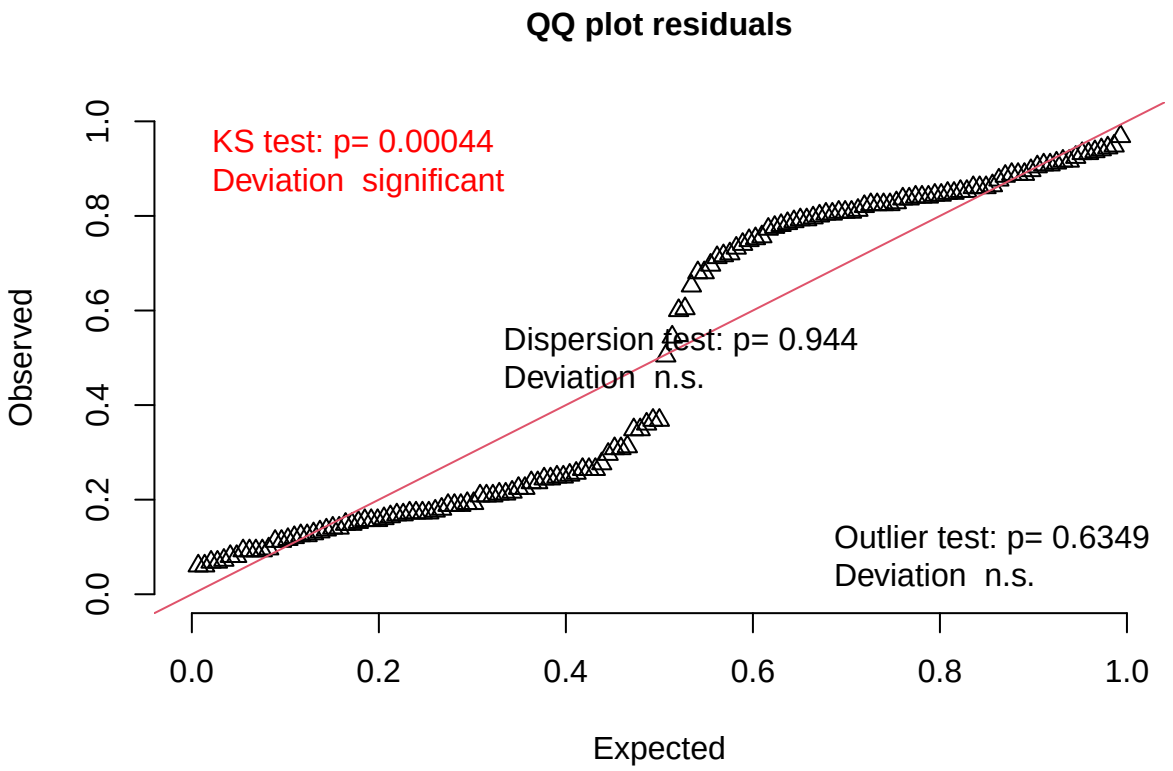


Would you agree that sparrows with shorter tarsi were favoured by natural selection?

```
# run a linear model
sparrow.lm = lm(survival ~ tarsus, data = sparrow.data)

# in this residual plot, there are three dashed lines: at the 0.25, 0.5, and
# 0.75 quantiles. We expect to see our solid lines closely following these
# straight lines. You can see how the upper and lower red lines deviate quite
# noticeably
DHARMa::plotResiduals(sparrow.lm)
```




```
#plot(DHARMa::simulateResiduals(sparrow.lm))

# Here we also see a significant deviation of the KS test (Kolmogorov-Smirnov).
# The KS test is an indication of goodness of fit. At least there is not
# significant overdispersion or outliers in the data.
DHARMa::plotQQunif(sparrow.lm)
```



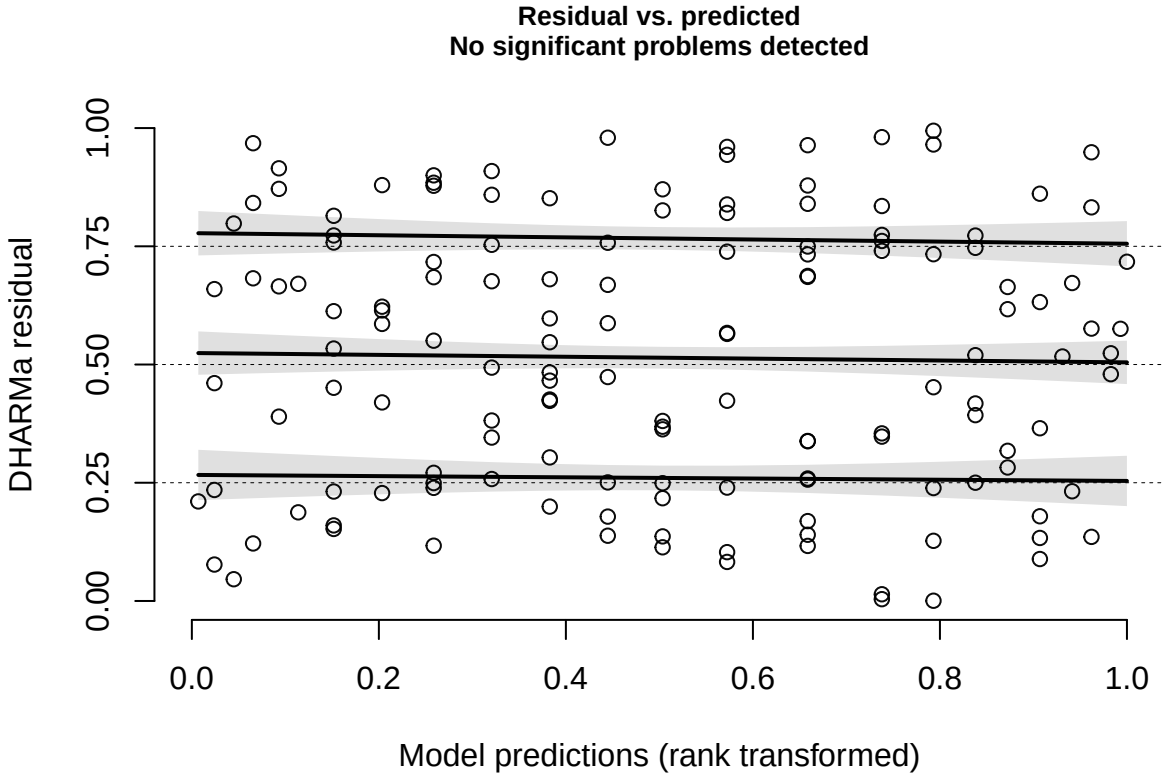
We can see clearly that a linear model is inappropriate for this data, and that its output will be unreliable. Let's dive into a GLM for count data! We'll specify that we want to use the **binomial** family, since we are dealing with "dead" or "alive" data, in the form of a zero or one.

```
# Let's run a GLM where we look at the effect of tarsus length on survival
sparrow.glm = glmmTMB::glmmTMB(survival ~ tarsus,
                  family = binomial(link = "logit"),
                  data = sparrow.data)

# we can also run a null model -> assumes that the log-odds of survival is
# consistent across all tarsal measurements (null hypothesis)
sparrow.glm.null = glmmTMB::glmmTMB(survival ~ 1,
                  family = binomial(link = "logit"),
                  data = sparrow.data)

# now we can compare the models
# You can see that model 2, which includes
# tarsus length, is a much better fit than the null model. This difference is
# also significant, as p < 0.05. This just means that tarsal length is a
# significant predictor of survival
anova(sparrow.glm.null, sparrow.glm, test = "Chisq")
```

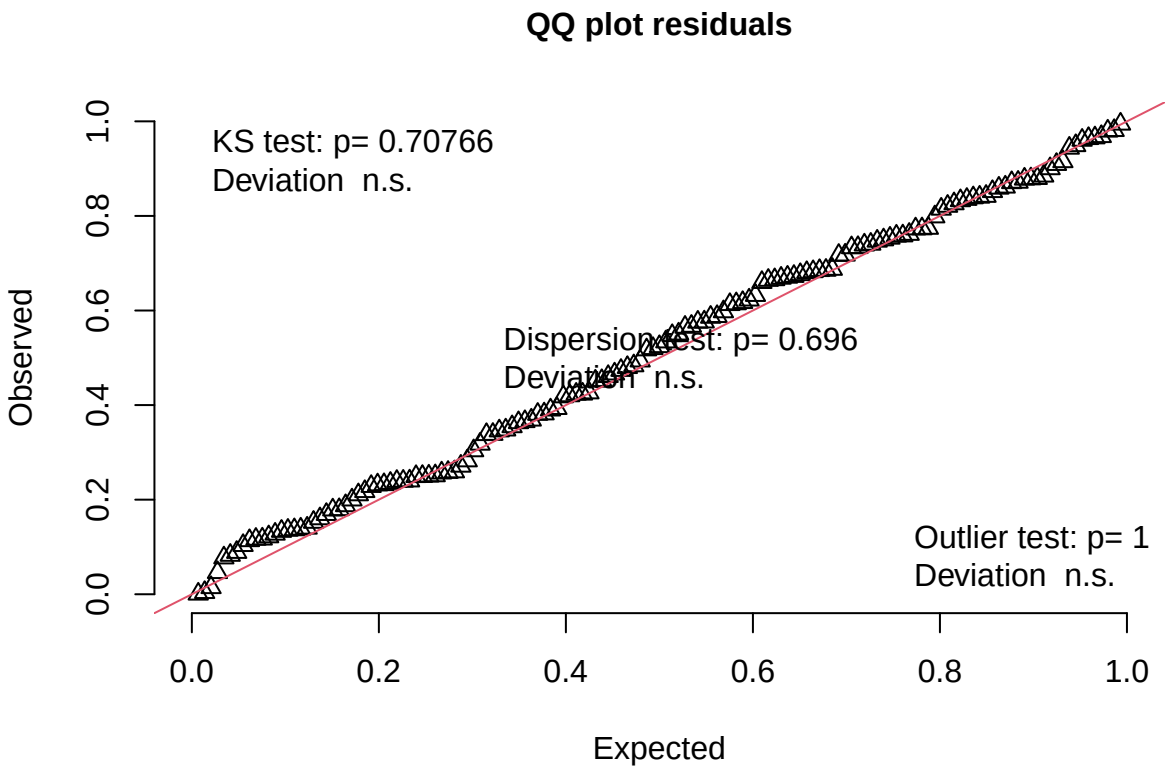
```
## Data: sparrow.data
## Models:
## sparrow.glm.null: survival ~ 1, zi=~0, disp=~1
## sparrow.glm: survival ~ tarsus, zi=~0, disp=~1
```

```
##                  Df    AIC    BIC   logLik deviance  Chisq Chi Df Pr(>Chisq)
## sparrow.glm.null  1 202.95 205.93 -100.475   200.95
## sparrow.glm       2 189.04 195.00  -92.521   185.04 15.908      1   6.65e-05
##
## sparrow.glm.null
## sparrow.glm      ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

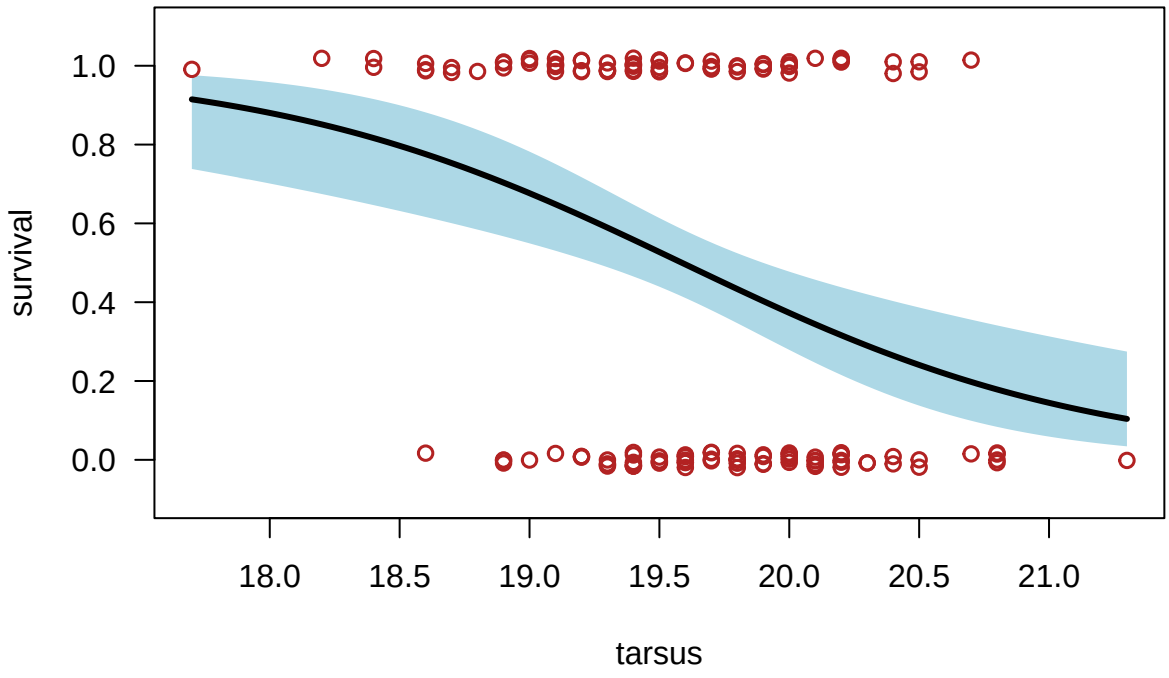
```
# Check the residuals and QQ plot
DHARMa::plotResiduals(sparrow.glm)
```




```
DHARMa::plotQQunif(sparrow.glm)
```

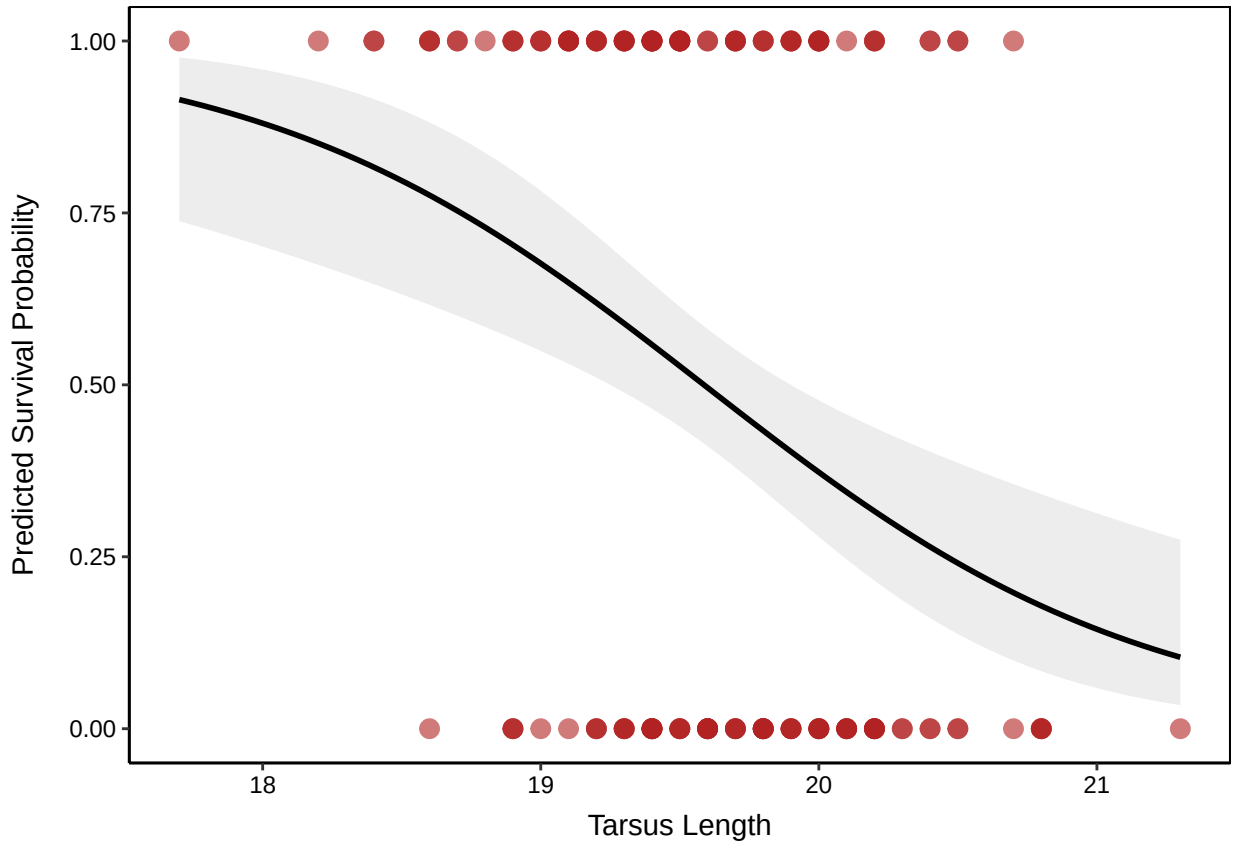



```
# take a quick peek at the model
visreg::visreg(sparrow.glm, xvar = "tarsus",
               scale = 'response', rug = FALSE,
               ylim = c(-.1, 1.1),
               line.par = list(col = 'black'),
               fill.par = list(col = 'lightblue'))
points(jitter(survival, 0.1) ~ tarsus,
       data = sparrow.data, pch = 1,
       col = "firebrick", lwd = 1.5)
```




```
# Let's plot this model using ggplot, for practice
# First, we'll add a column to our sparrow.data -> containing predictions
# These are probabilities of survival, based on our GLM
sparrow.data$predictions = predict(sparrow.glm, type = "response")

# Now let's plot in ggplot
ggplot(sparrow.data, aes(x = tarsus, y = predictions)) +
  geom_point(aes(x = tarsus, y = survival),
             color = "firebrick", size = 3, alpha = 0.6) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
              se = TRUE, color = "black", fill = "lightgrey") +
  labs(x = "Tarsus Length", y = "Predicted Survival Probability")
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```



You can see what a difference there is to the QQ plot and residuals after applying the GLM! Let's have a closer look at the stats now. The Estimate value for the Intercept and response variable (here it is tarsus) are referred to as the beta coefficients.

```
# tarsus length has a significant effect on survival
car::Anova(sparrow.glm, test = "Chisq", type = "II")
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: survival
##         Chisq Df Pr(>Chisq)
## tarsus 13.391  1  0.0002529 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(sparrow.glm)
```

```
##  Family: binomial  ( logit )
## Formula:          survival ~ tarsus
## Data: sparrow.data
##
##      AIC      BIC   logLik deviance df.resid
##    189.0    195.0    -92.5    185.0      143
##
##
```

```
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  24.6361     6.7454   3.652 0.000260 ***
## tarsus       -1.2578     0.3437  -3.659 0.000253 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gtsummary::tbl_regression(sparrow.glm, exponentiate = TRUE)
```

Characteristic	exp(Beta)	95% CI	p-value
tarsus	0.28	0.14, 0.56	<0.001

```
# get the exponential of the beta coefficient, since we ran a logistic
# regression. Note that the sign was negative, which just tells us the
# direction of the effect. R gives the log-odds, we need to get it as an
# odds ratio
odds.ratio = exp(-1.2578)

# percentage
odds.ratio.perc = (odds.ratio - 1)*100

confint(sparrow.glm)
```

```
##                2.5 %     97.5 % Estimate
## (Intercept) 11.41542 37.8568127 24.63612
## tarsus      -1.93147 -0.5841102 -1.25779
```

```
# get the exponential of the confidence intervals. Note the negative signs
conf.int.lower = exp(-1.97)
conf.int.upper = exp(-0.58)
```

Here we see that tarsus length has a significant effect on survival ($\chi^2 = 13.4$, d.f. $= 1$, $p < 0.001$). The beta coefficient value of 0.28 (**exp(-1.2578)**) means that for every millimeter increase in tarsus length, the probability of survival decreases (due to the negative sign) by a factor of **0.28**. As a percentage, this is an approximate **72% reduction** in the odds ([0.28 - 1])*100. The 95% confidence interval suggests that a tarsal length increase of 1 mm will result in a decreased odds of survival at a factor of between 0.14 and 0.56.

Perhaps we want to find out what the probability of survival is for sparrows that have tarsal lengths of 30, 20.5, 15, and 14 mm:

```
predict(sparrow.glm, newdata = data.frame(tarsus = c(22, 20.5, 15, 14)),
        type = "response")
```

```
## [1] 0.04585805 0.24074914 0.99688767 0.99911325
```

Run the same analysis on one or more of the other morphological measurements, and have a look what effect they have on survival.

GLM using count data

POISSON AND NEGATIVE BINOMIAL

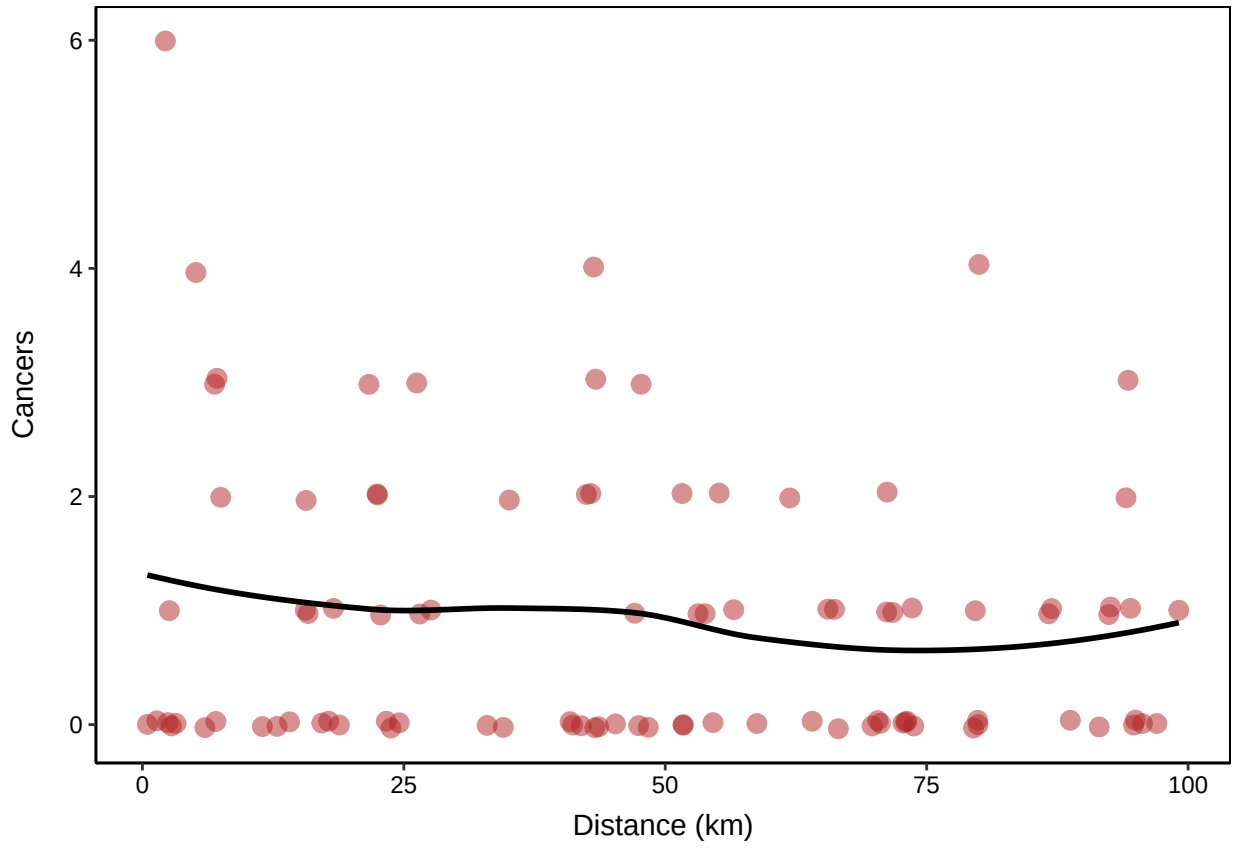
Let's have a look at a quick example of how a GLM can be used to analyse the relationship between the distance to a nuclear power plant (continuous predictor variable, in km), and the number of cases of cancer per year per clinic (count response variable). This data set was taken from the R book second edition by Michael Crawley, chapter 14.

We'll first apply the **poisson** family to this GLM, since we are dealing with cancer counts. Sometimes there is a lot of variation in count data, which results in overdispersion. In a normal Poisson model, we expect mean to equal variance, but when it is overdispersed, variance > mean. A negative binomial model is then more suited.

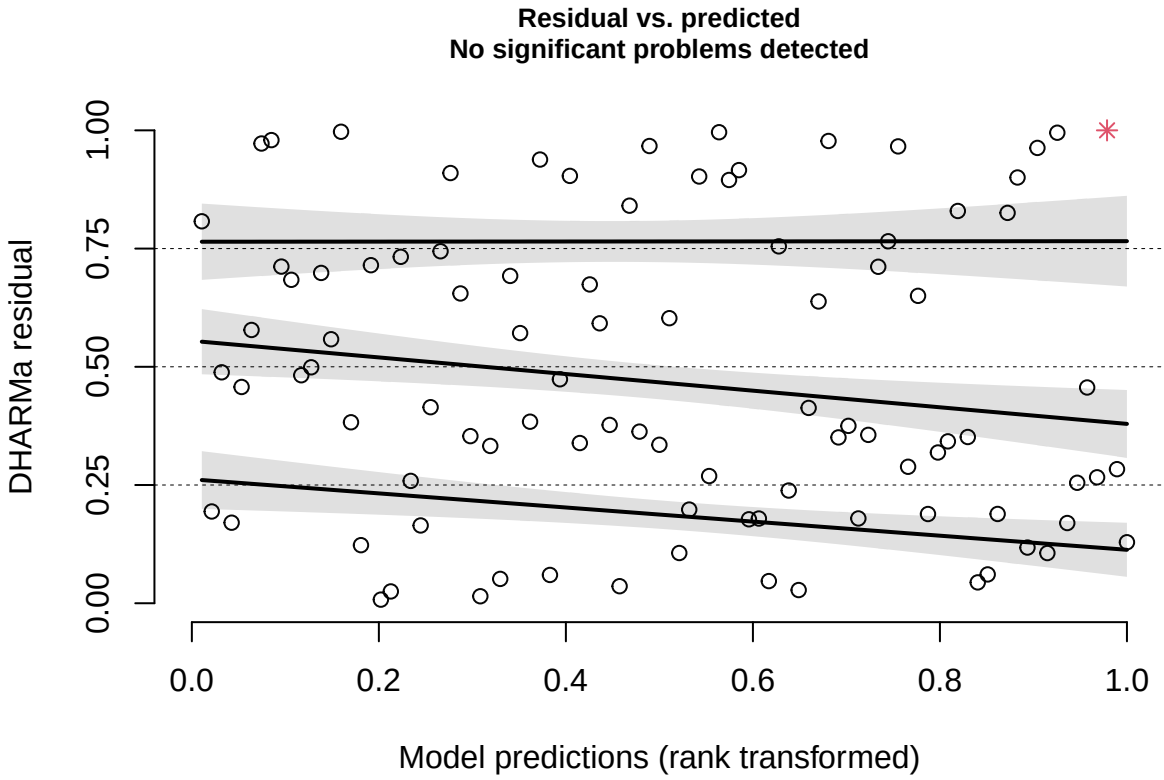
```
# read in the data
cancer.data = read.table("data/therbook/clusters.txt", header = T)

head(cancer.data)
```

```
##   Cancers Distance
## 1       0 11.46952
## 2       0 66.55395
## 3       0 47.46230
## 4       0 48.38129
## 5       0 73.76534
## 6       0 70.57555
```

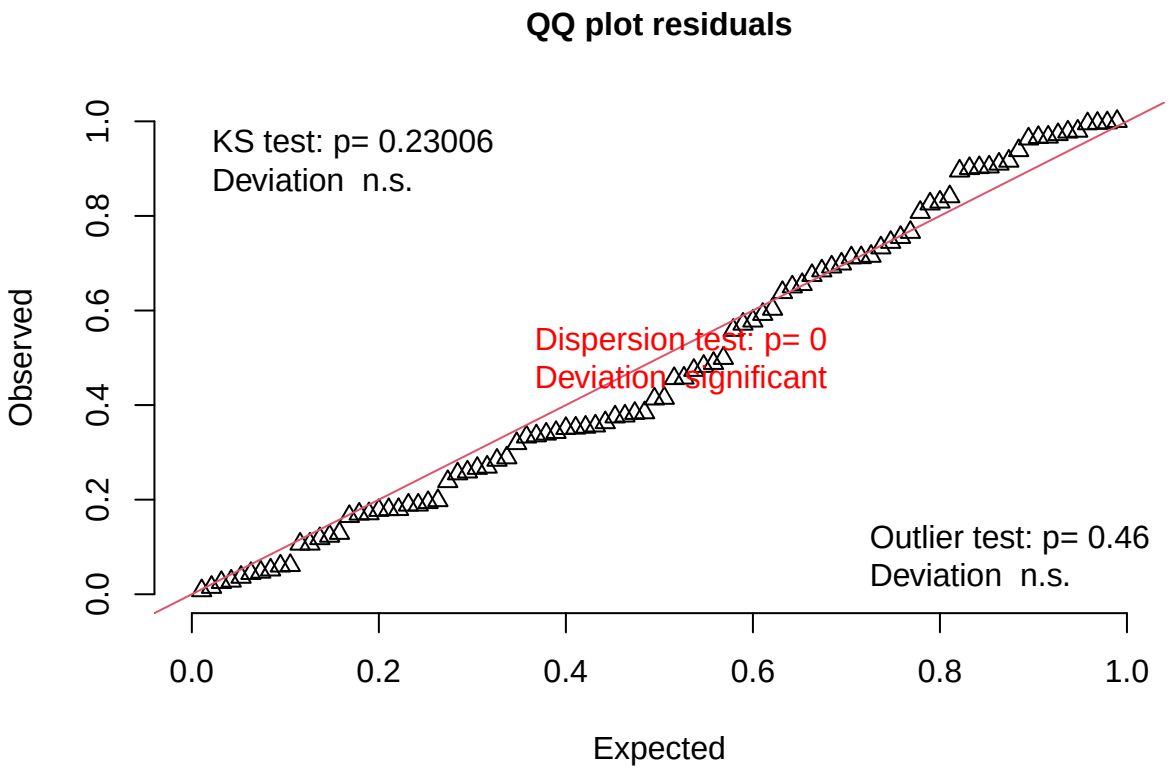
```
# Let's plot it
ggplot2::ggplot(data = cancer.data, aes(x = Distance, y = Cancers)) +
  geom_jitter(color = "firebrick", size = 3,
              height = 0.04, width = 0,
              alpha = 0.5) +
  # this adds a trendline
  geom_smooth(method = "loess", linewidth = 1, col = "black", se = FALSE) +
  xlab("Distance (km)") +
  ylab("Cancers")
```



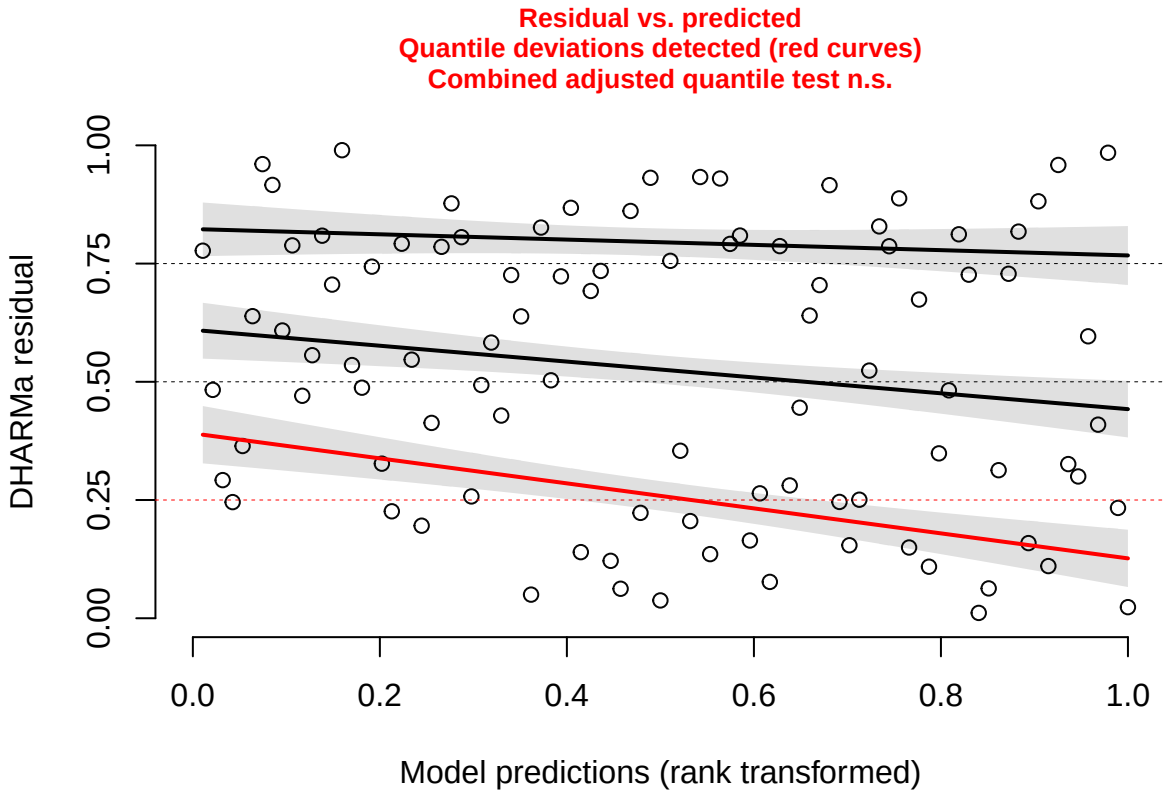

```r
# Run a poisson GLM
cancer.glm.poisson = glm(Cancers ~ Distance, family = poisson,
                         data = cancer.data)

# Another approach, using the glmmTMB library
cancer.glm.poisson.2 = glmmTMB::glmmTMB(Cancers ~ Distance,
                          family = poisson(link = "log"),
                          data = cancer.data)

# Check the residuals and QQ plot
DHARMa::plotResiduals(cancer.glm.poisson)
```



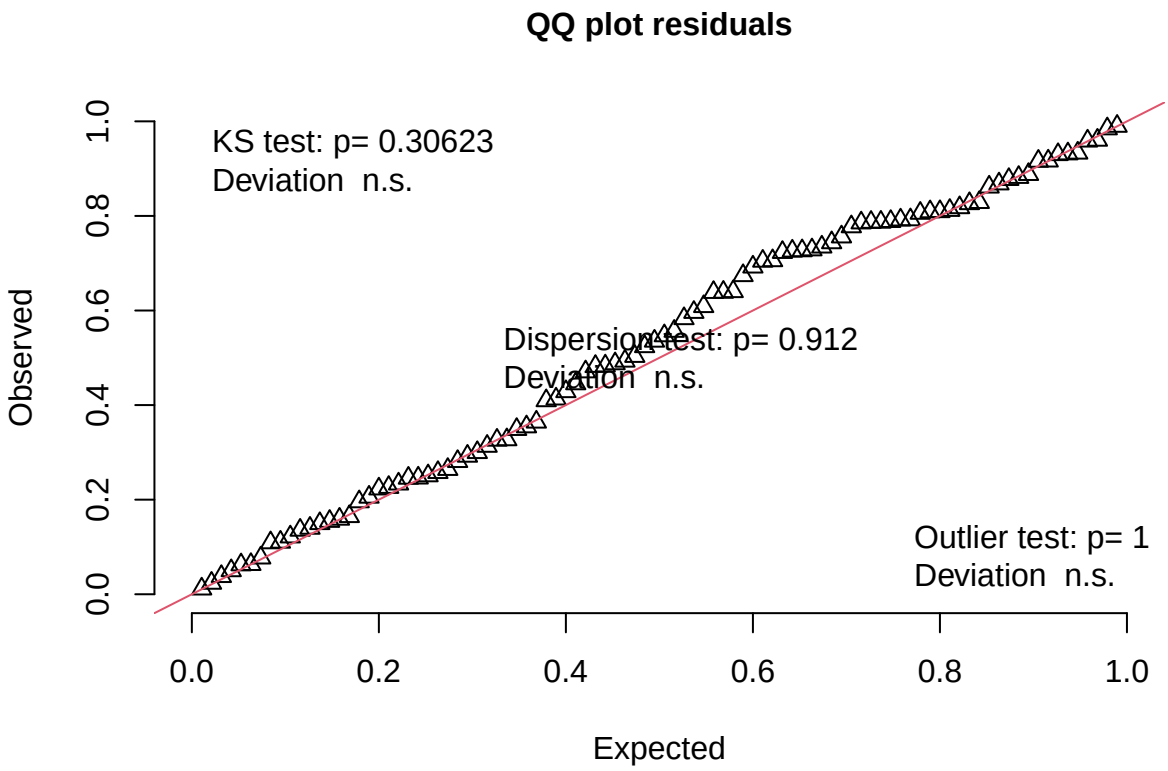

```
# notice that there is significant overdispersion in the data
DHARMa::plotQQunif(cancer.glm.poisson)
```



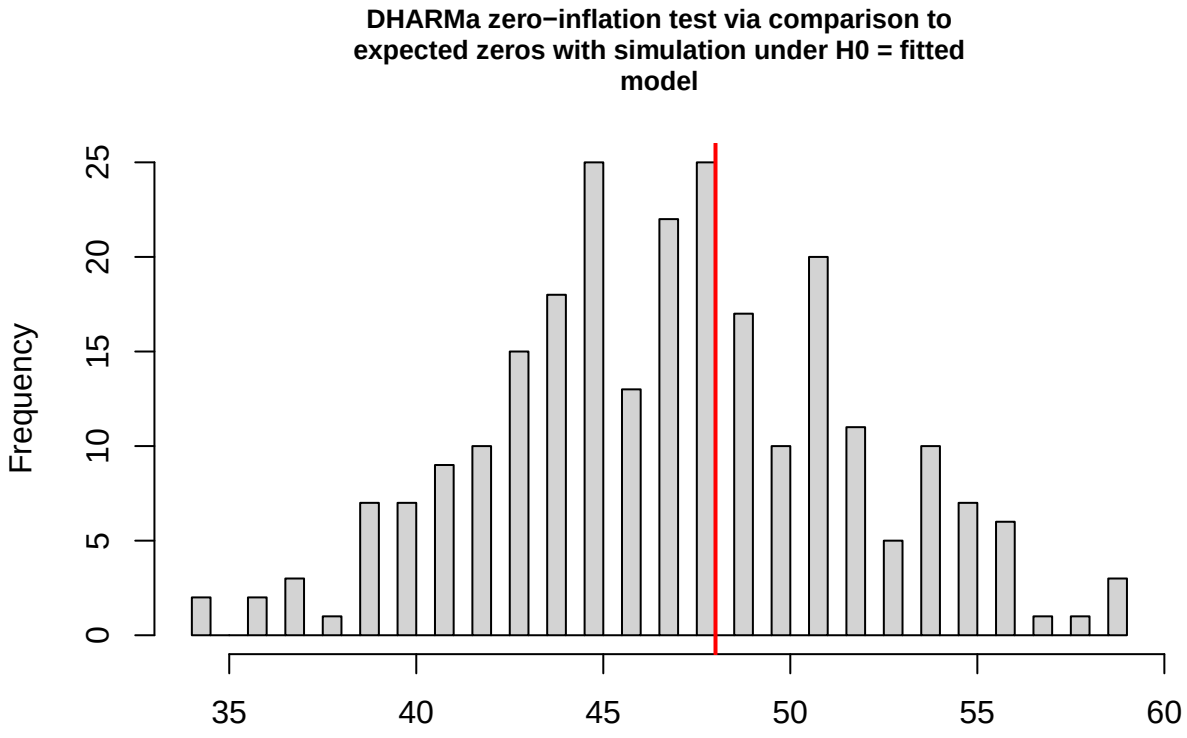

```
# Run a negative binomial GLM -> this assumes that variance > mean, and is
# often a better fit for most count data, particularly field-collected
cancer.glm.negbin = glmmTMB::glmmTMB(Cancers ~ Distance,
                        family = nbinom2(link = "log"),
                        data = cancer.data)

# Check the residuals and QQ plot
# although the bottom line doesn't look perfect, this model is fine
DHARMa::plotResiduals(cancer.glm.negbin)
```




```
# no more overdispersion
DHARMa::plotQQunif(cancer.glm.negbin)
```




```
# let's run a test to check whether our data has too many zero values: termed
# "zero inflation"
# since p > 0.05, we do not have an issue with zeroes
DHARMa::testZeroInflation(cancer.glm.negbin)
```




```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 1.0197, p-value = 0.928
## alternative hypothesis: two.sided
```

Let's compare the Poisson and negative binomial models:

```
gtsummary::tbl_regression(cancer.glm.negbin, exponentiate = TRUE)
```

Characteristic	exp(Beta)	95% CI	p-value
Distance	0.99	0.98, 1.00	0.2

```
anova(cancer.glm.poisson.2, cancer.glm.negbin, test = "Chisq")
```

```
## Data: cancer.data
## Models:
## cancer.glm.poisson.2: Cancers ~ Distance, zi=~0, disp=~1
## cancer.glm.negbin: Cancers ~ Distance, zi=~0, disp=~1
##                       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
```

```
## cancer.glm.poisson.2  2 262.40 267.49 -129.20   258.40
## cancer.glm.negbin     3 253.19 260.82 -123.59   247.19 11.214      1  0.0008117
##
## cancer.glm.poisson.2
## cancer.glm.negbin    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(cancer.glm.negbin, test = "Chisq", type = "II")
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: Cancers
##           Chisq Df Pr(>Chisq)
## Distance 1.6577  1     0.1979
```

```
summary(cancer.glm.negbin)
```

```
##  Family: nbinom2  ( log )
## Formula:          Cancers ~ Distance
## Data: cancer.data
##
##      AIC      BIC   logLik deviance df.resid
##    253.2    260.8   -123.6    247.2       91
##
##
## Dispersion parameter for nbinom2 family (): 1.36
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.182490   0.250903   0.727    0.467
## Distance    -0.006041   0.004692  -1.288    0.198
```

```
# beta 1 coefficient
exp(-0.006041)
```

```
## [1] 0.9939772
```

The anova() output shows that the negative binomial model provided a significantly better fit than the Poisson model ($p < 0.05$), and we can also see that it yielded lower AIC, BIC values, and a log likelihood value closer to zero.

Looking at the summary of the negative binomial model, we can conclude that a one kilometer increase in distance results in a decrease in the number of cancer patients by a factor of 0.9939772 -> exp(-0.006041). Or, as a percentage: [0.9939772 - 1] *100 = -0.6% -> i.e. a 0.6% decrease in the number of cancer records for every 1 km increase in distance from the nuclear power plant. However, this result was not significant ($\chi^2 = 1.7$, d.f. $= 1$, $p = 0.198$).

More count data -> horseshoe crabs

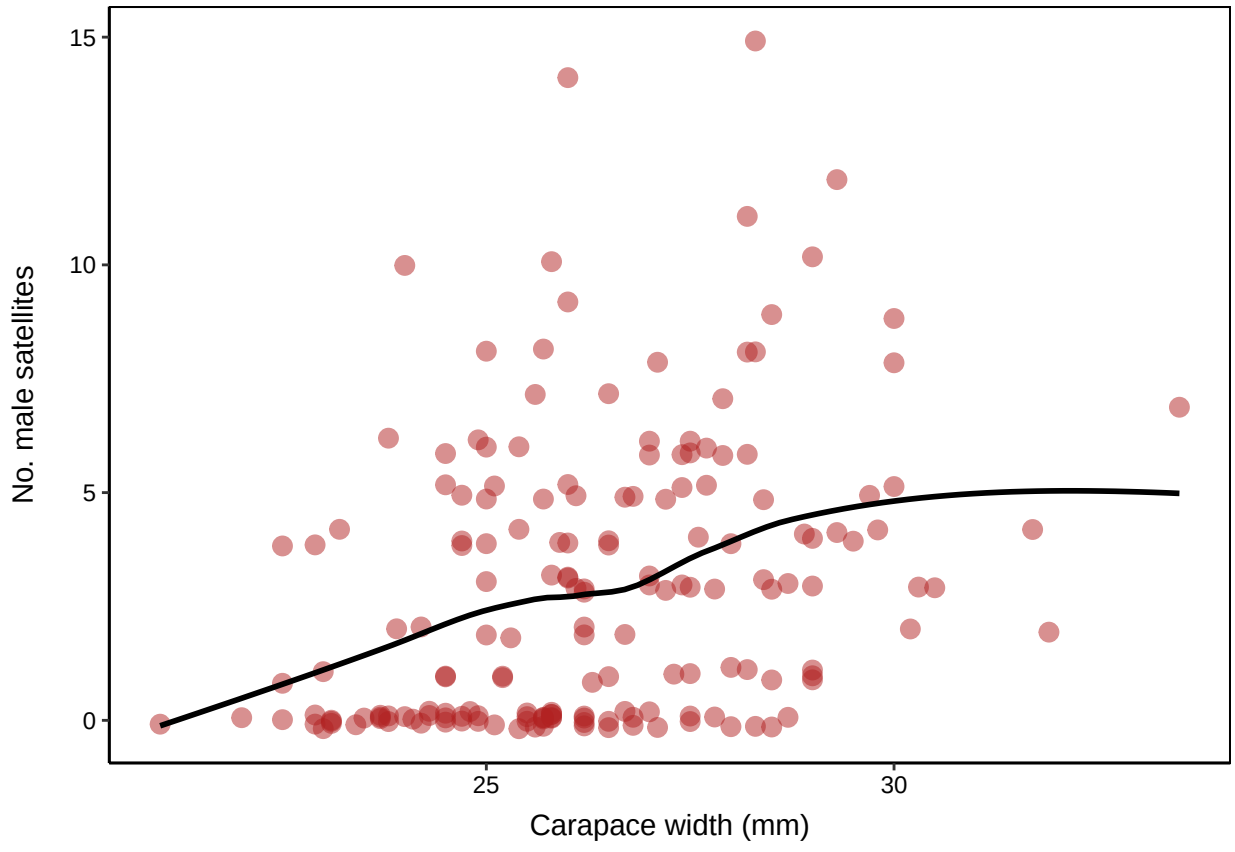
POISSON AND NEGATIVE BINOMIAL

Let's have a look at another example of modelling using counts. Brockman (1996) collected morphological data to explore female horseshoe crab (*Limulus polyphemus*) attractiveness to satellite males. The author measured female colour, spine condition, carapace width (cm), mass (kg), and the number of satellite males.

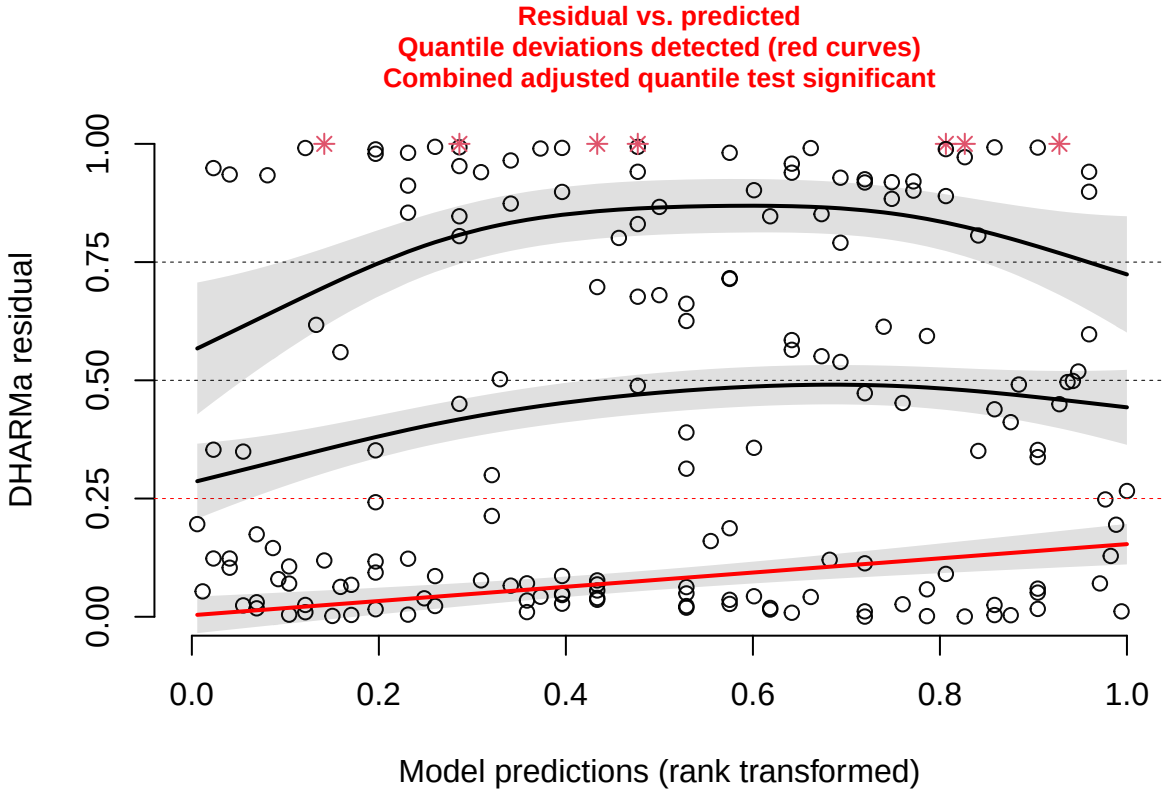
```
# read in the data
crab.data = read.csv("data/satellites.csv")
head(crab.data)
```

```
##          color    spine width.cm nsatellites mass.kg
## 1       medium both.bad     28.3           8    3.05
## 2  dark-medium both.bad     22.5           0    1.55
## 3 light-medium     good     26.0           9    2.30
## 4  dark-medium both.bad     24.8           0    2.10
## 5  dark-medium both.bad     26.0           4    2.60
## 6       medium both.bad     23.8           0    2.10
```

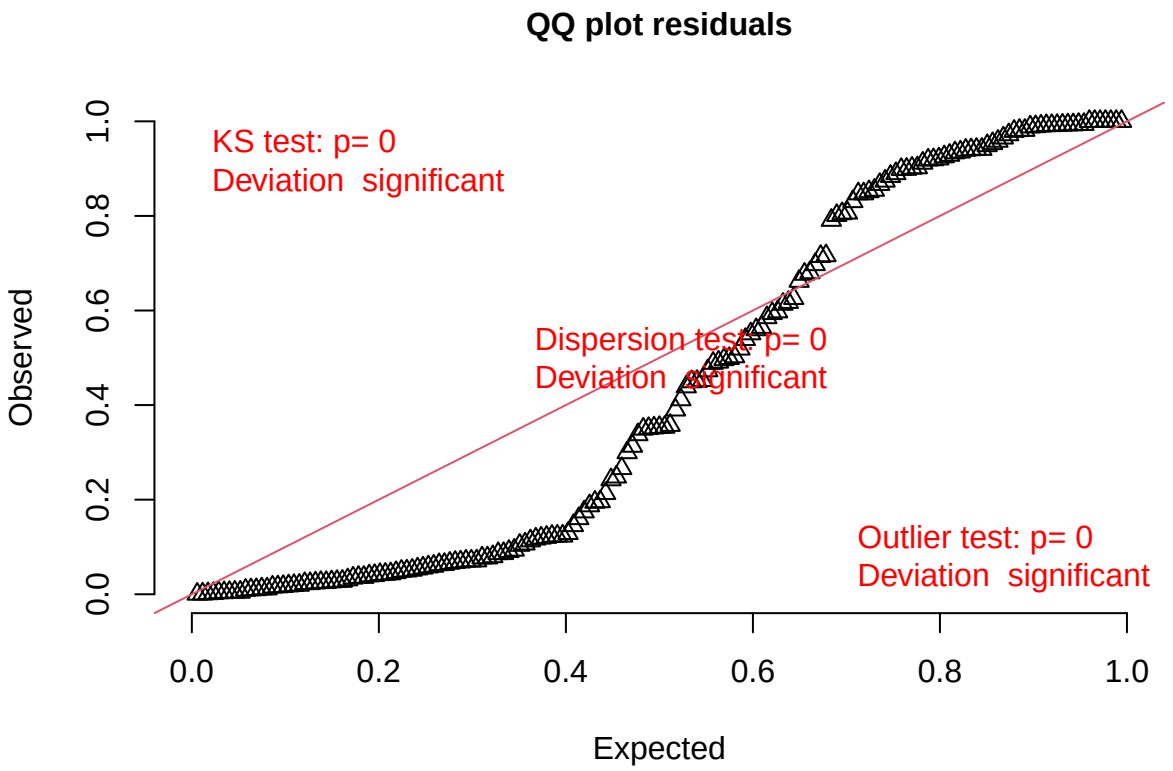
```
# let's have a look at carapace width versus the number of males
ggplot(data = crab.data, aes(width.cm, nsatellites)) +
  geom_jitter(color = "firebrick", size = 3, height = 0.2, width = 0, alpha = 0.5) +
  geom_smooth(method = "loess", linewidth = 1, col = "black", se = FALSE) +
  labs(x = "Carapace width (mm)", y = "No. male satellites")
```



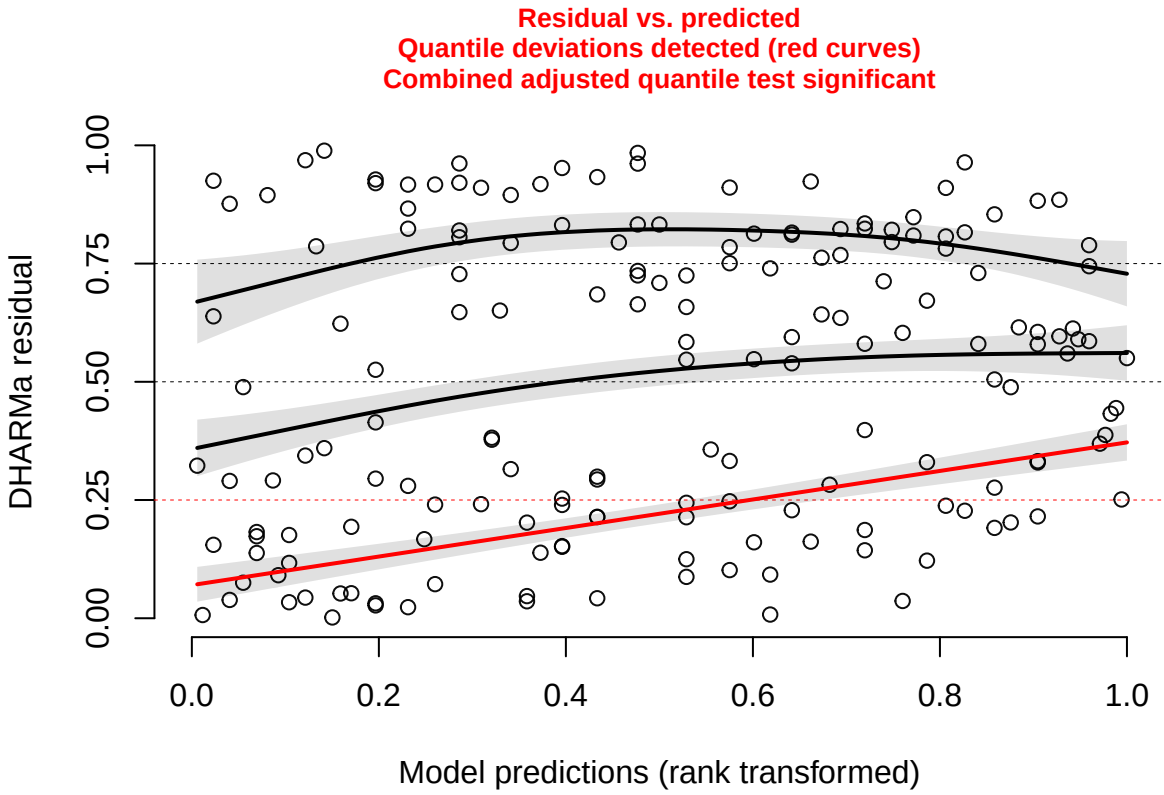
```r
# There appears to be an upward trend!
# Let's fit a poisson GLM to find out whether this is significant
crab.glm.poisson = glmmTMB::glmmTMB(nsatellites ~ width.cm,
                        family = poisson(link = "log"), data = crab.data)

# Check the residuals and QQ plot
DHARMa::plotResiduals(crab.glm.poisson)
```



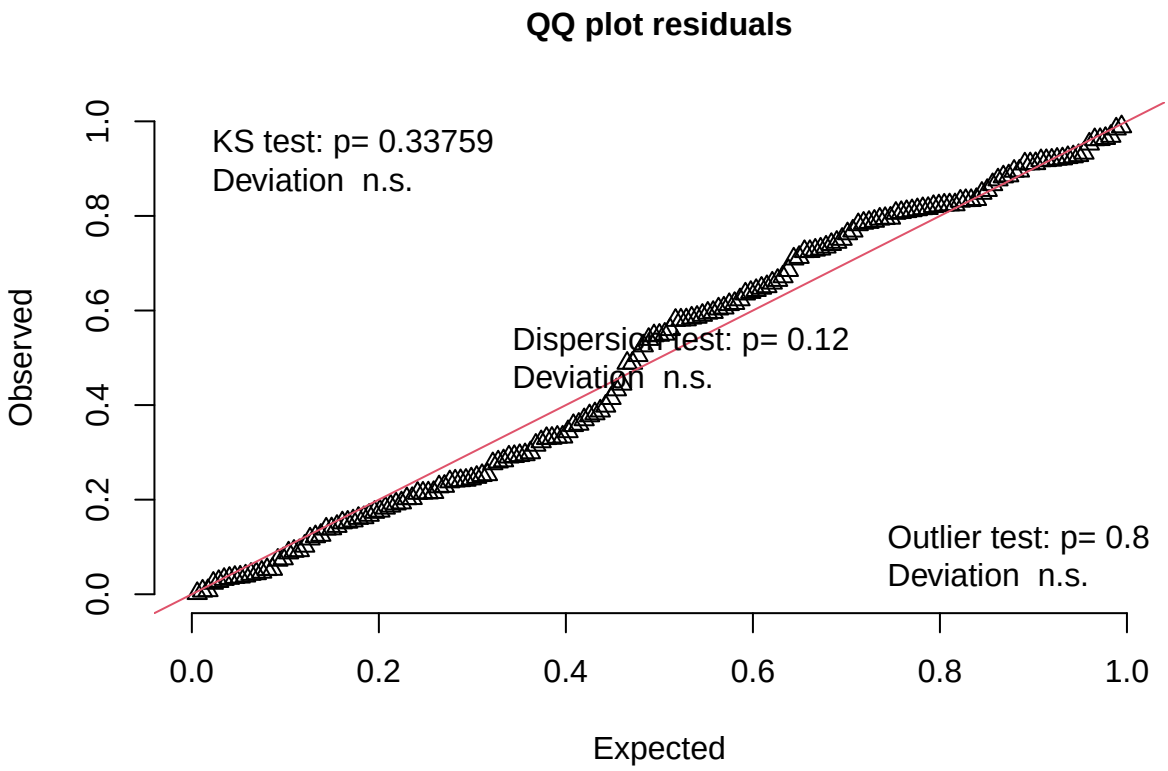

```r
DHARMa::plotQQunif(crab.glm.poisson)
```




```
# Let's fit a negative binomial model
crab.glm.negbin = glmmTMB::glmmTMB(nsatellites ~ width.cm,
                        family = nbinom2(link = "log"),
                        data = crab.data)

DHARMa::plotResiduals(crab.glm.negbin)
```




```
DHARMa::plotQQunif(crab.glm.negbin)
```




```
# another view of the summary output
gtsummary::tbl_regression(crab.glm.negbin, exponentiate = TRUE)
```

Characteristic	exp(Beta)	95% CI	p-value
width.cm	1.21	1.10, 1.33	<0.001

```
# compare models -> the negative binomial is significantly better!
anova(crab.glm.poisson, crab.glm.negbin)
```

```
## Data: crab.data
## Models:
## crab.glm.poisson: nsatellites ~ width.cm, zi=~0, disp=~1
## crab.glm.negbin: nsatellites ~ width.cm, zi=~0, disp=~1
##                  Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## crab.glm.poisson  2 927.18 933.48 -461.59   923.18
## crab.glm.negbin   3 757.29 766.75 -375.65   751.29 171.89      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# width does have a significant effect on the number of males
car::Anova(crab.glm.negbin, test = "Chisq", type = "II")
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
```

```
##
## Response: nsatellites
##           Chisq Df Pr(>Chisq)
## width.cm 16.275  1  5.479e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(crab.glm.poisson)
```

```
##  Family: poisson  ( log )
## Formula:          nsatellites ~ width.cm
## Data: crab.data
##
##      AIC      BIC   logLik deviance df.resid
##    927.2    933.5   -461.6    923.2      171
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.30476    0.54217  -6.095 1.09e-09 ***
## width.cm     0.16405    0.01996   8.218  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(crab.glm.negbin)
```

```
##  Family: nbinom2  ( log )
## Formula:          nsatellites ~ width.cm
## Data: crab.data
##
##      AIC      BIC   logLik deviance df.resid
##    757.3    766.8   -375.6    751.3      170
##
##
## Dispersion parameter for nbinom2 family (): 0.905
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.05250    1.26399  -3.206  0.00135 **
## width.cm     0.19207    0.04761   4.034 5.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# beta coefficient
exp(0.19207)
```

```
## [1] 1.211755
```

```
(1.211755 - 1)*100
```

```
## [1] 21.1755
```

```r
# 95% confidence interval
confint(crab.glm.negbin)
```

```
##                   2.5 %     97.5 %   Estimate
## (Intercept) -6.52987449 -1.5751187 -4.0524966
## width.cm     0.09875608  0.2853893  0.1920727
```

```r
exp(0.09875608)
```

```
## [1] 1.103797
```

```r
exp(0.2853893)
```

```
## [1] 1.33028
```

The beta coefficient suggests that for every cm increase in female carapace width, there will be an increase in male satellites by a factor of 1.2 (exp(0.19207)). This equates to an increased odds of an additional male by 21% **(1.211755 - 1)*100**. The 95% confidence interval suggests that a 1 cm increase in female carapace width will yield an increase in males by a factor of between 1.1 and 1.3. Since $p < 0.001$, we can conclude that female carapace width has a significant effect on the number of satellite males ($\chi^2 = 16.3$, d.f. $= 1$, $p < 0.001$).

GLM notation, and a slightly more complex example

So far we have only looked at modelling one predictor and one response variable. What if we want to look at multiple variables, with interactions?

In modeling formulae, you will often find these symbols:

Plus (+) inclusion of a variable into the model. This makes for an **additive** model

Asterisk (*) or colon (:) inclusion of a variable, and its interactions. This makes for a **multiplicative/interaction** model.

For example:

larvae ~ female_mass + female_length means that we want to model how the number of larvae are affected individually by female mass **AND** length (i.e. the effect of female mass on larval output does not depend on the effect of female length)

larvae ~ female_mass * female_length means that we want to look at how both female mass and length individually affect the number of larvae produced, **AND** whether female mass depends on female length, and vice-versa. A better way of writing this formula is:
larvae ~ female_mass + female_length + female_mass:female_length

Let's look at an example where we are interested in looking at the effect of soil pH (categorical variable; low, mid, and high) and biomass (continuous variable) on the number of species recorded on plots of land. This was taken from the R Book, by Michael Crawley.

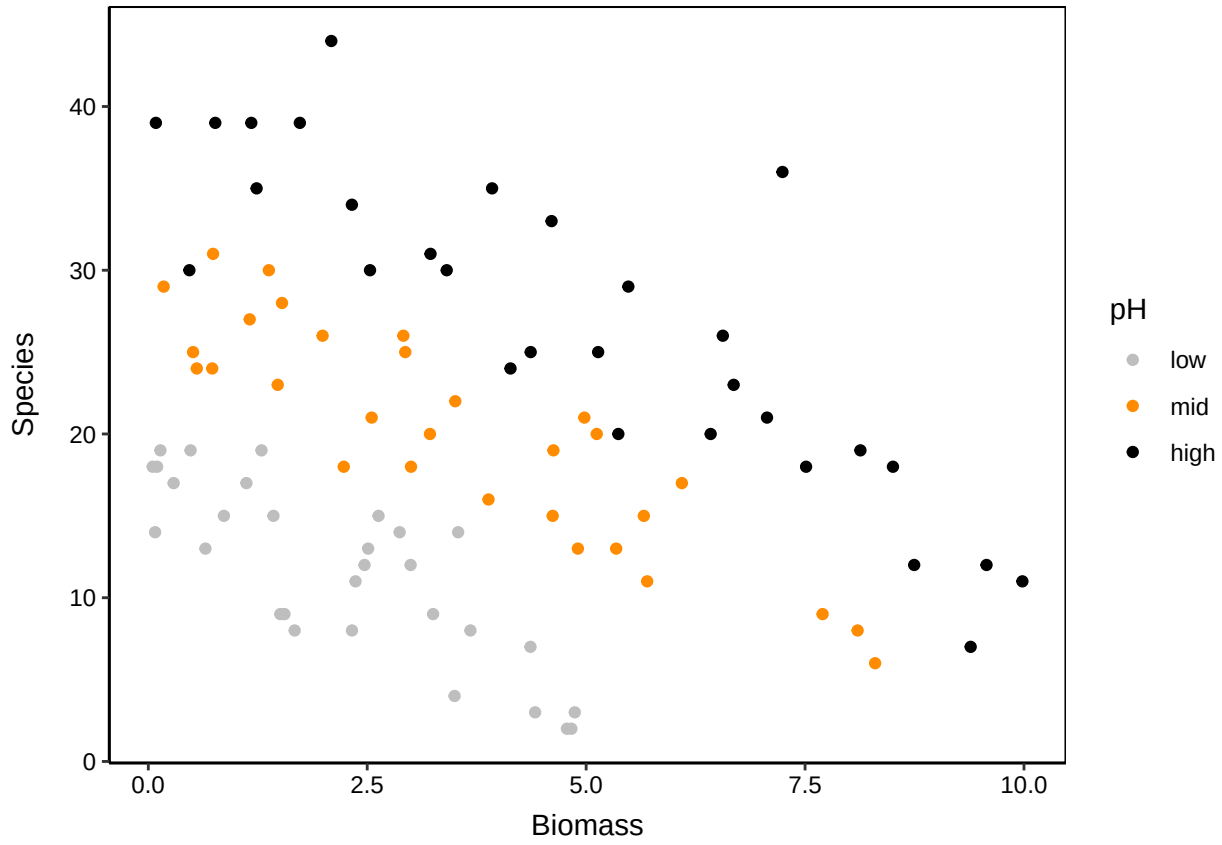
```r
plant.species.data <- read.table ("data/therbook/species.txt", header = TRUE)
head(plant.species.data)
```

```
##     pH   Biomass Species
## 1 high 0.4692972      30
## 2 high 1.7308704      39
## 3 high 2.0897785      44
## 4 high 3.9257871      35
## 5 high 4.3667927      25
## 6 high 5.4819747      29
```

```r
# make pH levels a factor, and force the order we want. Otherwise R orders
# them alphabetically
plant.species.data$pH = factor(plant.species.data$pH,
                               levels = c("low", "mid", "high"))
str(plant.species.data)
```

```
## 'data.frame':    90 obs. of  3 variables:
##  $ pH     : Factor w/ 3 levels "low","mid","high": 3 3 3 3 3 3 3 3 3 ...
##  $ Biomass: num  0.469 1.731 2.09 3.926 4.367 ...
##  $ Species: int  30 39 44 35 25 29 23 18 19 12 ...
```

```r
# Plot -> we can already see a negative relationship
ggplot2::ggplot(data = plant.species.data,
                aes(x = Biomass, y = Species,
                    colour = pH), alpha = 0.7 ) +
  scale_colour_manual(values = c("grey", "darkorange", "black")) +
  geom_point() +
  theme(legend.position = "right")
```



```r
# if you want to plot just one pH group, use dplyr to filter the data:
# dplyr::filter(plant.species.data, pH == "high")

# Let's run a Poisson GLM, since we are dealing with species counts
# Here, we will just model biomass as a predictor
plant.glm.1 = glmmTMB::glmmTMB(Species ~ Biomass,
                               data = plant.species.data,
                               family = poisson)

summary(plant.glm.1)
```

```
##  Family: poisson  ( log )
## Formula:          Species ~ Biomass
## Data: plant.species.data
##
##      AIC      BIC   logLik deviance df.resid
##    830.9    835.9   -413.4    826.9       88
##
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.184094   0.039159   81.31  < 2e-16 ***
## Biomass     -0.064441   0.009838   -6.55 5.74e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Let's add pH in as well
plant.glm.2 = glmmTMB::glmmTMB(Species ~ Biomass + pH,
                               data = plant.species.data,
                               family = poisson)

# if we compare model 1 and 2, model 2 is a significantly better fit. I.e.
# pH has a significant effect on species numbers
anova(plant.glm.1, plant.glm.2, test = "Chisq")
```

```
## Data: plant.species.data
## Models:
## plant.glm.1: Species ~ Biomass, zi=~0, disp=~1
## plant.glm.2: Species ~ Biomass + pH, zi=~0, disp=~1
##             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## plant.glm.1  2 830.86 835.86 -413.43   826.86
## plant.glm.2  4 526.43 536.43 -259.22   518.43 308.43      2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# And now let's introduce an interaction between biomass and pH
plant.glm.3 = glmmTMB::glmmTMB(Species ~ Biomass + pH + Biomass:pH,
                               data = plant.species.data,
                               family = poisson)

# compare model 2 and 3: the interaction between
# biomass and pH has a significant effect on species numbers on the plots,
```

```r
# since model 3 performs significantly better than model 2
anova(plant.glm.2, plant.glm.3, test = "Chisq", type = "III")
```

```
## Data: plant.species.data
## Models:
## plant.glm.2: Species ~ Biomass + pH, zi=~0, disp=~1
## plant.glm.3: Species ~ Biomass + pH + Biomass:pH, zi=~0, disp=~1
##             Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## plant.glm.2  4 526.43 536.43 -259.22   518.43
## plant.glm.3  6 514.39 529.39 -251.20   502.39 16.04      2  0.0003288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

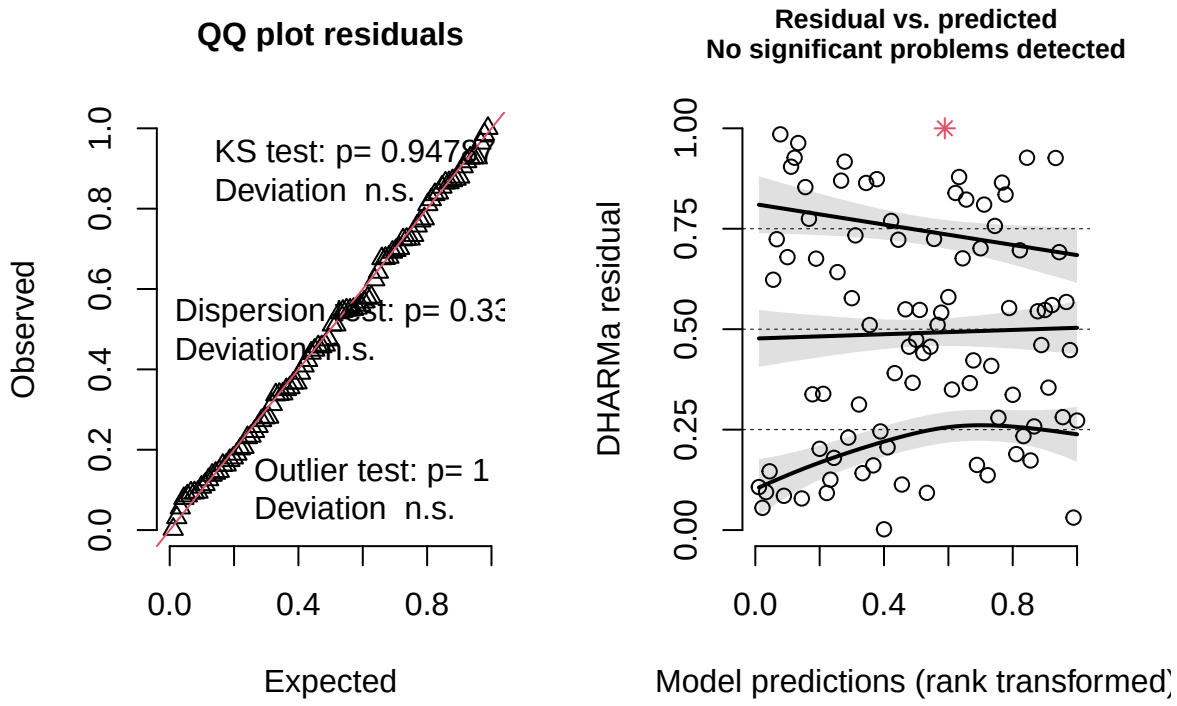
```r
# let's run a Likelihood Ratio Test (LRT) on our ADDITIVE model
# here we see that both biomass and pH have a significant effect on the number
# of species recorded
# we assume here that the relationship between biomass and species number is
# the same across all three pH levels
car::Anova(plant.glm.2, test.statistic = "Chisq", type = "III")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: Species
##               Chisq Df Pr(>Chisq)
## (Intercept) 2254.25  1  < 2.2e-16 ***
## Biomass      158.23  1  < 2.2e-16 ***
## pH           288.64  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

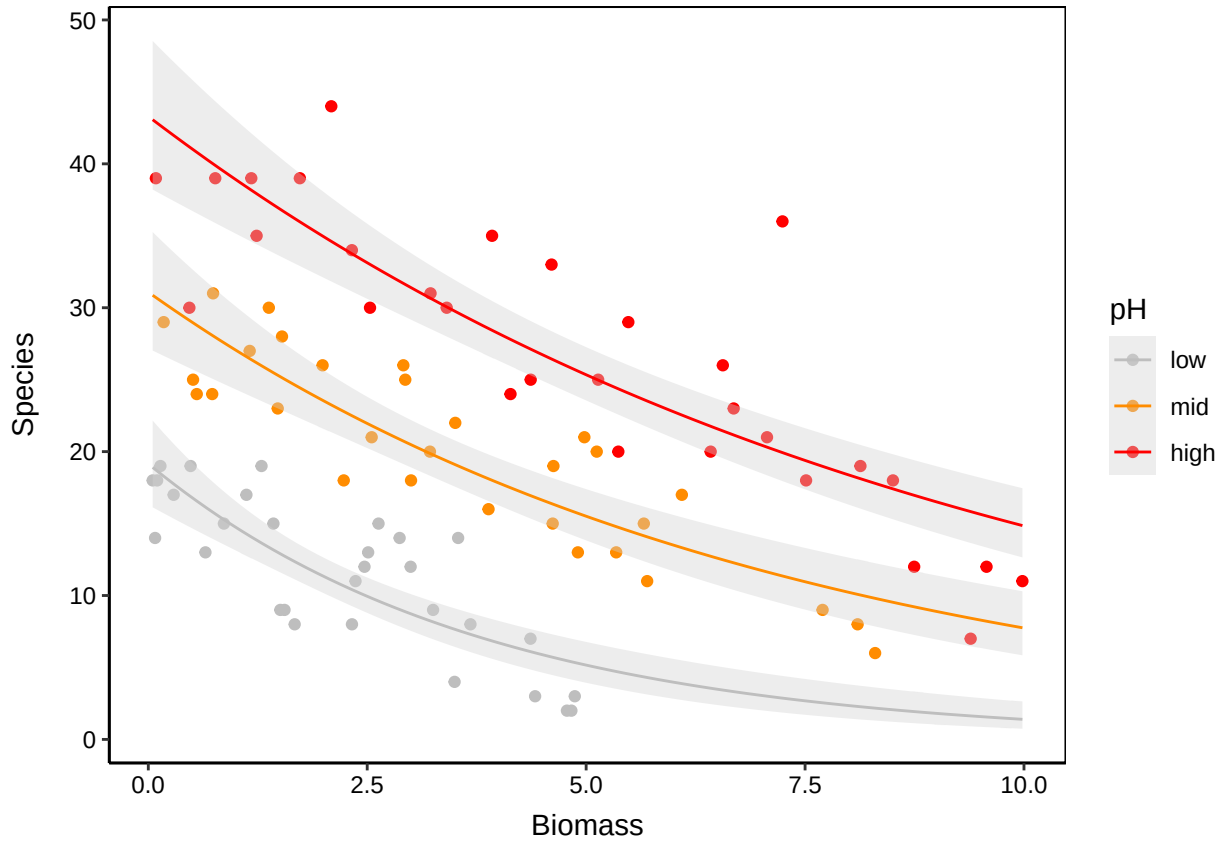
```r
# let's run a LRT on the MULTIPLICATIVE model
# here, we see a significant interaction between biomass and pH
# this means that the effect of biomass on species number differs between
# the three pH levels
# here, we see that the effect of biomass (our predictor variable)
# on species number (response variable) is allowed to vary across pH levels
car::Anova(plant.glm.3, test.statistic = "Chisq", type = "III")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: Species
##                Chisq Df Pr(>Chisq)
## (Intercept) 1284.004  1  < 2.2e-16 ***
## Biomass       47.511  1  5.470e-12 ***
## pH            63.115  2  1.971e-14 ***
## Biomass:pH    15.543  2  0.0004216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# model diagnostics
plot(DHARMa::simulateResiduals(plant.glm.3) )
```




```
# plot the fitted model
# looks like a good model to use!
ggplot2::ggplot(data = plant.species.data,
                aes(x = Biomass, y = Species,
                    colour = pH), alpha = 0.7 ) +
  scale_colour_manual(values = c("grey", "darkorange", "red")) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "poisson"),
              fill = "lightgrey", linewidth = 0.5, fullrange = TRUE) +
  theme(legend.position = "right")
```



What can you conclude from this plot? Our LR test on model 3 suggests that biomass ($\chi^2 = 47.5$, d.f. $= 1$, $p < 0.001$), pH ($\chi^2 = 63.1$, d.f. $= 2$, $p < 0.001$), and the effect of biomass on species number across pH levels($\chi^2 = 15.5$, d.f. $= 2$, $p < 0.001$) were all significant.

Let's have a look at the summary output for our third (interaction) GLM:

```
summary(plant.glm.3)
```

```
##  Family: poisson  ( log )
## Formula:          Species ~ Biomass + pH + Biomass:pH
## Data: plant.species.data
##
##      AIC      BIC   logLik deviance df.resid
##    514.4    529.4   -251.2    502.4       84
##
##
## Conditional model:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.95255    0.08240   35.83  < 2e-16 ***
## Biomass        -0.26216    0.03803   -6.89 5.47e-12 ***
## pHmid           0.48411    0.10723    4.51 6.34e-06 ***
## pHhigh          0.81557    0.10284    7.93 2.18e-15 ***
## Biomass:pHmid   0.12314    0.04270    2.88 0.003927 **
## Biomass:pHhigh  0.15503    0.04003    3.87 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# another way of presenting the output:
gtsummary::tbl_regression(plant.glm.3, exponentiate = TRUE)
```

Characteristic	exp(Beta)	95% CI	p-value
Biomass	0.77	0.71, 0.83	<0.001
pH			
low	—	—	
mid	1.62	1.32, 2.00	<0.001
high	2.26	1.85, 2.77	<0.001
Biomass * pH			
Biomass * mid	1.13	1.04, 1.23	0.004
Biomass * high	1.17	1.08, 1.26	<0.001

The summary output for a more complex model can be tricky to interpret. Here, we can say that in the low pH group, when biomass = 0, species richness is 19.15 (**exp(2.95255)**) (**y Intercept**). In the low pH group, for every unit increase in biomass, species number **decreases** by a log odds of **0.26216** -> which is an odds ratio of 0.77 (exp(**-0.26216**)). I.e. species number decreases by a factor of 0.77 for every unit increase in biomass (in the low pH group), or put another way, biomass decreases by 23% (1-0.77) x 100.

The **pHmid** value of **0.48411** means that species richness increases by a factor of 1.62 (**exp(0.48411)**) more than the pH low group, when biomass = 0 (i.e. the y-intercept of the pH mid group is 19.5 * 1.62 = 31.59). Similarly, the **pHhigh** value of **0.81557** means that species richness increases by a factor of 2.26 (**exp(0.81557)**) more than the pH low group, when biomass = 0 (i.e. the y-intercept of the pH high group is 19.5 * 2.26 = 44.07).

The **Biomass:pHmid** value of **0.12314** means that a one unit increase in biomass in the pH mid group will result in an increase in species richness by a factor of 1.13 (**exp(0.12314)**) more than the pH low group. Similarly, the value of **0.15503** in the **Biomass:pHhigh** means that a one unit increase in biomass in the pH high group will result in an increase in species richness by a factor of 1.17 (**exp(0.15503)**) more than the pH low group.

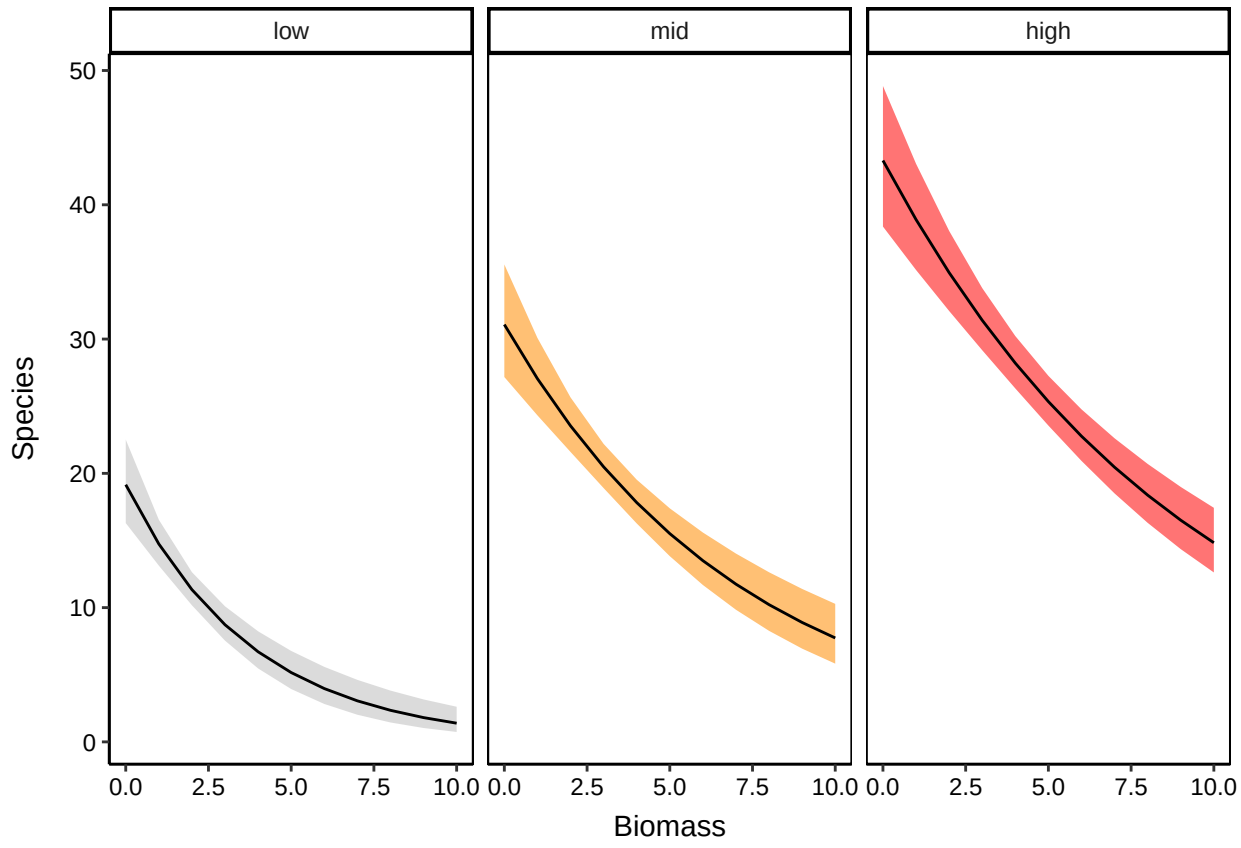
Let's look at marginal predictions for our three pH groups:

```r
preds.spp = ggeffects::ggpredict(plant.glm.3, terms = c("Biomass", "pH"))

preds = as.data.frame(preds.spp) %>%
  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.spp)
```

```
## # Predicted counts of Species
##
## pH: low
##
## Biomass | Predicted |       95% CI
## ---------------------------------
##       0 |     19.15 | 16.30, 22.51
```

```
##       1 |     14.74 | 13.13, 16.54
##
## pH: mid
##
## Biomass | Predicted |       95% CI
## ---------------------------------
##       0 |     31.08 | 27.17, 35.56
##       1 |     27.05 | 24.32, 30.08
##
## pH: high
##
## Biomass | Predicted |       95% CI
## ---------------------------------
##       0 |     43.30 | 38.38, 48.85
##       1 |     38.90 | 35.14, 43.06
```

```r
# plot
ggplot2::ggplot(data = dplyr::filter(preds.spp), aes(x = x, y = predicted)) +
  geom_ribbon(aes(ymin = conf.low, ymax = conf.high, fill = group, alpha = 0.5)) +
  scale_fill_manual(values = c("grey", "darkorange", "red")) +
  geom_line(aes(y = predicted)) +
  labs(x = "Biomass",
       y = "Species") +
  facet_wrap(~group)
```



Practice exercise 1 [optional]

The data set **pollinator_abundance_data.csv** contains pollinator abundances on three different flower species, across three seasons (summer, autumn, and winter). Run the appropriate GLM on this data, and have a look at whether:

(1) abundance differs across the three flower hosts

(2) abundance differs across seasons

(3) season affects abundance differently across the three flower species

Additionally, answer the following:

(1) What is the response variable?

(2) What is the predictor variable?

(3) Do we need to include an interaction term in this model?

(4) Which GLM family is appropriate for this data? Provide the relevant model diagnostic plots. (5) Provide box-and-whisker plots to graphically present the data

Practice exercise 2 [optional]

Have a look at the lung capacity data set from Kahn 2017, called `lungcap` in the `GLMsData` package.

(1) Generate a few box plots and scatter plots to explore the data. Plot FEV (lung capacity) in smokers vs non-smokers, and then account for height and age. What are your observations? Are they what you would expect?
(2) Fit some Gaussian GLMs with different predictors and interaction terms, and find a suitable model
(3) Interpret the results, and suggest reasons for any oddities

Practice exercise 3 [optional]

Use the `cheese` dataset in the `GLMsData` package to find out what makes a winning Cheddar cheese! The data comes from Moore and McCabe (1993).

Find out whether acetic acid, H2S, or lactic acid has an effect on taste scores by running a Gaussian GLM. Does the model fit well? Which ingredient/s had a significant effect on scores?

Practice exercise 4 [optional]

Use the `deposit` dataset in the `GLMsData` package to analyse the effect of three different insecticides at different dosages on insect mortality. Which insecticide, at which dosage, was the most effective? What type of GLM is best to use here?

Submit your answers to me at **clarke.vansteenderen@ru.ac.za** if you would like feedback.