Basic statistical modelling in R GLMs (Generalised Linear Models)

Clarke van Steenderen
Department of Zoology and Entomology
Rhodes University
South Africa

Clarke.vansteenderen@ru.ac.za

Generalised Linear Models (GLMs)

Linear models assume normality in residuals/error terms → i.e. a constant relationship between mean and variance. But what if this is not met?

We can apply different error structures to a linear model to account for deviations from normality. This means that we can use a GLM, and specify the "family" type depending on the type of data in question

Common families are:

- Binomial → two-state data
- Poisson → count data
- Negative binomial → count data that is overdispersed
- Gaussian → normally-distributed; essentially an LM or an ANOVA
- Exponential/beta → proportions

An ANOVA is a LM or GLM (Gaussian)
ANOVAs have one or more categorical predictors (1 or 2-way) and a continuous response variable

NB: offsets in GLMS → factors that are not strictly predictors, but are used to correct the model (e.g. sampling effort, population size and structure)

Sum-of-square tests assess the significance of each parameter in the model → use `anova(model)` or `car::Anova(model, type = x)`

Where `type = I, II, or III`
I: One predictor variable
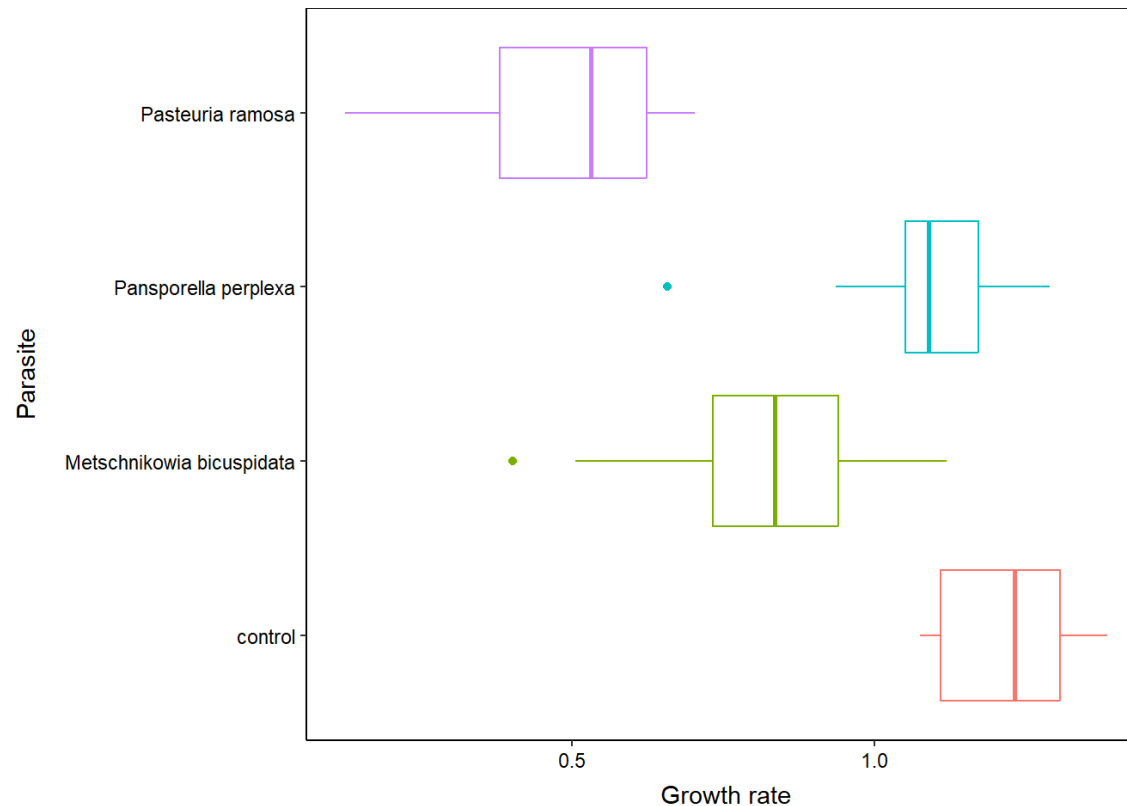II: Two or more predictors with no interaction terms
III: Two or more predictors with interaction terms

GLM: Gaussian

Growth rates (continuous) of *Daphnia* water fleas infected with different parasite species (categorical)
GLMs with a Beta distribution are sometimes used for rate data, but a Gaussian GLM is simpler, if it meets all the assumptions

```
daphnia.glm.gaus = glm(data = daphnia.data,
growth.rate ~ parasite,
family = gaussian)
```
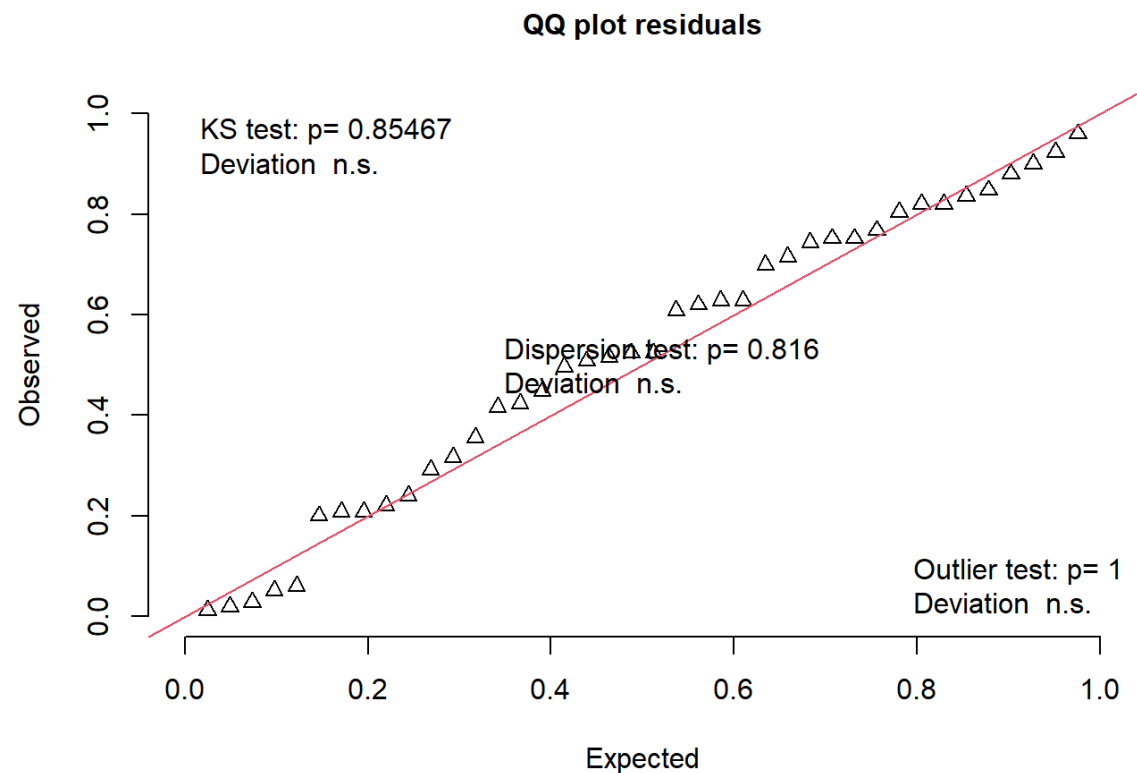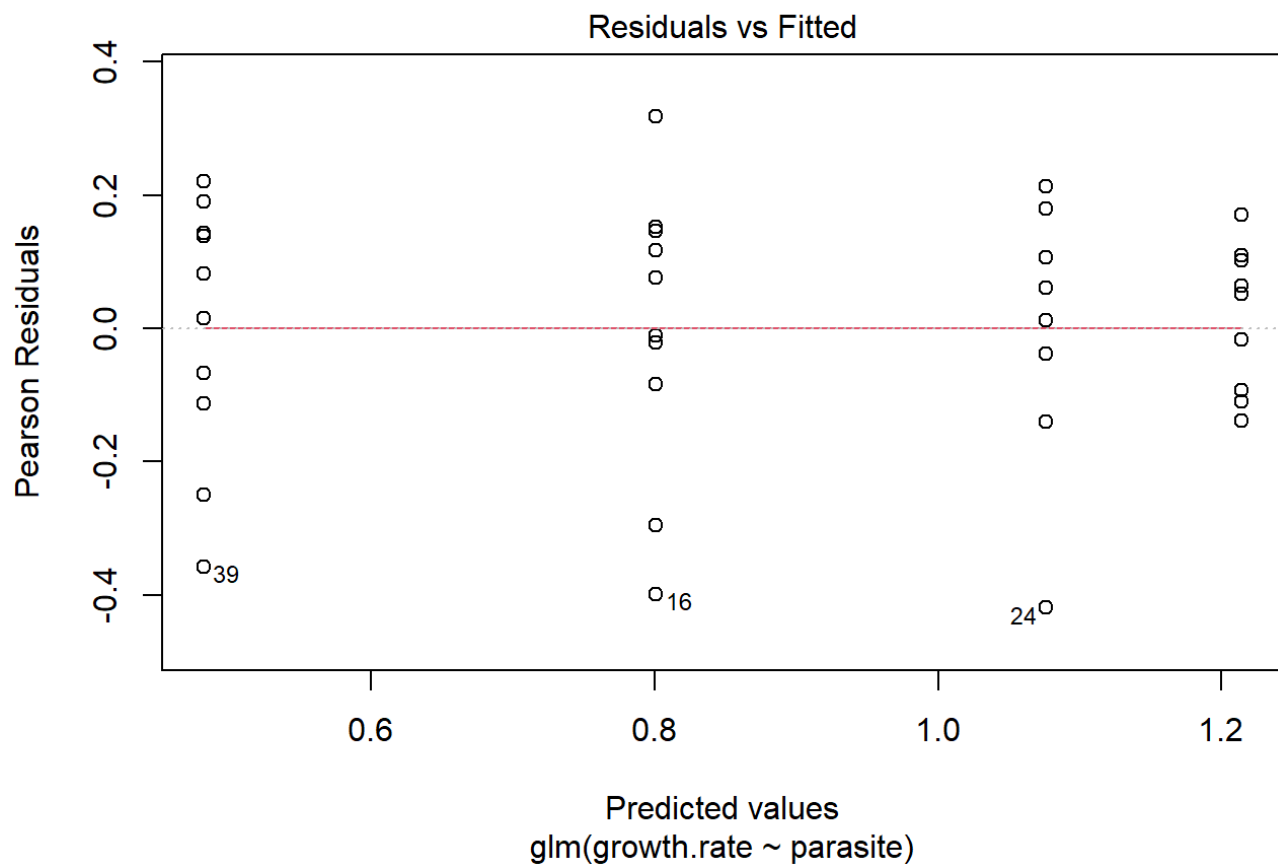
```
anova(daphnia.glm.gaus, test = "LR")
```

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

Growth rates vary significantly across parasite types

```
> daphnia.means = Rmisc::summarySE(daphnia.data, measurevar = "growth.rate",
+                                  groupvars = "parasite")
> daphnia.means
                 parasite  N growth.rate        sd         se        ci
1                 control 10   1.2139088 0.1141824 0.03610766 0.0816812
2 Metschnikowia bicuspidata 10   0.8011541 0.2158424 0.06825535 0.1544043
3    Pansporella perplexa 10   1.0763551 0.1795313 0.05677277 0.1284289
4        Pasteuria ramosa 10   0.4822030 0.1938393 0.06129737 0.1386643
```
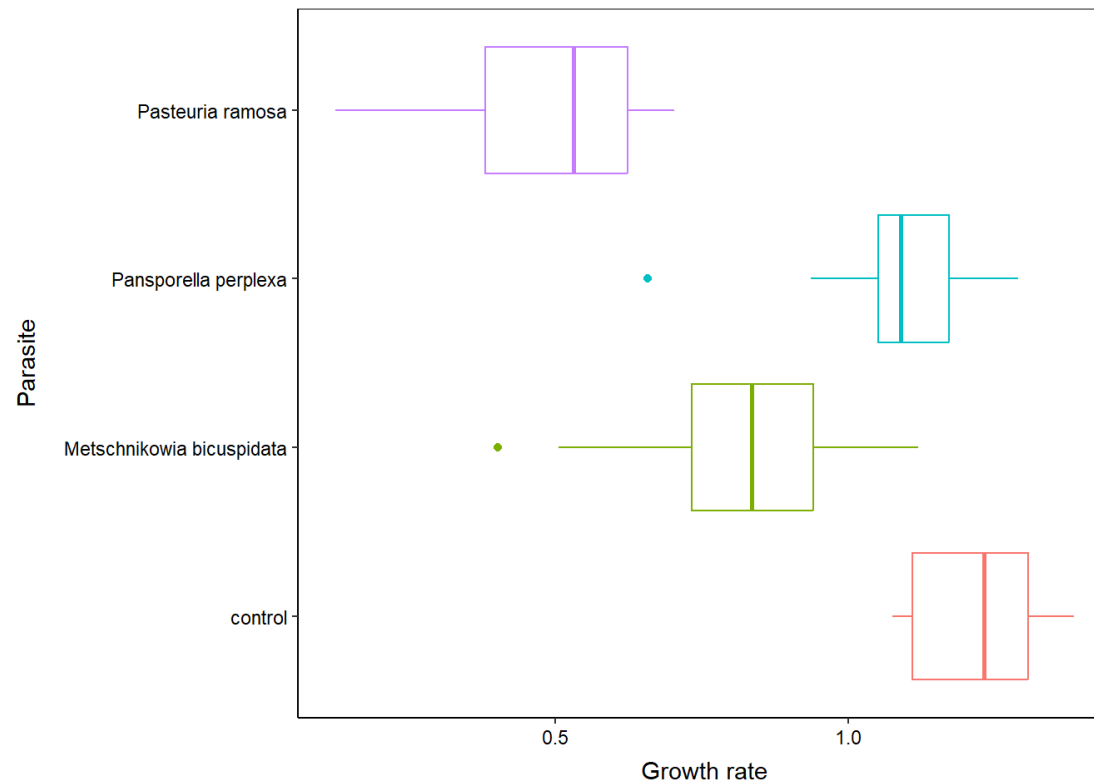
- The control group is the first "level" alphabetically, and is used as the base category for comparison → this is the (Intercept), and represents the baseline level for the control (growth rate = 1.21391). This is only the case when there is a single categorical predictor, and here, this value is then also the mean
- The other Estimate values show how far each other category/parasite is from the Control (i.e. relative distance)

```
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
```

- *M. bicuspidata* is 0.41275 units less than the control: i.e. 1.21391 – 0.41275 = 0.80116, which is significant (p < 0.001)

```
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
```
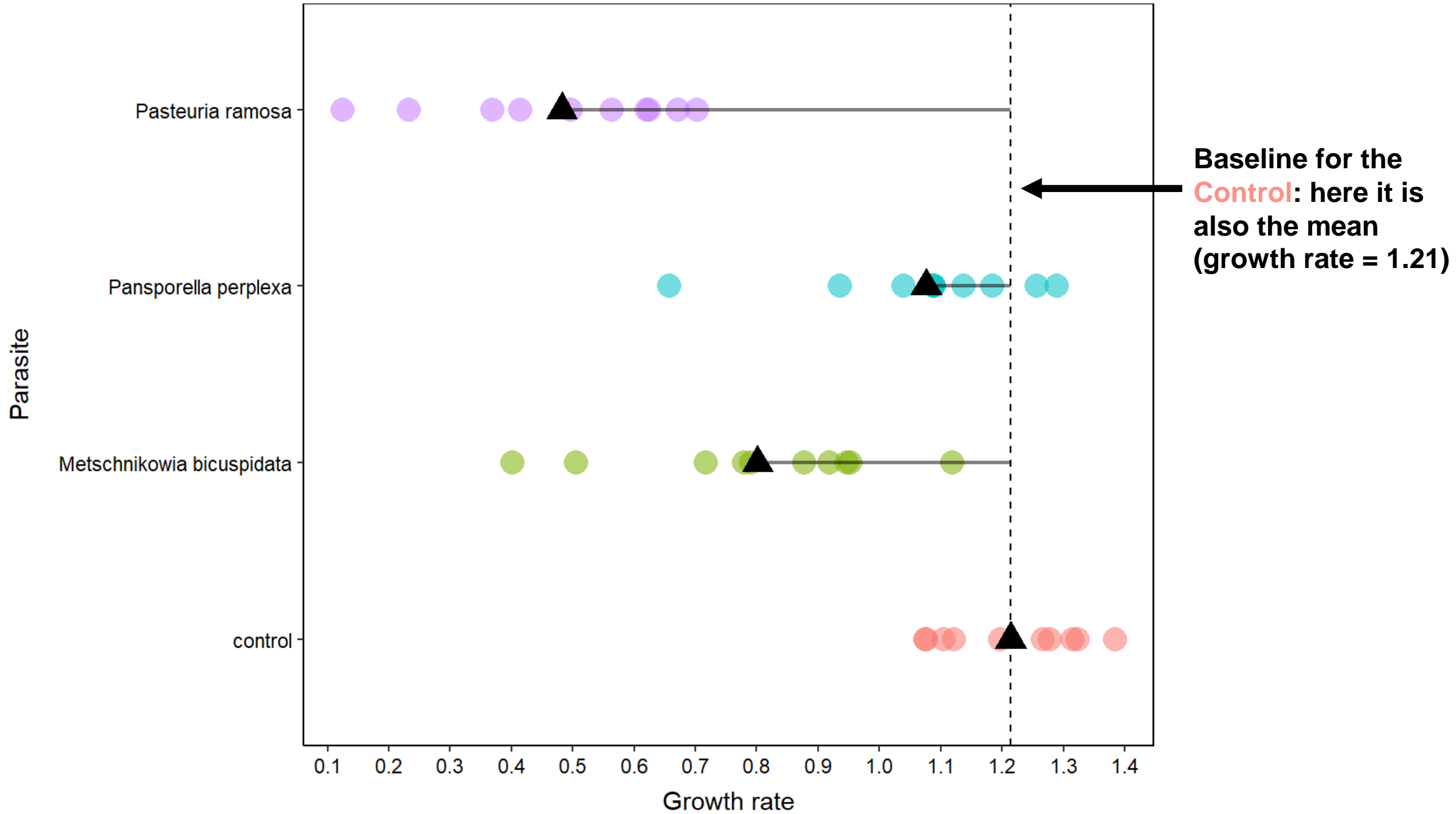
Baseline for the **Control**: here it is also the mean (growth rate = 1.21)

What if we had another variable, in addition to parasite?

```
glm(formula = growth.rate ~ parasite + temperature,
family = gaussian, data = daphnia.data)
```

Hypothetical output:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.50000    0.20000   7.500  < 2e-16 ***
parasiteMetschnikowia bicuspidata -0.30000  0.08000  -3.750  0.0002 ***
parasitePansporella perplexa     -0.10000    0.08500  -1.176  0.2417
parasitePasteuria ramosa         -0.50000    0.09000  -5.556  < 2e-16 ***
temperature                       0.05000    0.01000   5.000  5.6e-06 ***
```

Here, the baseline for the Control would be a mean growth rate of 1.5 when temperature = 0

M. bicuspidata, at 0 degrees, would have a growth rate 0.3 units less → i.e. 1.5 – 0.3 = 1.2

The temperature coefficient (β_1 = 0.05) means that for every 1 degree increase, growth rate increases by 0.05 across all parasites

We know that growth rates vary significantly across parasites, but which parasites exactly?

Post-hoc tests show comparisons between all our variables, and highlight which ones are significant. The Tukey test is a very popular choice
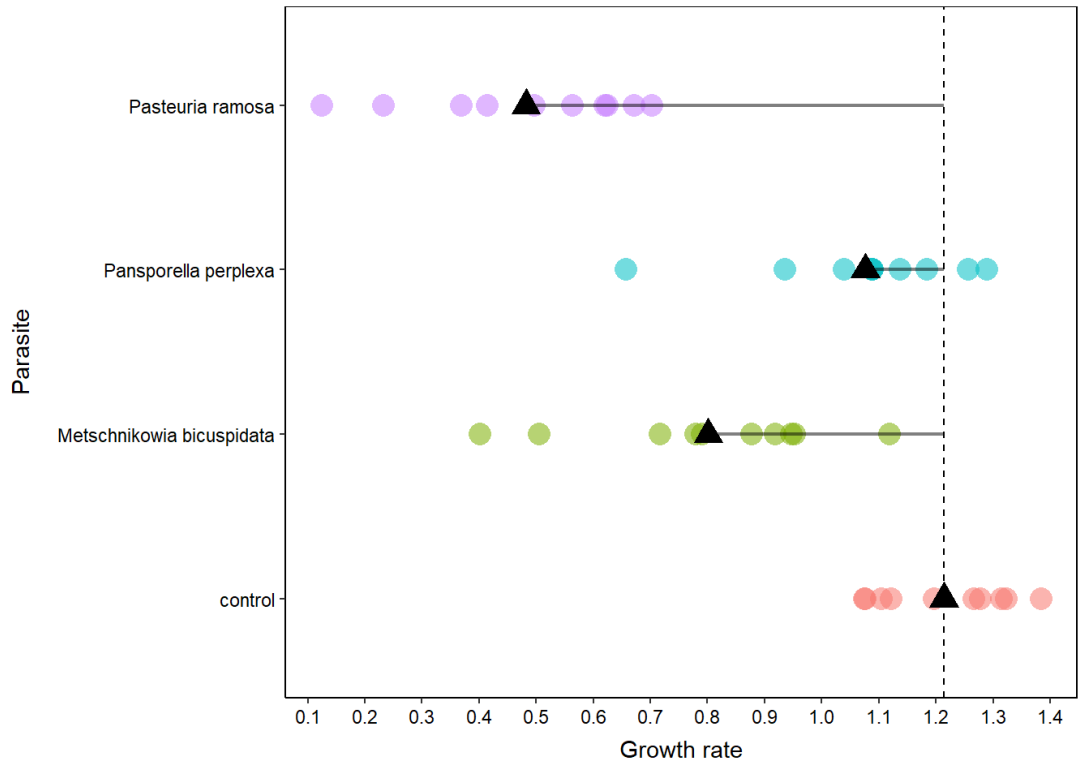
```
posthoc.daphnia = emmeans::emmeans(daphnia.glm.gaus, pairwise ~ parasite, adjust = "tukey")
```

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
control - Metschnikowia bicuspidata	0.413	0.0804	36	0.1961	0.6294	5.131	0.0001
control - Pansporella perplexa	0.138	0.0804	36	-0.0791	0.3542	1.710	0.3335
control - Pasteuria ramosa	0.732	0.0804	36	0.5150	0.9484	9.096	<.0001
Metschnikowia bicuspidata - Pansporella perplexa	-0.275	0.0804	36	-0.4919	-0.0585	-3.421	0.0082
Metschnikowia bicuspidata - Pasteuria ramosa	0.319	0.0804	36	0.1023	0.5356	3.965	0.0018
Pansporella perplexa - Pasteuria ramosa	0.594	0.0804	36	0.3775	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

The estimate values (magnitude and sign) indicate differences in size – compare to the plot

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
control - Metschnikowia bicuspidata	0.413	0.0804	36	0.1961	0.6294	5.131	0.0001
control - Pansporella perplexa	0.138	0.0804	36	-0.0791	0.3542	1.710	0.3335
control - Pasteuria ramosa	0.732	0.0804	36	0.5150	0.9484	9.096	<.0001
Metschnikowia bicuspidata - Pansporella perplexa	-0.275	0.0804	36	-0.4919	-0.0585	-3.421	0.0082
Metschnikowia bicuspidata - Pasteuria ramosa	0.319	0.0804	36	0.1023	0.5356	3.965	0.0018
Pansporella perplexa - Pasteuria ramosa	0.594	0.0804	36	0.3775	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

GLM: binomial

Schluter & Smith (1986) had a look at whether different morphological traits of song sparrows had an effect on survival. We'll focus on tarsal length:
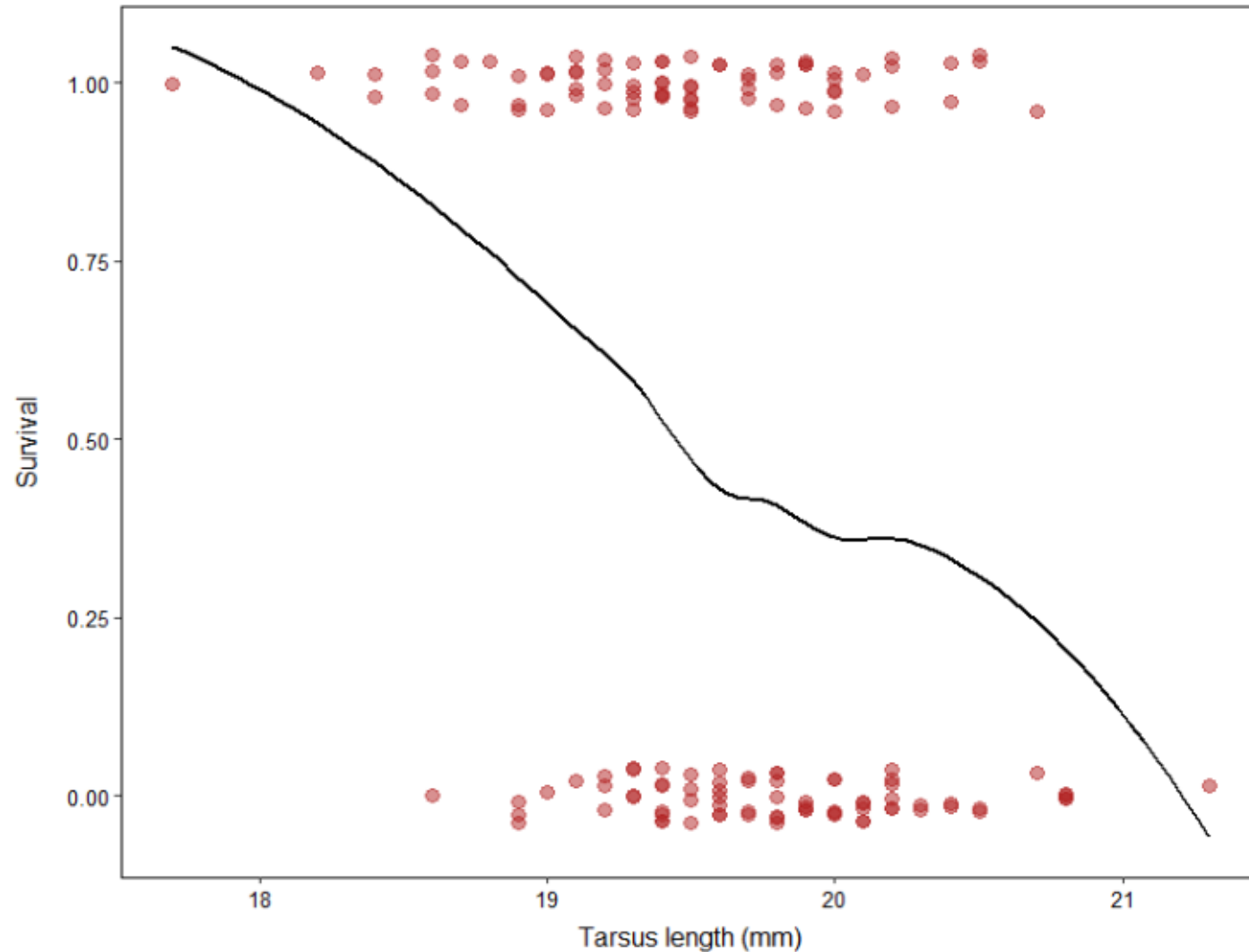
```
> head(sparrow.data)
  mass wing tarsus blength bdepth bwidth year sex survival predictions
1 23.7 67.0   17.7     9.1    5.9    6.8 1978   f        1   0.9147634
2 23.1 65.0   19.5     9.5    5.9    7.0 1978   f        0   0.5272762
3 21.8 65.2   19.6     8.7    6.0    6.7 1978   f        0   0.4958586
4 21.7 66.0   18.2     8.4    6.2    6.8 1978   f        1   0.8512375
```

What relationship do you notice?
Is it good for these sparrows to have small or large feet? Why?

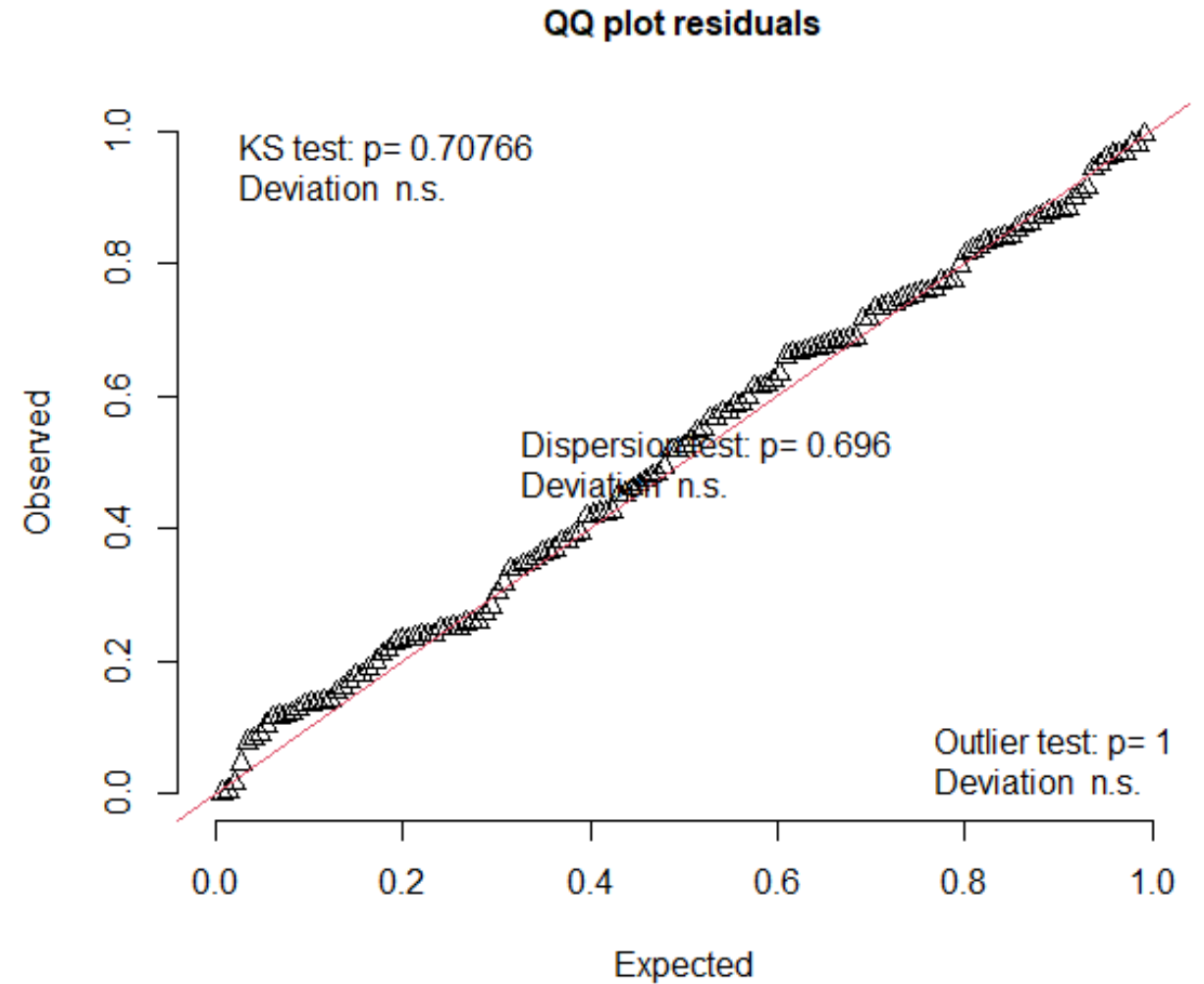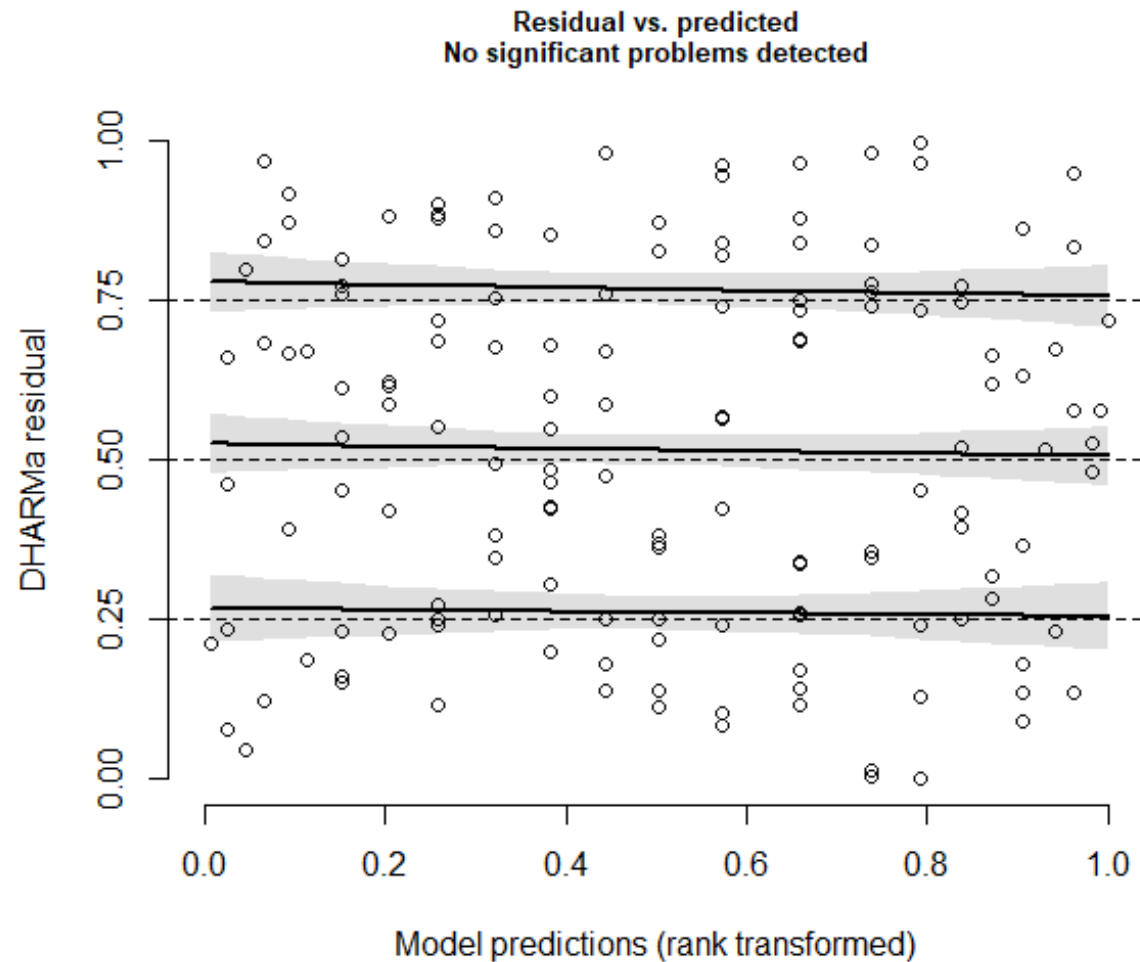**Since we have binomial data (dead or alive), we'll use a binomial GLM using the `glmmTMB` package**

```
sparrow.glm = glmmTMB::glmmTMB(survival ~ tarsus,
family = binomial(link = "logit"),
data = sparrow.data)
```

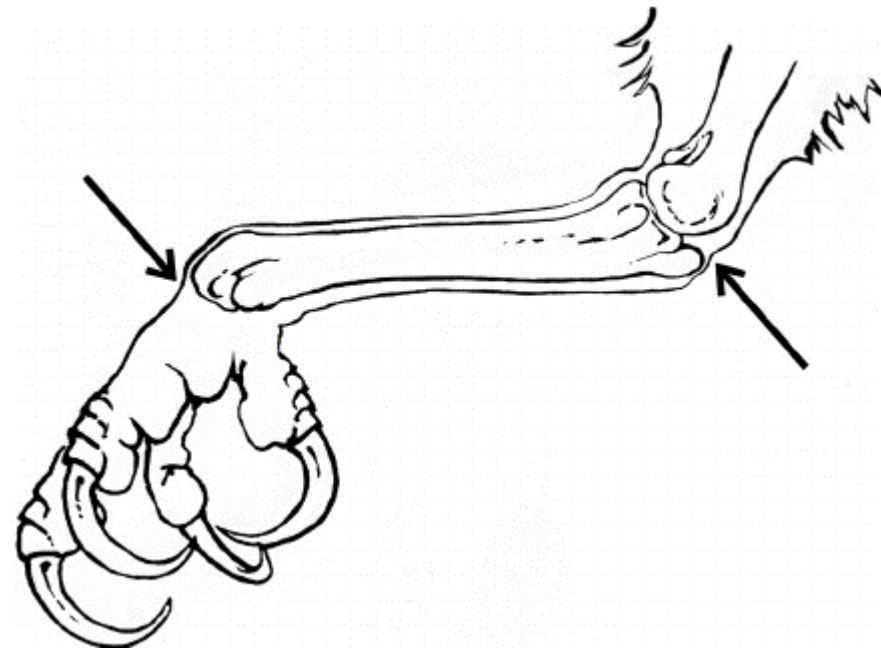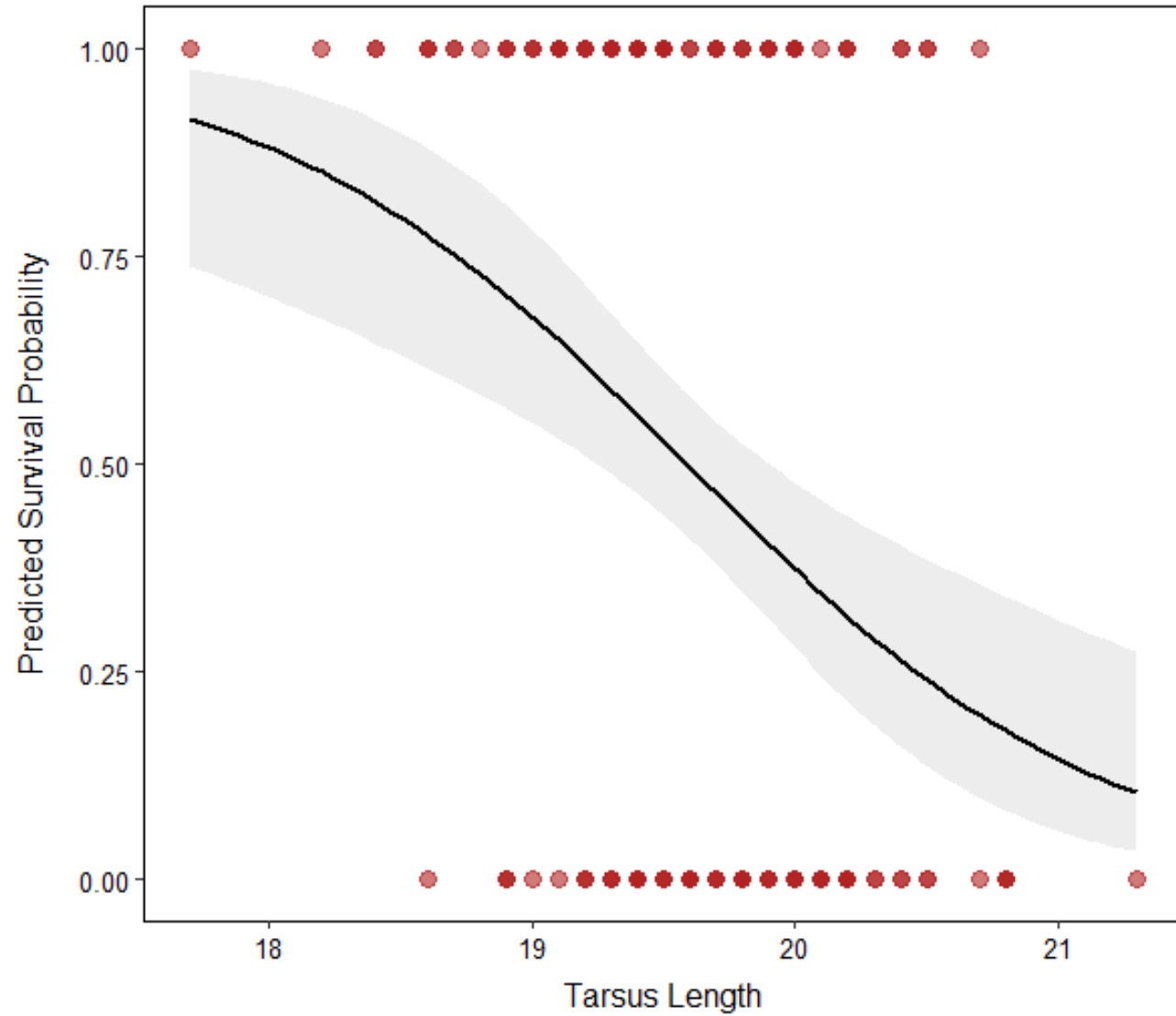**Check the model diagnostics using the `DHARMa` package**

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

Looks like a good model 👍

Residual vs. predicted
No significant problems detected

DHARMa residual vs. Model predictions (rank transformed)

QQ plot residuals

KS test: p= 0.70766
Deviation n.s.

Dispersion test: p= 0.696
Deviation n.s.

Outlier test: p= 1
Deviation n.s.

Observed vs. Expected

Plot our binomial GLM:

Check for significance using a Wald Chi-square test:

```
> 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>
```

Tarsus length has a significant effect on survival (χ^2 = 13.4, d.f. = 1, p < 0.001)

Let's look at the beta coefficients:

```
>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 β_0   6.7454    3.652 0.000260 ***
tarsus       -1.2578 β_1   0.3437   -3.659 0.000253 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since this was a logistic regression, we need to get the exponentiated value for the β_1 coefficient (log odds):

exp(-1.2578) = 0.28 (this is now the odds ratio. Note the - sign)
For every mm ↑ in tarsus length, the odds of survival ↓ by a factor of 0.28
As a percentage: (0.28 – 1) x 100 = -72% (i.e. a 72% decrease in the odds of survival)

95% confidence interval:
Lower = exp(-1.97) = 0.14
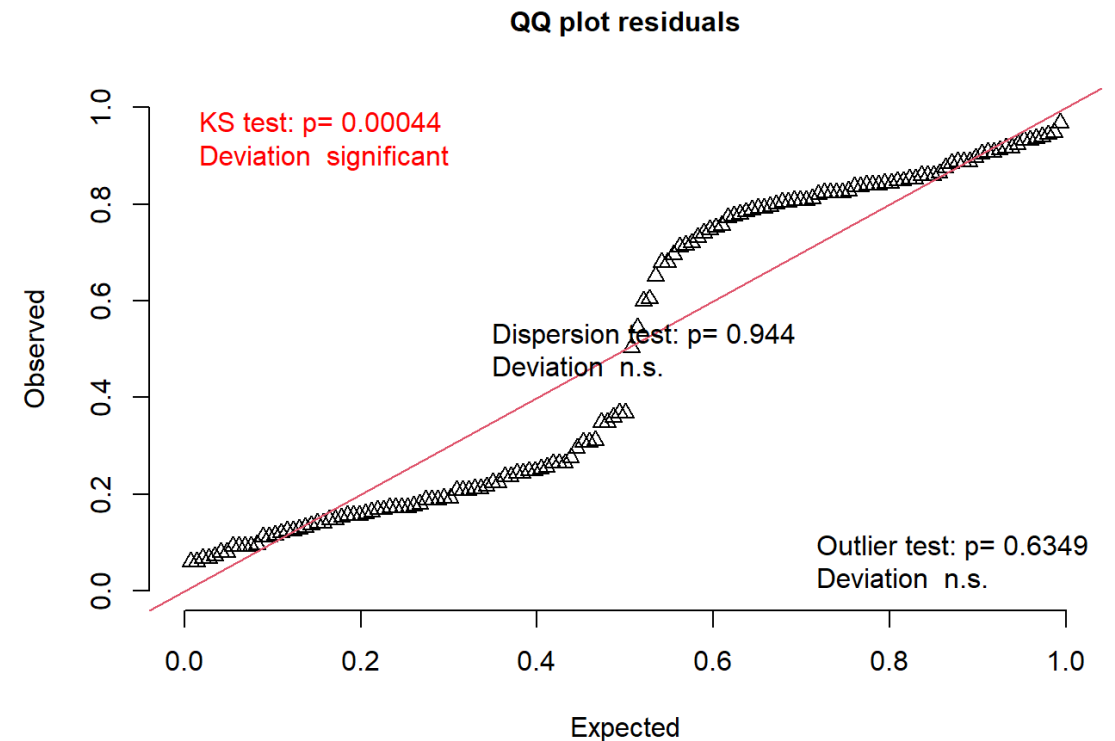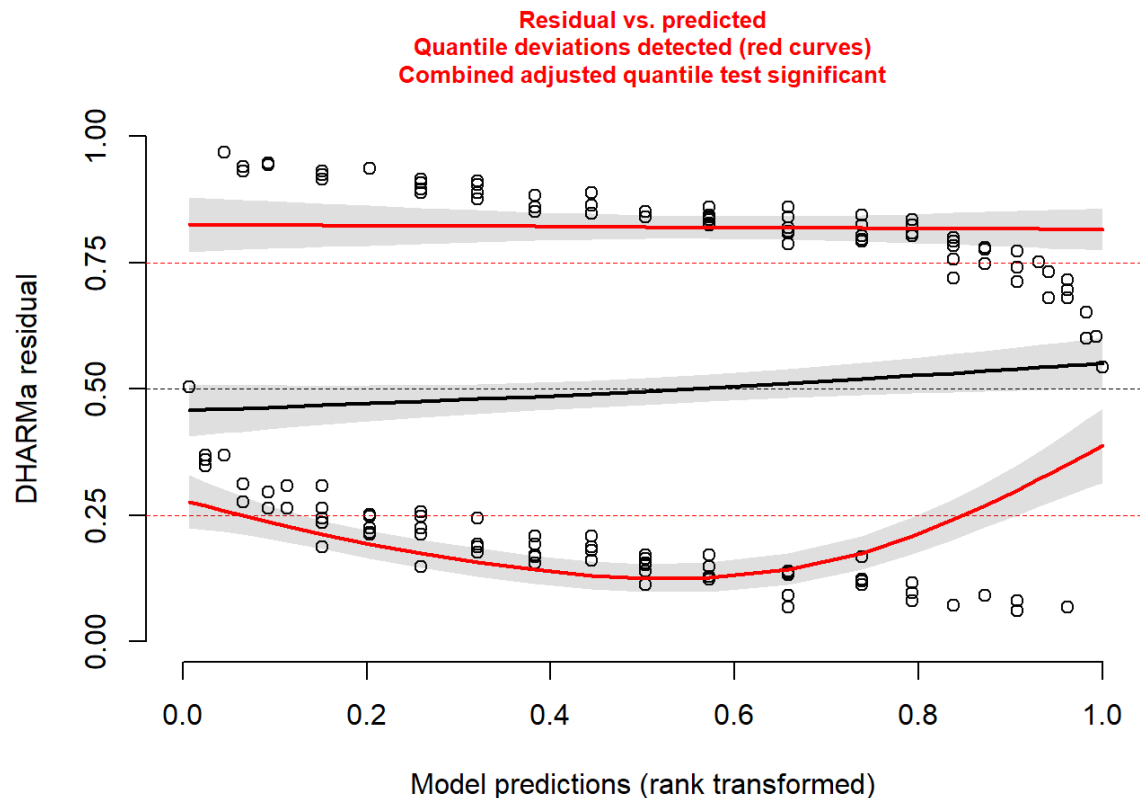Upper = exp(-0.58) = 0.56

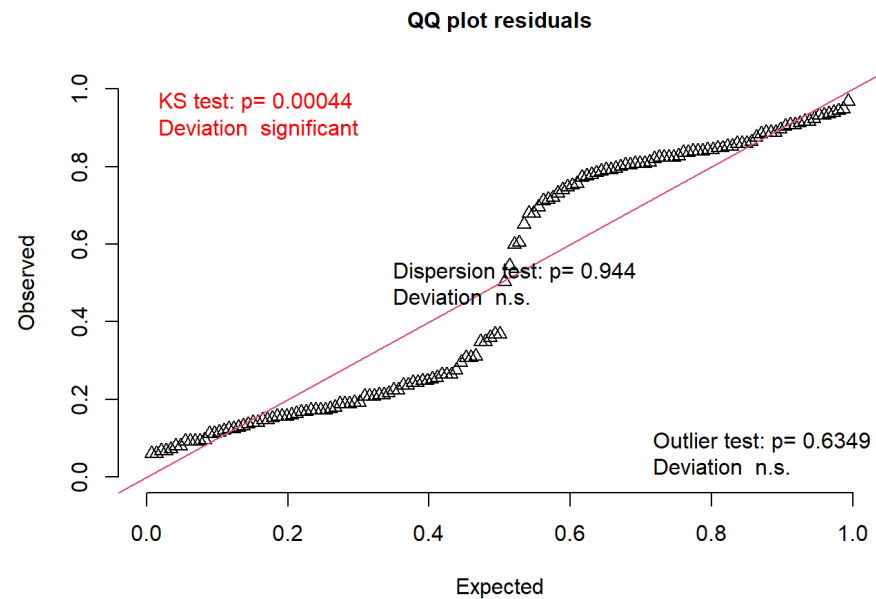| Characteristic	exp(Beta)	95% CI[1]	p-value
tarsus	0.28	0.14, 0.56	<0.001

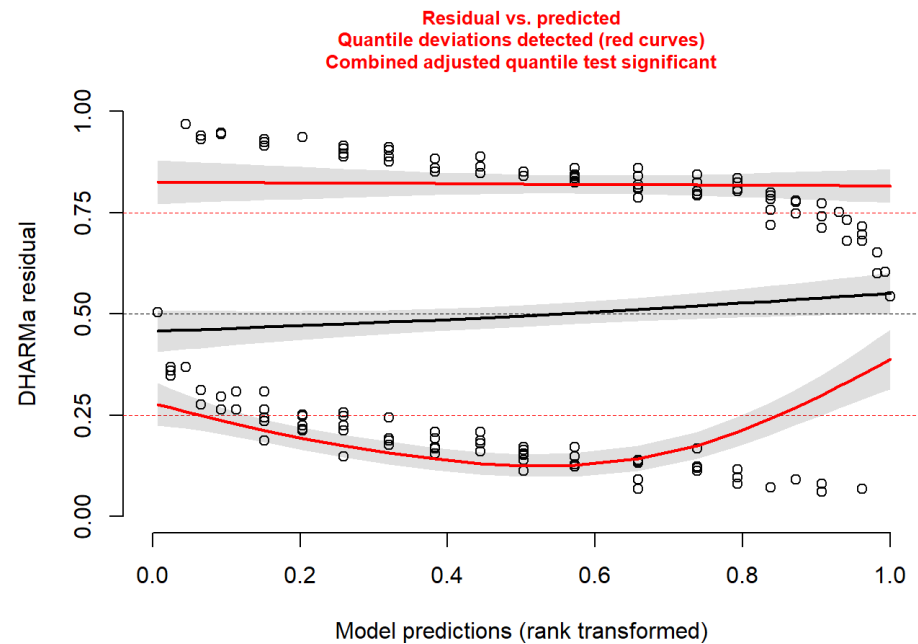[1] CI = Confidence Interval

95 out of 100 times, we will get an odds ratio between 0.14 and 0.56

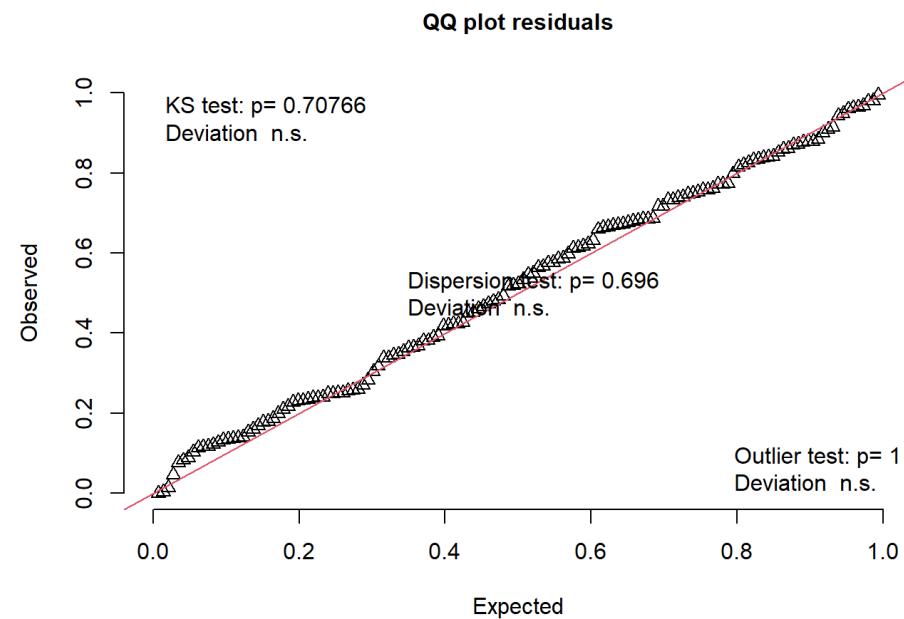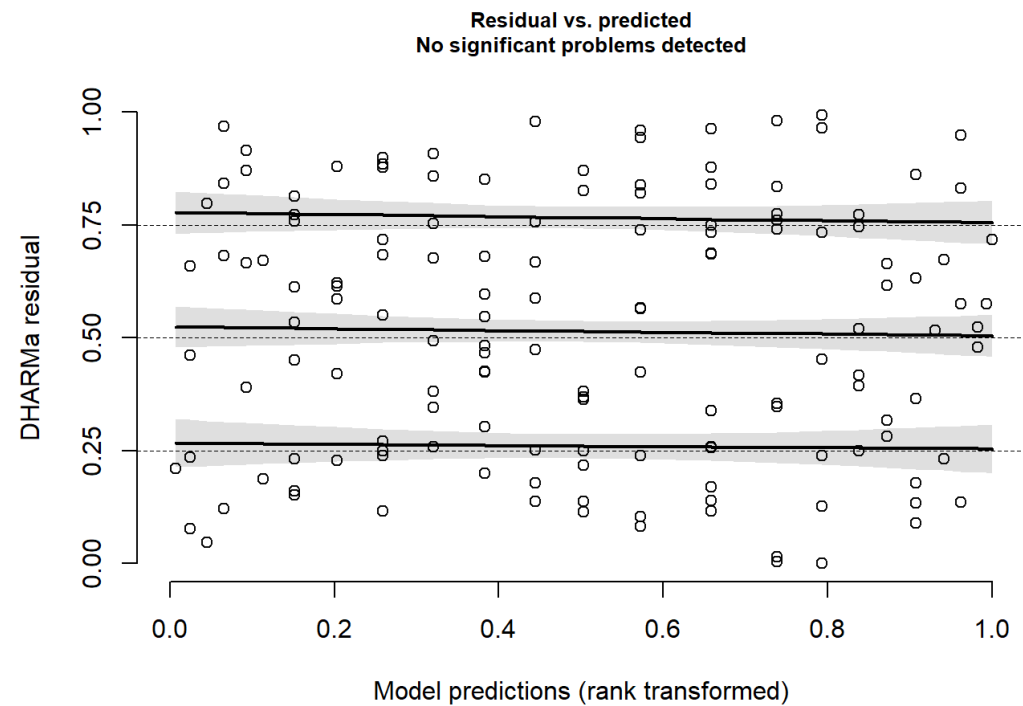# What would happen if we tried running a standard LM, rather than a binomial GLM?

```
sparrow.lm = lm(survival ~ tarsus, data = sparrow.data)
DHARMa::plotResiduals(sparrow.lm)
```

LINEAR MODEL

Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant

DHARMa residual

Model predictions (rank transformed)

QQ plot residuals

KS test: p= 0.00044
Deviation significant

Dispersion test: p= 0.944
Deviation n.s.

Outlier test: p= 0.6349
Deviation n.s.

Observed

Expected

BINOMIAL GLM

Residual vs. predicted
No significant problems detected

DHARMa residual

Model predictions (rank transformed)

QQ plot residuals

KS test: p= 0.70766
Deviation n.s.

Dispersion test: p= 0.696
Deviation n.s.

Outlier test: p= 1
Deviation n.s.

Observed

Expected

A few more notes on GLMs

```
glm(larvae ~ mass,
family = poisson(link = "log"), data = d)


glm(larvae ~ mass + length,
family = poisson(link = "log"), data = d)
```

Interaction terms * and :

```
glm(larvae ~ mass * length,
family = poisson(link = "log"), data = d)


glm(larvae ~ mass + length + mass:length,
family = poisson(link = "log"), data = d)
```