03\_model-analysis\_infection-endpoint\_load

Bryony Allen

09/05/2019

Table of Contents

[Part 1: Bd: import data 1](#_Toc8741993)

[Part 2: Bd: visualise endpoint infection load data 2](#_Toc8741994)

[Part 3: Bd: apply models to endpoint infection load data 4](#_Toc8741995)

[Part 3a: Bd: model selection 5](#_Toc8741996)

[Part 3b: Bd: model comparison 9](#_Toc8741997)

[Part 3c: Bd: model fit 11](#_Toc8741998)

[Part 3d: Bd: model plotting 12](#_Toc8741999)

[Part 3e: Bd: model checks 14](#_Toc8742000)

[Part 4: Bd: post-hoc tests 15](#_Toc8742001)

[Part 5: Rv: visualise endpoint infection load data 17](#_Toc8742002)

[Part 6: Rv: apply models to endpoint infection load data 18](#_Toc8742003)

[Part 6a: Rv: model selection 19](#_Toc8742004)

[Part 6b: Rv: model comparison 23](#_Toc8742005)

[Part 6c: Rv: model fit 25](#_Toc8742006)

[Part 6d: Rv: model plotting 26](#_Toc8742007)

[Part 6e: Rv: model checks 27](#_Toc8742008)

[Part 7: Rv: post-hoc tests 29](#_Toc8742009)

# 

# Part 1: Bd: import data

**N.B.** At the moment this code chuck reads in a .csv file with endpoint infection data, originally compiled in excel and then cleaned in “02\_tidy\_data-qpcr”script. In the future you will read in the .csv file of merged qPCR outputs (created, checked and cleaned in “02\_tidy\_data-qpcr”script) and experiment metadata (checked and cleaned in “02\_tidy\_data-metadata”script).

## Observations: 321  
## Variables: 26  
## $ ID <fct> A1.1, A1.2, A1.3, A1.4, A1.5, A1.6, A1...  
## $ Species <fct> Bb, Bb, Bb, Bb, Bb, Bb, Bb, Bb, Bb, Bb...  
## $ ExperimentNo <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ Scenario <fct> Coinfection, Coinfection, Coinfection,...  
## $ Treatment <fct> Rv-Bd, Rv-Bd, Rv-Bd, Rv-Bd, Rv-Bd, Rv-...  
## $ Exposure.1 <fct> rv, rv, rv, rv, rv, rv, rv, rv, rv, rv...  
## $ Exposure.2 <fct> bd, bd, bd, bd, bd, bd, bd, bd, bd, bd...  
## $ endpoint.date <fct> 06/06/2018, 26/05/2018, 06/06/2018, 27...  
## $ endpoint.code <fct> EU, MORT, EU, MORT, EU, MORT, EU, MORT...  
## $ Bd.endpoint.status <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ Bd.endpoint.CTmean <dbl> 30.94989, 37.30877, 32.19301, 36.65141...  
## $ Bd.endpoint.Qmean <dbl> 23.6021347, 0.4950310, 11.0839987, 0.7...  
## $ Rv.MCPendpoint.status <int> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0,...  
## $ Rv.MCPendpoint.CTmean <dbl> 0.00000, 24.75441, 0.00000, 22.90014, ...  
## $ Rv.MCPendpoint.Qmean <dbl> 0.000000e+00, 2.462007e+05, 0.000000e+...  
## $ Rv.EBF3Nendpoint.status <int> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0,...  
## $ Rv.EBF3Nendpoint.CTmean <dbl> NA, 33.46544, NA, 31.07020, NA, 34.558...  
## $ Rv.EBF3Nendpoint.Qmean <dbl> NA, 357.7631, NA, 1496.9165, NA, 182.4...  
## $ EMA.YN <int> 1, 0, 1, 0, 1, 1, 1, NA, NA, NA, NA, N...  
## $ EMA.date <fct> , NA, , NA, , , , NA, NA, NA, NA, NA, ...  
## $ EMA.GE.EMA <dbl> 0.004041085, NA, 0.479822159, NA, 0.00...  
## $ EMA.GE.WS <dbl> 0.026650012, NA, 0.005265172, NA, 0.00...  
## $ Bd.endpoint.GE <dbl> 236.021347, 4.950310, 110.839987, 7.37...  
## $ viable.GE <dbl> 0.004041085, NA, 0.479822159, NA, 0.00...  
## $ dead.GE <dbl> 0.022608927, NA, -0.474556987, NA, -0....  
## $ Rv.endpoint.load <dbl> NA, 1.376334e+03, NA, 9.135792e+02, NA...

# Part 2: Bd: visualise endpoint infection load data

Here I only include Bd infection loads over the 0.1 GE threshold.

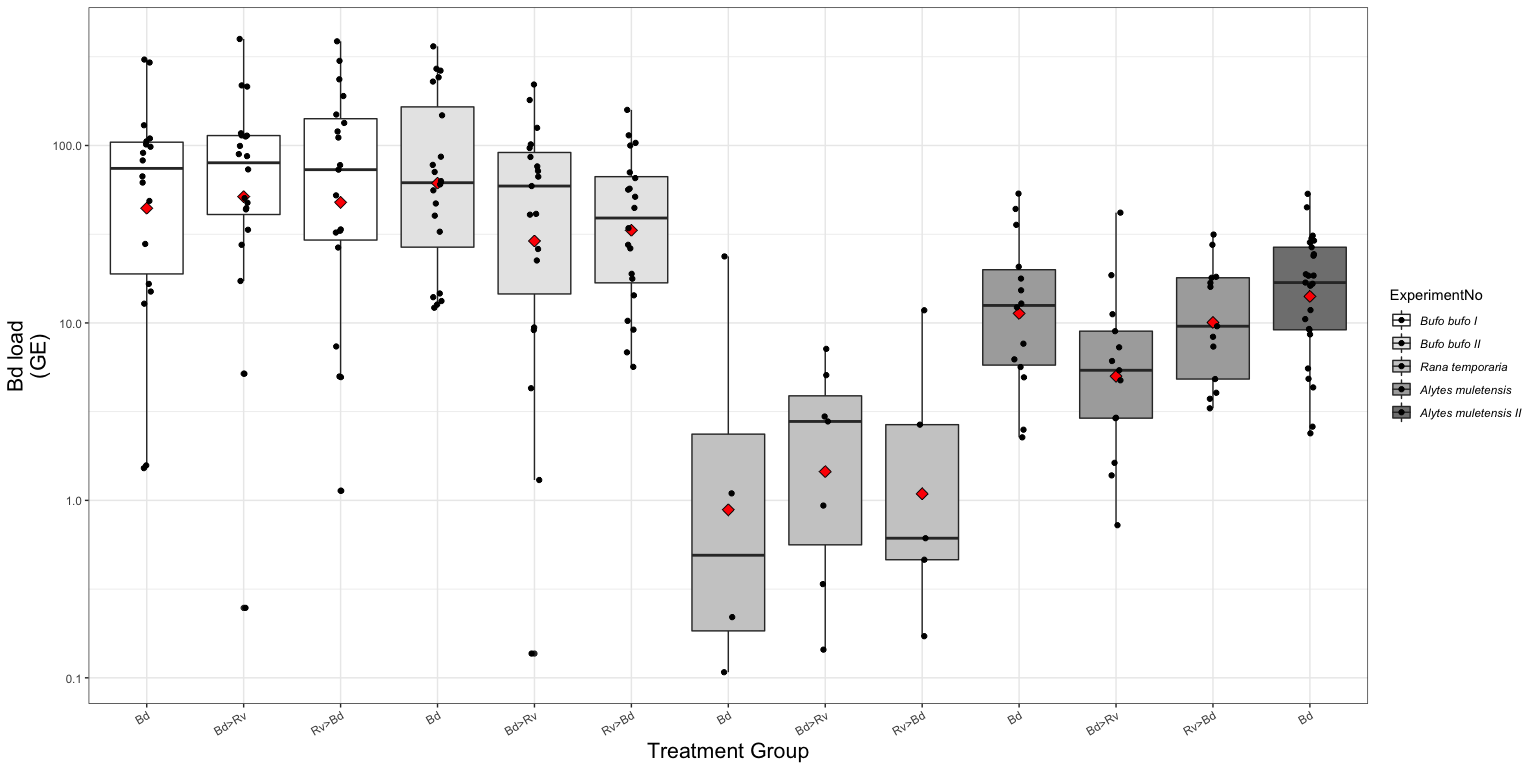


Fig.1. Boxplot of the endpoint Bd load for each treatment group across the three host species. Bd load is quantified as genomic equivalents (GE) where 1 GE represents 1 Bd zoospore. Where the black dots represent each sample and the red diamond the mean for that group.

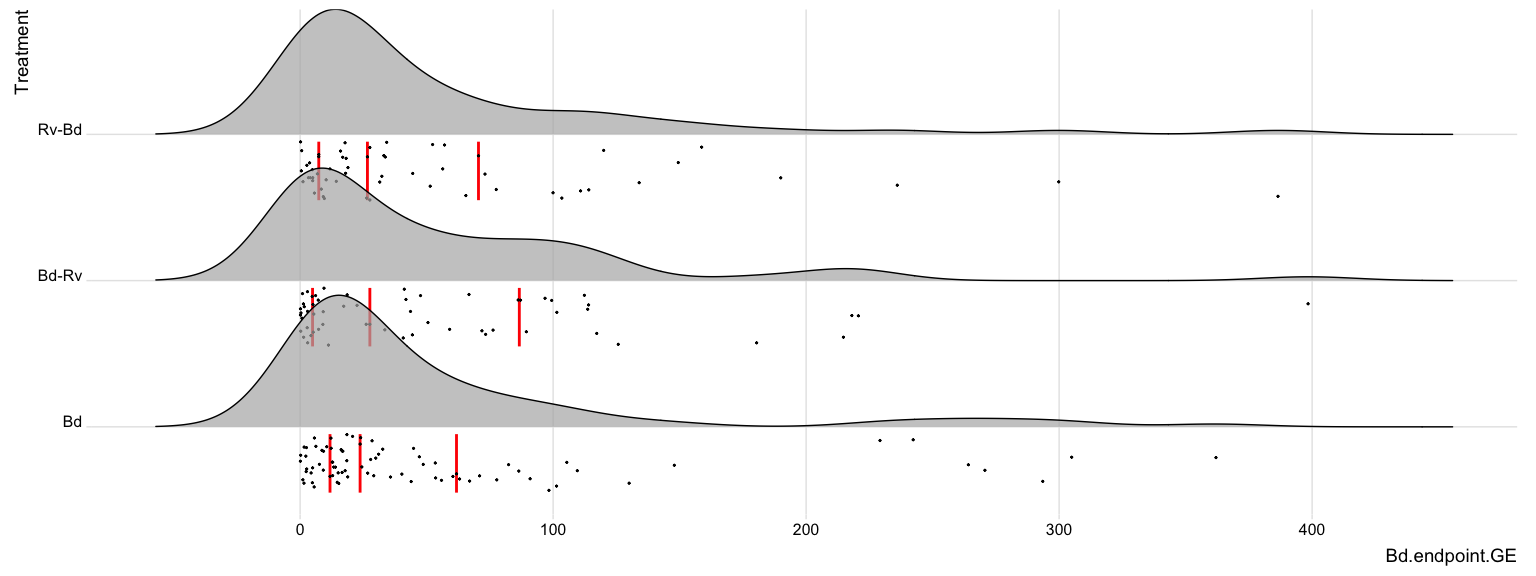


Fig. 2a. The distribution of Bd load expressed as GE from endpoint tissue samples.

… and when Bd load is logged

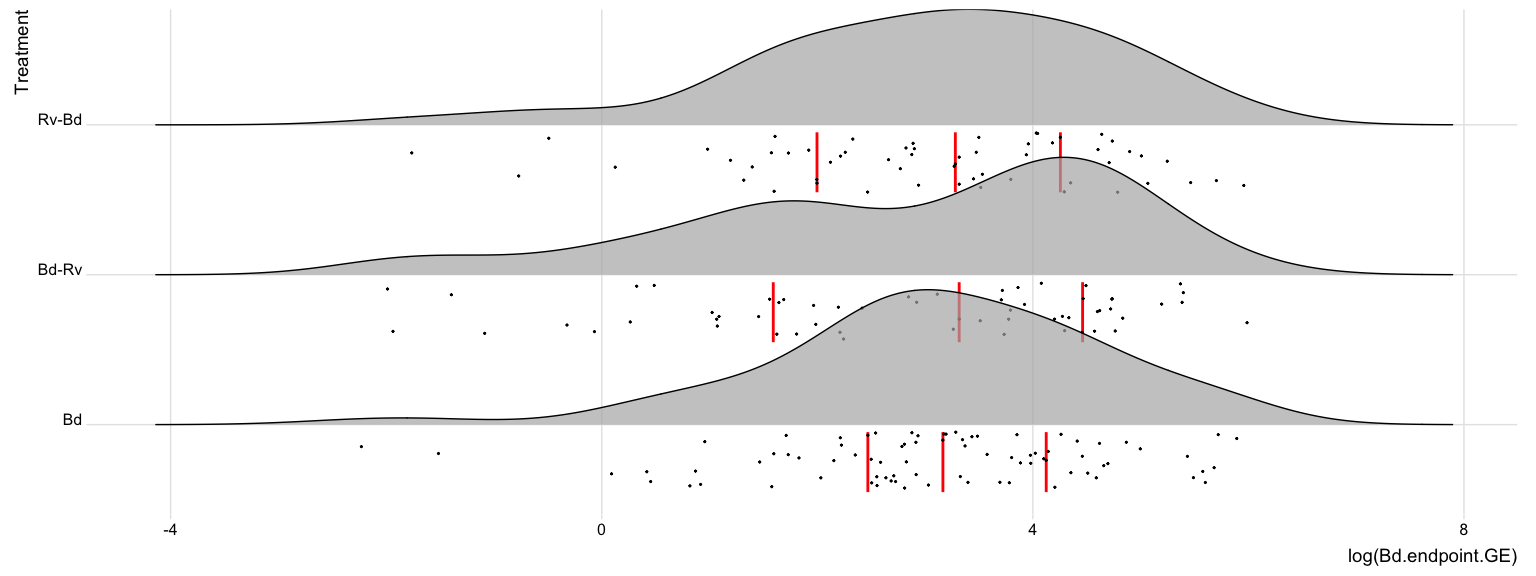


Fig. 2b. The logged distribution of Bd load expressed as genomic equivalents (GE), where one GE is equivalent to a single zoospore, from endpoint tissue samples.

# Part 3: Bd: apply models to endpoint infection load data

GLM’s where  
 response variable = endpoint load [continuous; ]  
 explanatory variable(s) = treatment [categorical, levels = 4] & species [categorical, levels = 5]

**N.B.** I use ExperimentNo as a proxy for species where

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment No. | Species | total Bd zsp’s | min. temp. (oC) | max. temp. (oC) |
| 1 | *Bufo bufo* I | 3,675,000 | 16.6 | 23.5 |
| 2 | *Bufo bufo* II | 1,443,750 | 16.7 | 27.6 |
| 3 | *Rana temporaria* | 2,336,250 | 16.7 | 27.6 |
| 4 | *Alytes muletensis* I | 472,500 | 15 | 16.6 |
| 5 | *Alytes muletensis* II | 294,759 | 15 | 16.6 |

… as this also accounts for Bd dose and room temperature variation between experiments.

Here I create a dataframe with only individuals that am infected with Bd, at a level over the detection threshold of 0.1GE. Note that *Alytes muletensis* II have been removed as they only have one treatment group (Bd only).

Bd.load <- data.endpoint %>%  
 filter(!ExperimentNo=='5') %>% # removal of Alytes babies  
 filter((Bd.endpoint.status=='1' & Bd.endpoint.GE > 0.1)) %>%   
 dplyr::select(ID, Species, ExperimentNo, Scenario, Treatment, Bd.endpoint.status, Bd.endpoint.GE)   
  
droplevels(Bd.load)

## Part 3a: Bd: model selection

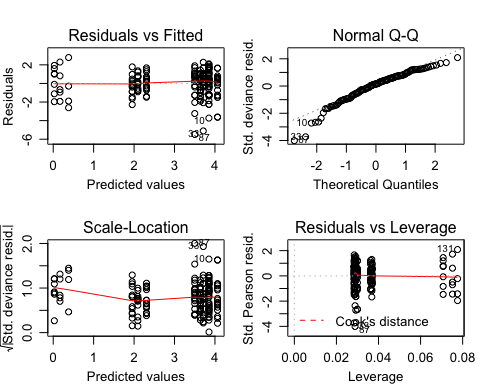
Choosing a error structure family. I decided to do this with a semi-maximal model to help the fitting as much as possible.

1. gaussian distribution with data log transformed

Bd.load2.1 <- glm(log(Bd.endpoint.GE) ~ Treatment + ExperimentNo, data=Bd.load, family= "gaussian")  
summary(Bd.load2.1) #AIC = 609.11 #Residual deviance: 320.39 on 166 degrees of freedom

##   
## Call:  
## glm(formula = log(Bd.endpoint.GE) ~ Treatment + ExperimentNo,   
## family = "gaussian", data = Bd.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.4879 -0.7431 0.2091 0.9583 2.7853   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.0639 0.2406 16.894 < 2e-16 \*\*\*  
## TreatmentBd-Rv -0.3526 0.2599 -1.357 0.177   
## TreatmentRv-Bd -0.2133 0.2615 -0.816 0.416   
## ExperimentNo2 -0.2110 0.2581 -0.817 0.415   
## ExperimentNo3 -3.6835 0.3936 -9.359 < 2e-16 \*\*\*  
## ExperimentNo4 -1.7573 0.2867 -6.130 6.19e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 1.93006)  
##   
## Null deviance: 552.77 on 171 degrees of freedom  
## Residual deviance: 320.39 on 166 degrees of freedom  
## AIC: 609.11  
##   
## Number of Fisher Scoring iterations: 2

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Bd.load2.1)



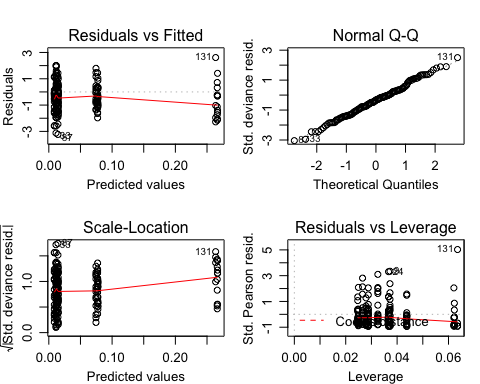
1. Gamma distribution

> this error structure is suitable for continous, positive, right skewed data where variance is near-constant on the log-scale

# Gamma distribution  
Bd.load2.3 <- glm(Bd.endpoint.GE ~ Treatment + ExperimentNo, data=Bd.load, family=Gamma())  
summary(Bd.load2.3) #AIC = 1633.2 #Residual deviance: 225.26 on 166 degrees of freedom

##   
## Call:  
## glm(formula = Bd.endpoint.GE ~ Treatment + ExperimentNo, family = Gamma(),   
## data = Bd.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2336 -1.1169 -0.3656 0.2824 2.6140   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.008727 0.001969 4.432 1.69e-05 \*\*\*  
## TreatmentBd-Rv 0.002579 0.002777 0.929 0.35441   
## TreatmentRv-Bd 0.003009 0.002846 1.057 0.29189   
## ExperimentNo2 0.003125 0.002405 1.299 0.19569   
## ExperimentNo3 0.254890 0.071672 3.556 0.00049 \*\*\*  
## ExperimentNo4 0.065786 0.013100 5.022 1.31e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 1.163935)  
##   
## Null deviance: 361.90 on 171 degrees of freedom  
## Residual deviance: 225.26 on 166 degrees of freedom  
## AIC: 1633.2  
##   
## Number of Fisher Scoring iterations: 7

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Bd.load2.3)



The diagnostic plots are fine

* residual vs. fitted do show patterning meaning the variance is non-consistent
* QQ plot is really good
* scale-location the scatter of the residuals decreases with the fitted values
* residual deviance is better than log(Bd load) with gaussian
* AIC is not great

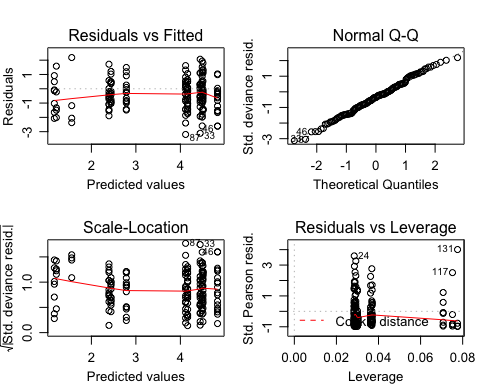
1. Gamma distribution with log-link

> apparent it is quite common to fit a log-link with a Gamma error structure … so that is left-skewed?

Bd.load2.4 <- glm(Bd.endpoint.GE ~ Treatment + ExperimentNo, data=Bd.load, family="Gamma"(link='log'))  
summary(Bd.load2.4) #AIC = 1629.9 #Residual deviance: 221.67 on 166 degrees of freedom

##   
## Call:  
## glm(formula = Bd.endpoint.GE ~ Treatment + ExperimentNo, family = Gamma(link = "log"),   
## data = Bd.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1959 -1.0667 -0.3603 0.3185 2.1843   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.8351 0.1797 26.906 <2e-16 \*\*\*  
## TreatmentBd-Rv -0.3904 0.1941 -2.011 0.0460 \*   
## TreatmentRv-Bd -0.3448 0.1954 -1.765 0.0794 .   
## ExperimentNo2 -0.3277 0.1928 -1.700 0.0911 .   
## ExperimentNo3 -3.2777 0.2940 -11.148 <2e-16 \*\*\*  
## ExperimentNo4 -2.0499 0.2141 -9.572 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 1.077049)  
##   
## Null deviance: 361.90 on 171 degrees of freedom  
## Residual deviance: 221.67 on 166 degrees of freedom  
## AIC: 1629.9  
##   
## Number of Fisher Scoring iterations: 7

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Bd.load2.4)



**Conclusion** Gamma with log-link as “the most appropriate link function is one which produces minimum residual deviance” and the diagnostic plots are better looking than the gaussian model.

## Part 3b: Bd: model comparison

The four models (all using Gamma with log-link structure)

Bd.load1 <- glm(Bd.endpoint.GE ~ Treatment \* ExperimentNo, data=Bd.load, family="Gamma"(link='log'))  
  
Bd.load2 <- glm(Bd.endpoint.GE ~ Treatment + ExperimentNo, data=Bd.load, family="Gamma"(link='log'))  
  
Bd.load3 <- glm(Bd.endpoint.GE ~ Treatment, data=Bd.load, family="Gamma"(link='log'))  
  
Bd.load4 <- glm(Bd.endpoint.GE ~ ExperimentNo, data=Bd.load, family="Gamma"(link='log'))  
  
Bd.load.N <- glm(Bd.endpoint.GE ~ 1, data=Bd.load, family="Gamma"(link='log'))

anova(Bd.load1, Bd.load2, test="Chisq") # start by comparing the interaction terms

## Analysis of Deviance Table  
##   
## Model 1: Bd.endpoint.GE ~ Treatment \* ExperimentNo  
## Model 2: Bd.endpoint.GE ~ Treatment + ExperimentNo  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 160 215.58   
## 2 166 221.67 -6 -6.0909 0.4362

anova(Bd.load2, Bd.load3, test="Chisq") # compares Trt and Species to just Trt

## Analysis of Deviance Table  
##   
## Model 1: Bd.endpoint.GE ~ Treatment + ExperimentNo  
## Model 2: Bd.endpoint.GE ~ Treatment  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 166 221.67   
## 2 169 359.64 -3 -137.97 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(Bd.load2, Bd.load4, test="Chisq") # compares Trt and Species to just Species

## Analysis of Deviance Table  
##   
## Model 1: Bd.endpoint.GE ~ Treatment + ExperimentNo  
## Model 2: Bd.endpoint.GE ~ ExperimentNo  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 166 221.67   
## 2 168 226.88 -2 -5.2094 0.08907 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(Bd.load4, Bd.load.N, test="Chisq") # compares Species model to null model

## Analysis of Deviance Table  
##   
## Model 1: Bd.endpoint.GE ~ ExperimentNo  
## Model 2: Bd.endpoint.GE ~ 1  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 168 226.88   
## 2 171 361.90 -3 -135.02 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ANOVA 1: suggests we should reject the more complex model with interaction terms (Treatment \* ExperimentNo) in favour for the model with just the terms (pvalue = 0.4362)

ANOVA 2: suggests we should favour the more complex model (Treatment + ExperimentNo) over the model with Treatment only (pvalue = < .001) so ExperimentNo leads to significantly improved fit

ANOVA 3: suggests we should reject the more complex model (Treatment + ExperimentNo) and choose the model with just ExperimentNo. (pvalue = 0.08907)

ANOVA 4: suggests we should favour the more complex model (ExperimentNo.) over the null model (pvalue = < .001) so ExperimentNo leads to significantly improved fit

**Conclusion**: we should choose the model with just ExperimentNo (aka species)

## Part 3c: Bd: model fit

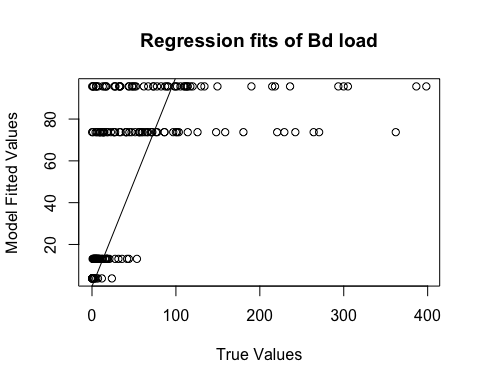


Fig. 6. Plot of relationship between the fitted values from the model and the true data values.

**To Do** Figure this one out. Values should fall around the line but they seem to be clumped by ExperimentNo. tried to do this with the null model but that just grouped everything

## Part 3d: Bd: model plotting

# Bd.load4 <- glm(Bd.endpoint.GE ~ ExperimentNo, data=Bd.load, family="Gamma"(link='log'))  
  
# create a dataframe of "new" data   
newdat <- expand.grid(ExperimentNo=c("1", "2", "3", "4"),Treatment=c("Bd", "Bd-Rv", "Rv-Bd"))  
  
# predict the value/result of the new data using the glm  
newdat <-cbind(newdat, predict(object = Bd.load4, # the model   
 newdata=newdat, se=TRUE, type="response", print.matrix=T)) # dataframe of new data   
newdat

## ExperimentNo Treatment fit se.fit residual.scale  
## 1 1 Bd 95.624850 13.753365 1.085864  
## 2 2 Bd 73.727916 10.422732 1.085864  
## 3 3 Bd 3.763935 1.021781 1.085864  
## 4 4 Bd 13.105349 2.250061 1.085864  
## 5 1 Bd-Rv 95.624850 13.753365 1.085864  
## 6 2 Bd-Rv 73.727916 10.422732 1.085864  
## 7 3 Bd-Rv 3.763935 1.021781 1.085864  
## 8 4 Bd-Rv 13.105349 2.250061 1.085864  
## 9 1 Rv-Bd 95.624850 13.753365 1.085864  
## 10 2 Rv-Bd 73.727916 10.422732 1.085864  
## 11 3 Rv-Bd 3.763935 1.021781 1.085864  
## 12 4 Rv-Bd 13.105349 2.250061 1.085864

expl.var <- c(1:3) # chose the range for the x-axis (Experiment No.)  
exp.labs <- c("1" = "Bufo bufo I", "2" = "Bufo bufo II", "3" = "Rana temporaria", "4" = "Alytes muletensis I")  
  
newdat1<- subset(newdat, ExperimentNo== "1") # need to subset the data so you can plot each seperatly   
newdat2<- subset(newdat, ExperimentNo=="2")  
newdat3<- subset(newdat, ExperimentNo=="3")  
newdat4<- subset(newdat, ExperimentNo=="4")  
newdat5<- subset(newdat, ExperimentNo=="5")

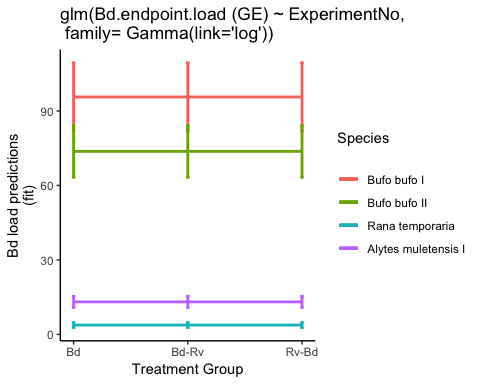
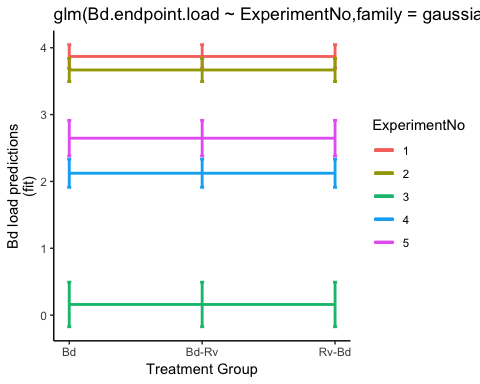


Fig. 7. Probability of Bd infection load (GE) predicted by model

Here is the old plot with *Alytes* “babies” included and a gaussian error family.



The Bd load predictions are interesting…

The two *Bufo bufo* experiments have the highest predicted loads, and reflect the Bd dosed (1 = 3,675,000 zsp’s; 2 = 1,443,750 zsp’s). As anticipated from the Bd status data and predictions on species susceptibility the *Rana temporaria* loads are low.

Oddly, the *Alytes muletensis* load prediction is lower than I had expected based off of the Bd status predictions. Similar numbers of *Alytes* and *Bufo* are infected with Bd but the loads differ. Two possibilities here, species susceptibility differences in pathogen burden or (more likely) a reflection of Bd dose (472,500 zsp’s compared to millions of zsp’s).

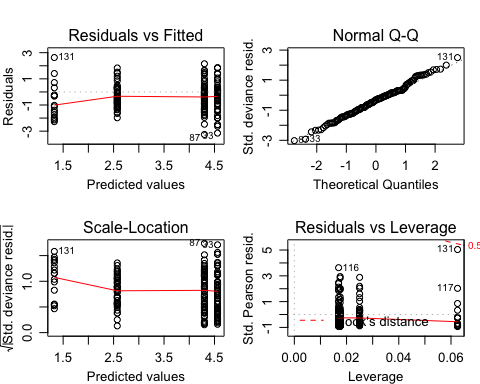
**QUESTION** all Bd doses are consisted high but are we seeing dose-dependent Bd loads within this high dose class?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment No. | Species | total Bd zsp’s | min. temp. (oC) | max. temp. (oC) |
| 1 | *Bufo bufo* I | 3,675,000 | 16.6 | 23.5 |
| 2 | *Bufo bufo* II | 1,443,750 | 16.7 | 27.6 |
| 3 | *Rana temporaria* | 2,336,250 | 16.7 | 27.6 |
| 4 | *Alytes muletensis* I | 472,500 | 15 | 16.6 |
| 5 | *Alytes muletensis* II | 294,759 | 15 | 16.6 |

## Part 3e: Bd: model checks

Here I check the chosen model Bd.load4 <- glm(Bd.endpoint.GE ~ ExperimentNo, data=Bd.load, family="Gamma"(link='log'))

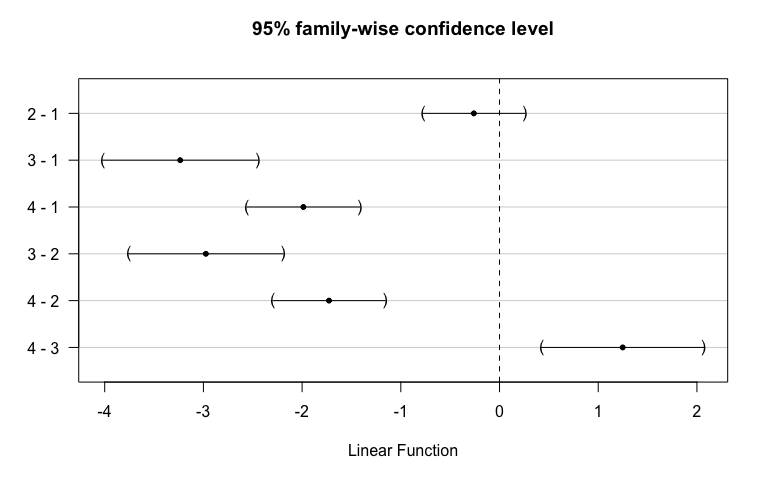
##   
## Call:  
## glm(formula = Bd.endpoint.GE ~ ExperimentNo, family = Gamma(link = "log"),   
## data = Bd.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2526 -1.1251 -0.3381 0.2420 2.6297   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5604 0.1438 31.708 < 2e-16 \*\*\*  
## ExperimentNo2 -0.2601 0.2017 -1.289 0.199   
## ExperimentNo3 -3.2350 0.3072 -10.530 < 2e-16 \*\*\*  
## ExperimentNo4 -1.9874 0.2240 -8.873 1.02e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 1.179102)  
##   
## Null deviance: 361.90 on 171 degrees of freedom  
## Residual deviance: 226.88 on 168 degrees of freedom  
## AIC: 1630.6  
##   
## Number of Fisher Scoring iterations: 7



# Part 4: Bd: post-hoc tests

Tukry’s honest significant difference

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Tukey Contrasts  
##   
##   
## Fit: glm(formula = Bd.endpoint.GE ~ ExperimentNo, family = Gamma(link = "log"),   
## data = Bd.load)  
##   
## Linear Hypotheses:  
## Estimate Std. Error z value Pr(>|z|)   
## 2 - 1 == 0 -0.2601 0.2017 -1.289 0.56225   
## 3 - 1 == 0 -3.2350 0.3072 -10.530 < 1e-04 \*\*\*  
## 4 - 1 == 0 -1.9874 0.2240 -8.873 < 1e-04 \*\*\*  
## 3 - 2 == 0 -2.9749 0.3061 -9.720 < 1e-04 \*\*\*  
## 4 - 2 == 0 -1.7274 0.2224 -7.767 < 1e-04 \*\*\*  
## 4 - 3 == 0 1.2476 0.3212 3.884 0.00048 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)



Where “1” = “Bufo bufo I”, “2” = “Bufo bufo II”, “3” = “Rana temporaria”, “4” = “Alytes muletensis”

**Conclusions** Tukey summary suggests everything is significant apart from the two *Bufo* experiments.

The plots suggest:

* 2-1 : no significant difference between *Bufo bufo* experiments
* 3-2 & 3-1 : significant difference between *Rana* and both *Bufo bufo* experiments (to an equal extent)
* 4-2 & 4-1: significant difference between *Alytes* and both *Bufo bufo* experiments (to an equal extent) <<< see above for my thoughts on this
* 4-3: significant (+ve) difference between *Alytes* and *Rana*

Also did a pairwise.t.test() comparing the factor levels of ExperimentNo …. not entirely sure wht

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Bd.endpoint.GE and ExperimentNo   
##   
## 1 2 3   
## 2 0.20527 - -   
## 3 5.8e-05 0.00209 -   
## 4 6.0e-07 0.00024 0.66080  
##   
## P value adjustment method: holm

# Part 5: Rv: visualise endpoint infection load data

For the plots below I only plot Rv infection loads over 0 and filter out one extremely high Rv viral load (greater than 2000).

**To Do** At the moment there are 3 samples missing EBF3N qPCR data so as a temporary fix I have averaged the EBF3N score for all the samples and used that to calculate a rough viral load for these 3 samples

**N.B.** previous thoughts on filtering out samples with negative infection status vs. including and using log(n+1). AF response “usually settle for running one binomial ‘status’ (presence/absence of infection) analysis, and a separate ‘intensity’ (load among infecteds only) analysis. Zero-inflated/Hurdle models effectively do this in a single model, but I typically think splitting into 2 separate analyses is clearer and easier to explain/interpret.”

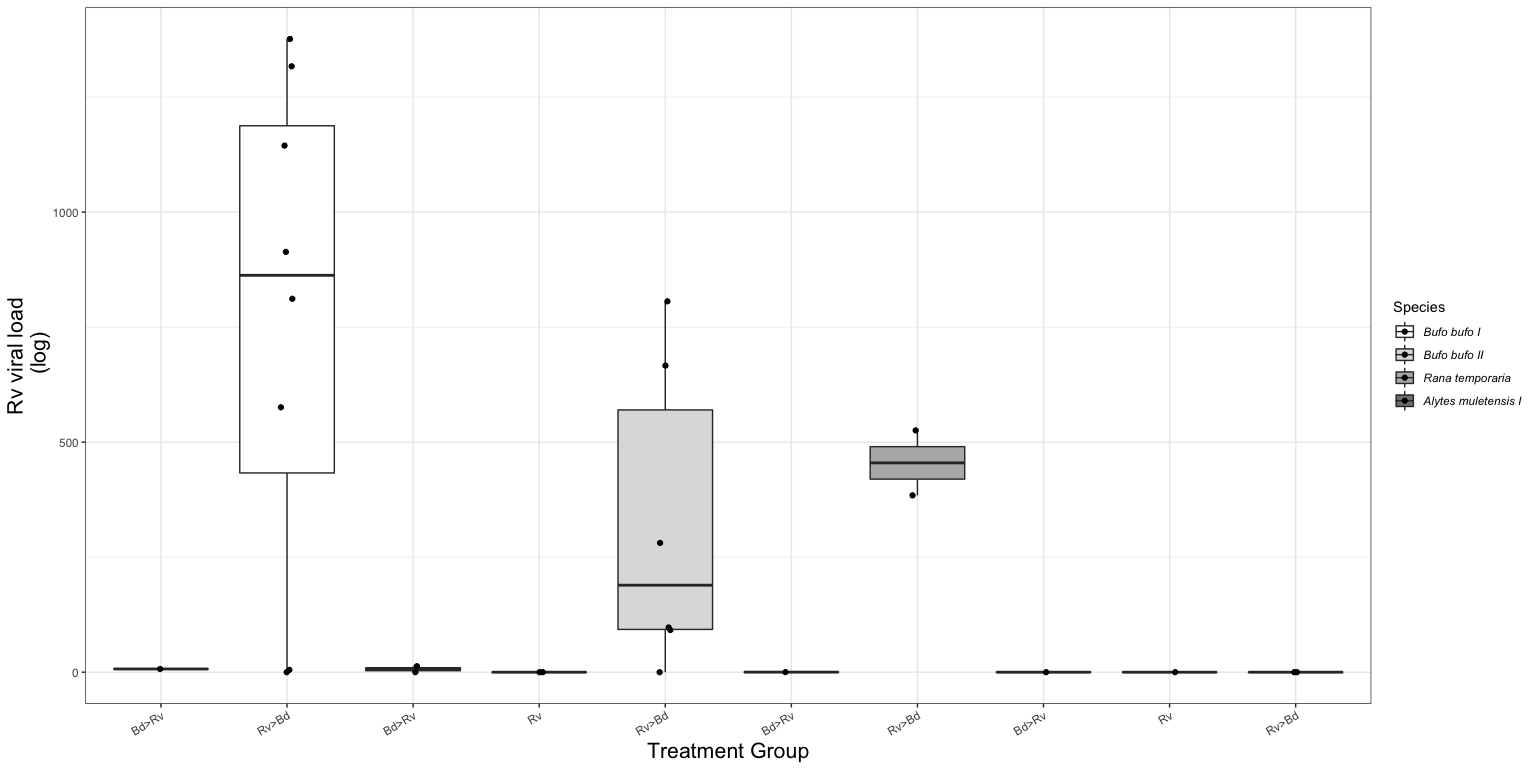


Fig.3. Boxplot of the endpoint Ranaviral load (excluding uninfected individuals) for each treatment group across the three host species. Viral load has been normalised using Leung et al.’s (2017) method. The black dots represent each sample.

The distribution of Rv viral load (excluding uninfected individuals)

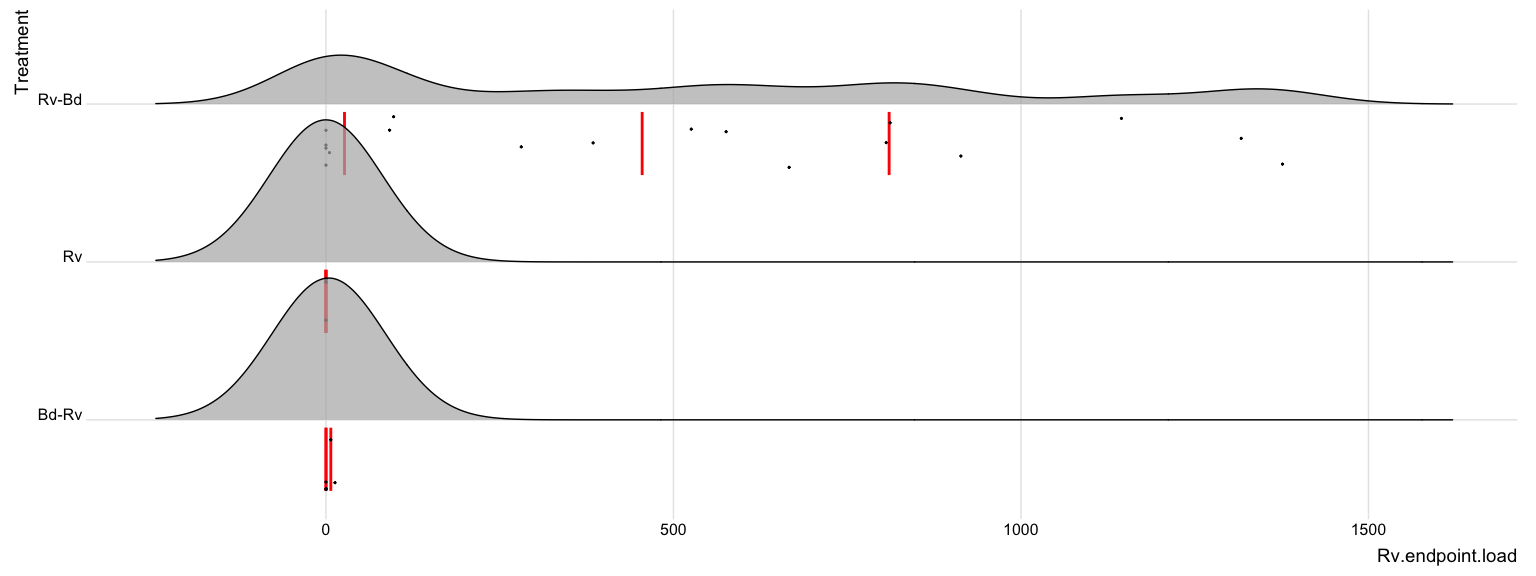


Fig. 4a. The distribution of Rv load from endpoint tissue samples.

…and when Rv viral load is logged because the distribution of Rv load is skewed. Again excluding uninfected individuals.

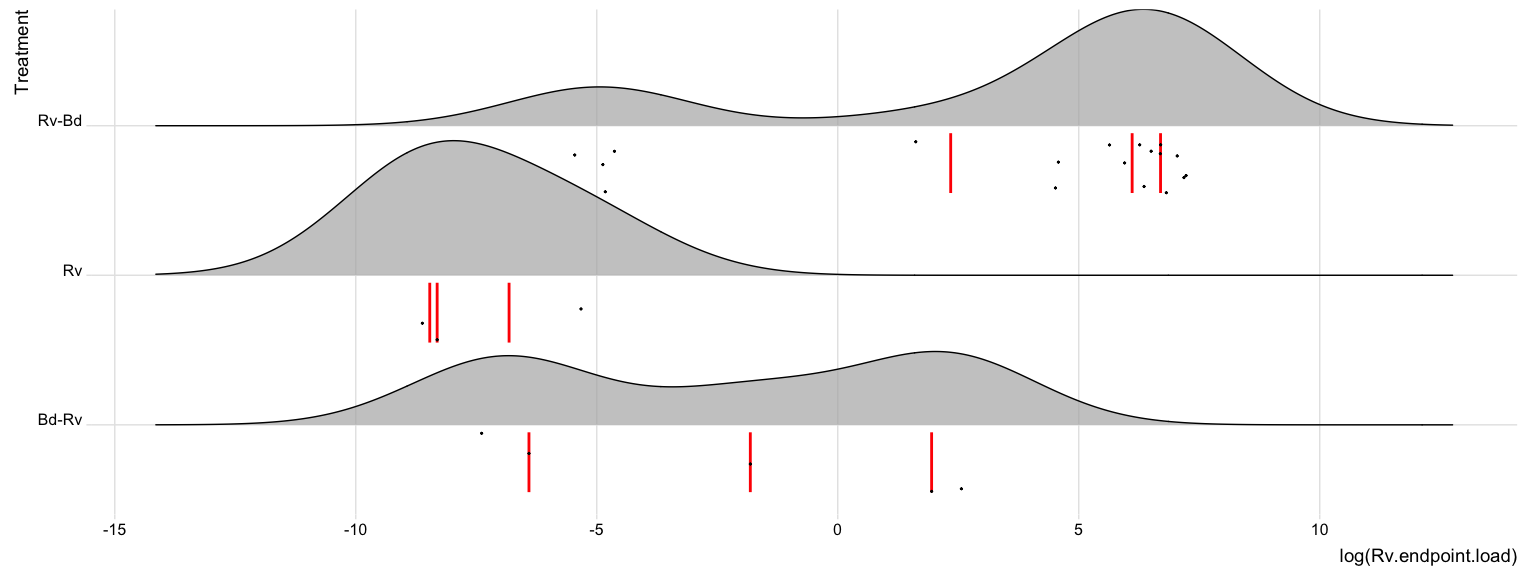


Fig. 4b. The logged distribution of Rv load from endpoint tissue samples.

Just eyeballing this the Rv-Bd group has the highest infection loads compared to the other treatment groups. Both the Rv only and the Bd-Rv group were exposed to Rv at the same time point at the end of the experiments exposure window so 9 days before these sample were taken.

?? **QUESTION** is the sequence/timing of the Rv exposure allowing enough time for Rv loads to build or is it the nauture of that coinfection combination??

# Part 6: Rv: apply models to endpoint infection load data

Here I create a dataframe with only individuals that am infected with Rv

Rv.load <- data.endpoint %>%  
 filter((Rv.endpoint.load > 0.1)) %>%   
 dplyr::select(ID, Species, ExperimentNo, Scenario, Treatment, Rv.MCPendpoint.status, Rv.EBF3Nendpoint.status, Rv.endpoint.load)   
  
droplevels(Rv.load)

**To Do** check whether there is a threshold for ranavirus detection???

## Part 6a: Rv: model selection

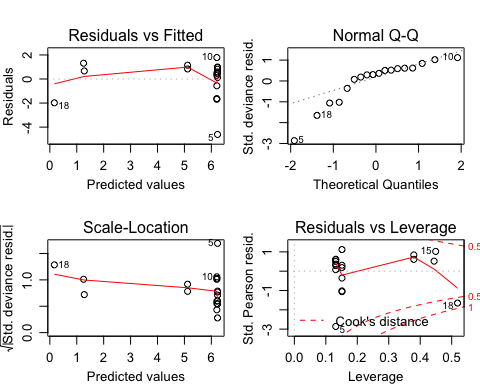
Choosing a error structure family. Again, I decided to do this with a semi-maximal model to help the fitting as much as possible.

1. gaussian distribution with data log transformed

Rv.load2.1 <- glm(log(Rv.endpoint.load) ~ Treatment + ExperimentNo, data=Rv.load, family= "gaussian")  
summary(Rv.load2.1) #AIC = 76.262 #Residual deviance: 41.834 on 14 degrees of freedom

##   
## Call:  
## glm(formula = log(Rv.endpoint.load) ~ Treatment + ExperimentNo,   
## family = "gaussian", data = Rv.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.6115 -0.3936 0.5372 0.9230 1.7846   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.27919 1.15206 1.110 0.285556   
## TreatmentRv-Bd 4.95117 1.11610 4.436 0.000564 \*\*\*  
## ExperimentNo2 -0.02557 0.89487 -0.029 0.977607   
## ExperimentNo3 -1.11255 1.19317 -0.932 0.366915   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 2.988172)  
##   
## Null deviance: 111.617 on 17 degrees of freedom  
## Residual deviance: 41.834 on 14 degrees of freedom  
## AIC: 76.262  
##   
## Number of Fisher Scoring iterations: 2

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Rv.load2.1)



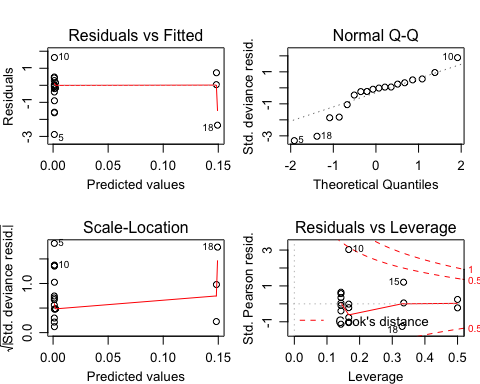
* residuals vs fitted: wedged
* QQ plot is not great, (not normally distributed?)
* residual vs. leverage reflects the data, low sample numbers means a few data points are influencing the model

1. Gamma distribution

# Gamma distribution  
Rv.load2.3 <- glm(Rv.endpoint.load ~ Treatment + ExperimentNo, data=Rv.load, family=Gamma())  
summary(Rv.load2.3) #AIC = 257.35 #Residual deviance: 23.678 on 14 degrees of freedom

##   
## Call:  
## glm(formula = Rv.endpoint.load ~ Treatment + ExperimentNo, family = Gamma(),   
## data = Rv.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.88592 -0.77833 -0.04491 0.24714 1.63001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.483e-01 8.110e-02 1.828 0.0889 .  
## TreatmentRv-Bd -1.471e-01 8.110e-02 -1.814 0.0912 .  
## ExperimentNo2 8.565e-05 6.236e-04 0.137 0.8927   
## ExperimentNo3 1.075e-03 1.534e-03 0.700 0.4951   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 0.893)  
##   
## Null deviance: 46.427 on 17 degrees of freedom  
## Residual deviance: 23.678 on 14 degrees of freedom  
## AIC: 257.35  
##   
## Number of Fisher Scoring iterations: 7

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Rv.load2.3)



Not great

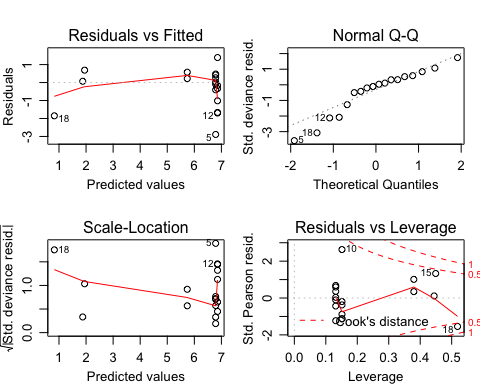
* residuals vs fitted: patterning
* QQ plot is not great, suggesting not normally distributed
* residual vs. leverage is better but still reflects the data, low sample numbers means a few data points are influencing the model

1. Gamma distribution with log-link

Rv.load2.4 <- glm(Rv.endpoint.load ~ Treatment + ExperimentNo, data=Rv.load, family="Gamma"(link='log'))  
summary(Rv.load2.4) #AIC = 255.8 #Residual deviance: 22.027 on 14 degrees of freedom

##   
## Call:  
## glm(formula = Rv.endpoint.load ~ Treatment + ExperimentNo, family = Gamma(link = "log"),   
## data = Rv.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.88952 -0.86174 -0.02881 0.38392 1.39261   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87732 0.57792 3.248 0.00583 \*\*   
## TreatmentRv-Bd 4.91046 0.55988 8.771 4.62e-07 \*\*\*  
## ExperimentNo2 0.06971 0.44890 0.155 0.87881   
## ExperimentNo3 -1.04891 0.59854 -1.752 0.10156   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 0.7519493)  
##   
## Null deviance: 46.427 on 17 degrees of freedom  
## Residual deviance: 22.027 on 14 degrees of freedom  
## AIC: 255.8  
##   
## Number of Fisher Scoring iterations: 9

par(mfrow=c(2,2), mar=c(3,3,3,1), mgp=c(2,0.8,0))  
plot(Rv.load2.4)



**Conclusion** Gamma with log-link has the best (QQ) plots and lowest residual deviance

**To Do** AF suggests splitting the continuous Rv load data into zero-low-high or similar categories? ref bimodal distributions in Fig 3b

## Part 6b: Rv: model comparison

The four models (all using log(Bd load) and gaussian error structure)

Rv.load1 <- glm(Rv.endpoint.load ~ Treatment \* ExperimentNo, data=Rv.load, family="Gamma"(link='log'))  
  
Rv.load2 <- glm(Rv.endpoint.load ~ Treatment + ExperimentNo, data=Rv.load, family="Gamma"(link='log'))  
  
Rv.load3 <- glm(Rv.endpoint.load ~ Treatment, data=Rv.load, family="Gamma"(link='log'))  
  
Rv.load4 <- glm(Rv.endpoint.load ~ ExperimentNo, data=Rv.load, family="Gamma"(link='log'))  
  
Rv.load.N <- glm(Rv.endpoint.load ~ 1, data=Rv.load, family="Gamma"(link='log'))

anova(Rv.load1, Rv.load2, test="Chisq") # start by comparing the interaction terms

## Analysis of Deviance Table  
##   
## Model 1: Rv.endpoint.load ~ Treatment \* ExperimentNo  
## Model 2: Rv.endpoint.load ~ Treatment + ExperimentNo  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 12 17.652   
## 2 14 22.027 -2 -4.3747 0.08616 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(Rv.load2, Rv.load3, test="Chisq") # compares Trt and Species to just Trt

## Analysis of Deviance Table  
##   
## Model 1: Rv.endpoint.load ~ Treatment + ExperimentNo  
## Model 2: Rv.endpoint.load ~ Treatment  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 14 22.027   
## 2 16 24.287 -2 -2.2602 0.2225

anova(Rv.load2, Rv.load4, test="Chisq") # compares Trt and Species to just Species

## Analysis of Deviance Table  
##   
## Model 1: Rv.endpoint.load ~ Treatment + ExperimentNo  
## Model 2: Rv.endpoint.load ~ ExperimentNo  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 14 22.027   
## 2 15 44.785 -1 -22.758 3.769e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(Rv.load3, Rv.load.N, test="Chisq") # compares Species model to null model

## Analysis of Deviance Table  
##   
## Model 1: Rv.endpoint.load ~ Treatment  
## Model 2: Rv.endpoint.load ~ 1  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 16 24.287   
## 2 17 46.427 -1 -22.14 3.869e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ANOVA 1: shows adding a interaction term between Treatment and ExperimentNo does not greatly imporve fit (p-value = 0.08616)

ANOVA 2: suggests we should reject the more complex model (Treatment + ExperimentNo) in favour of the model with Treatment only (pvalue = 0.2225)

ANOVA 3: suggests we should favour the more complex model (Treatment + ExperimentNo) over the model with ExperimentNo only (pvalue = < .001)

ANOVA 4: suggests we should accept the more complex model (Treatment) over the null model (pvalue = < .001) so ExperimentNo leads to significantly improved fit

**Conclusion**: we should choose the model with just Treatment (like Rv status)

## Part 6c: Rv: model fit

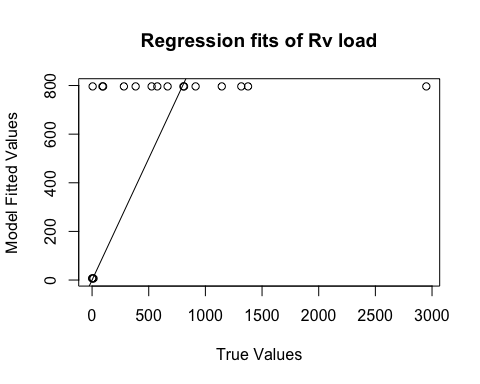


Fig. 6. Plot of relationship between the fitted values from the model and the true data values.

**To Do** split Rv load data into categories low-high or similar?

## Part 6d: Rv: model plotting

#Rv.load3 <- glm(Rv.endpoint.load ~ Treatment, data=Rv.load, family="Gamma"(link='log'))  
  
# create a dataframe of "new" data   
newdat <- expand.grid(ExperimentNo=c("1", "2", "3", "4", "5"),Treatment=c( "Bd-Rv", "Rv-Bd"))  
  
# predict the value/result of the new data using the glm  
Rv.newdat <-cbind(newdat, predict(object = Rv.load3, # the model   
 newdata=newdat, se=TRUE, type="response", print.matrix=T)) # dataframe of new data   
Rv.newdat

## ExperimentNo Treatment fit se.fit residual.scale  
## 1 1 Bd-Rv 6.727776 3.601161 0.9271107  
## 2 2 Bd-Rv 6.727776 3.601161 0.9271107  
## 3 3 Bd-Rv 6.727776 3.601161 0.9271107  
## 4 4 Bd-Rv 6.727776 3.601161 0.9271107  
## 5 5 Bd-Rv 6.727776 3.601161 0.9271107  
## 6 1 Rv-Bd 796.385476 190.637915 0.9271107  
## 7 2 Rv-Bd 796.385476 190.637915 0.9271107  
## 8 3 Rv-Bd 796.385476 190.637915 0.9271107  
## 9 4 Rv-Bd 796.385476 190.637915 0.9271107  
## 10 5 Rv-Bd 796.385476 190.637915 0.9271107

expl.var <- c(1:2) # chose the range for the x-axis (Treatment)  
exp.labs <- c("1" = "Bufo bufo I", "2" = "Bufo bufo II", "3" = "Rana temporaria", "4" = "Alytes muletensis I")  
  
newdat1<- subset(Rv.newdat, ExperimentNo== "1") # need to subset the data so you can plot each seperatly   
newdat2<- subset(Rv.newdat, ExperimentNo=="2")  
newdat3<- subset(Rv.newdat, ExperimentNo=="3")  
newdat4<- subset(Rv.newdat, ExperimentNo=="4")

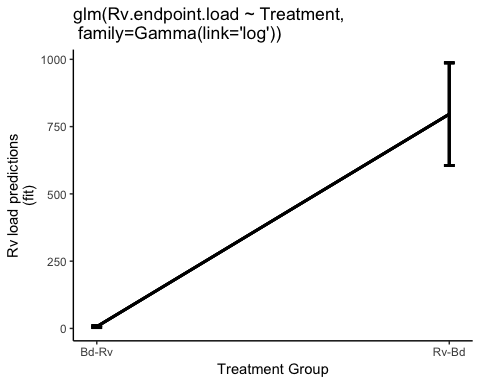


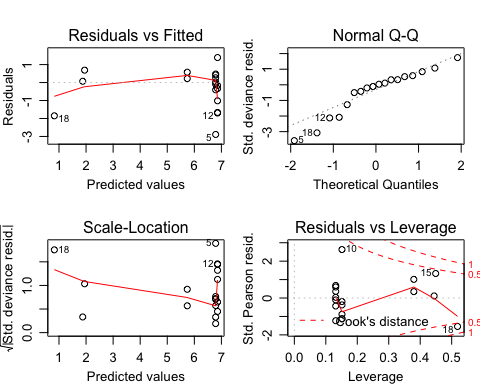
Fig. 7. Probability of Rv infection load predicted by model

## Part 6e: Rv: model checks

Here I check the two best models.

#### Maximal Model: Species + Treamtent (no interaction)

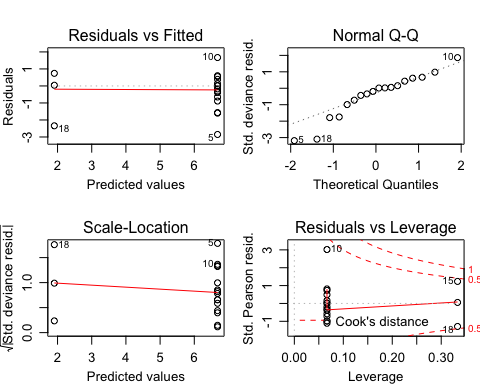
##   
## Call:  
## glm(formula = Rv.endpoint.load ~ Treatment + ExperimentNo, family = Gamma(link = "log"),   
## data = Rv.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.88952 -0.86174 -0.02881 0.38392 1.39261   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87732 0.57792 3.248 0.00583 \*\*   
## TreatmentRv-Bd 4.91046 0.55988 8.771 4.62e-07 \*\*\*  
## ExperimentNo2 0.06971 0.44890 0.155 0.87881   
## ExperimentNo3 -1.04891 0.59854 -1.752 0.10156   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 0.7519493)  
##   
## Null deviance: 46.427 on 17 degrees of freedom  
## Residual deviance: 22.027 on 14 degrees of freedom  
## AIC: 255.8  
##   
## Number of Fisher Scoring iterations: 9



#### Simplified Model: Treatment

Now fit a simpler model has an with only ExperimentNo.

##   
## Call:  
## glm(formula = Rv.endpoint.load ~ Treatment, family = Gamma(link = "log"),   
## data = Rv.load)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.85223 -0.82878 -0.08042 0.32466 1.66993   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.9062 0.5353 3.561 0.0026 \*\*   
## TreatmentRv-Bd 4.7738 0.5864 8.142 4.41e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Gamma family taken to be 0.8595342)  
##   
## Null deviance: 46.427 on 17 degrees of freedom  
## Residual deviance: 24.287 on 16 degrees of freedom  
## AIC: 253.89  
##   
## Number of Fisher Scoring iterations: 6



The AIC better for Treatment only model.

# Part 7: Rv: post-hoc tests

Tukey’s honest significant difference

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Tukey Contrasts  
##   
##   
## Fit: glm(formula = Rv.endpoint.load ~ Treatment, family = Gamma(link = "log"),   
## data = Rv.load)  
##   
## Linear Hypotheses:  
## Estimate Std. Error z value Pr(>|z|)   
## Rv-Bd - Bd-Rv == 0 4.7738 0.5864 8.142 4.44e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)

**Conclusions** Tukey summary suggests there is significant difference between the two coinfection treatment groups. The plot was pointless.