

Modelling zero-inflated (and overdispersed) count data

Francisco Rodríguez-Sánchez

<https://frodriguezsanchez.net>

How many eggs in nests?



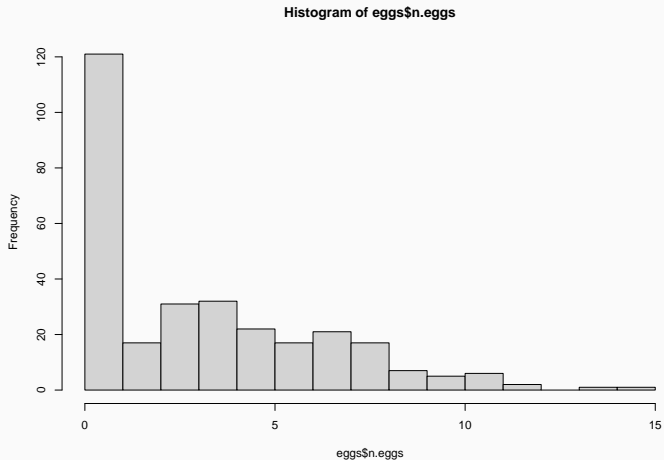
```
eggs <- read.csv("data/eggs.csv")
```

diameter	old	n.eggs
14	no	4
8	yes	0
7	yes	0

diameter: nest diameter (cm)

old: does nest look old/abandoned?

How many eggs in nests?



Many zeros does not mean you need a zero-inflated model!

Check model afterwards

How many eggs in nests?

- Nests may be occupied or not

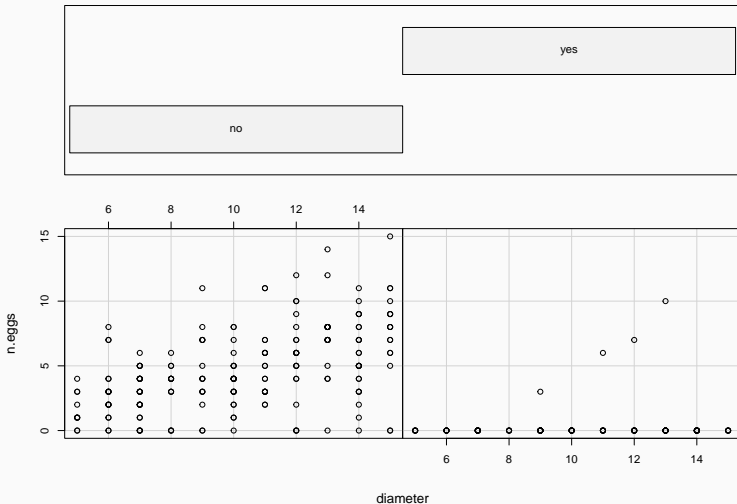
How many eggs in nests?

- Nests may be occupied or not
- Occupied nests may not have eggs (too soon, predation, etc)

Number of eggs ~ nest diameter * old appearance

```
coplot(n.eggs ~ diameter | old, data = eggs)
```

Given : old



```
eggs.poi <- glm(n.eggs ~ old * diameter,  
               data = eggs,  
               family = poisson)
```

Trying Poisson GLM

Call:

```
glm(formula = n.eggs ~ old * diameter, family = poisson, data = eggs)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.8905	-0.8784	-0.4514	0.3892	6.6795

Coefficients:

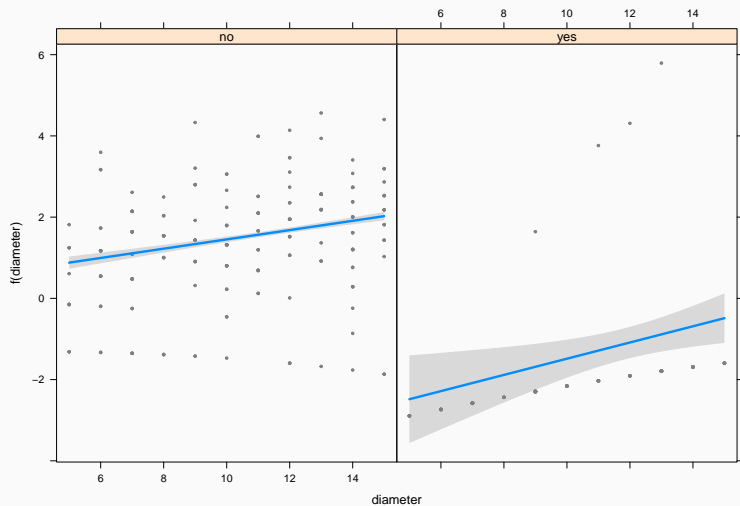
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.30773	0.12883	2.389	0.0169 *
oldyes	-3.78879	0.92230	-4.108	3.99e-05 ***
diameter	0.11441	0.01105	10.354	< 2e-16 ***
oldyes:diameter	0.08513	0.07634	1.115	0.2648

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1184.57 on 299 degrees of freedom
Residual deviance: 526.97 on 296 degrees of freedom
AIC: 1176.7

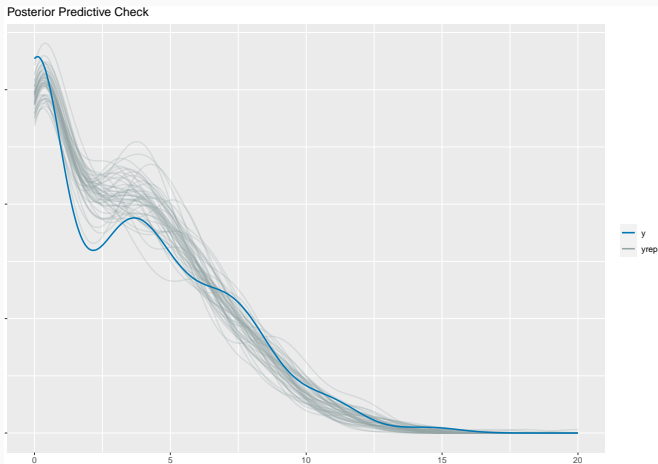
Visualising the fitted Poisson GLM



Checking Poisson GLM

Simulate data from fitted model (**yrep**) and compare with observed data (**y**)

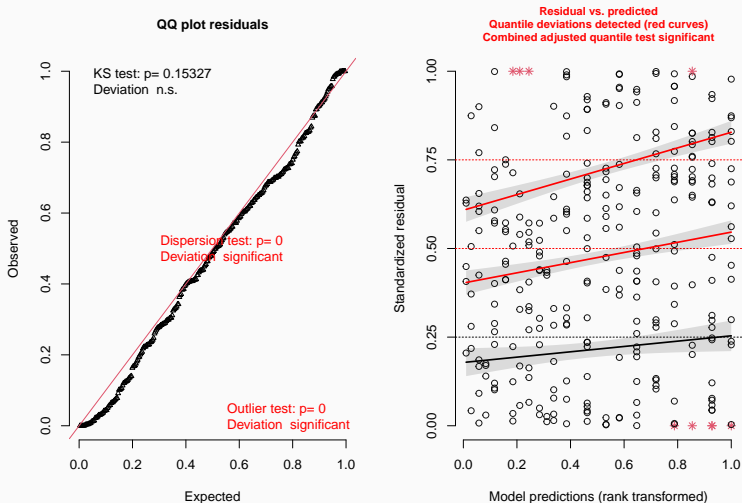
```
library("performance")  
pp_check(eggs.poi)
```



Checking Poisson GLM with DHARMA

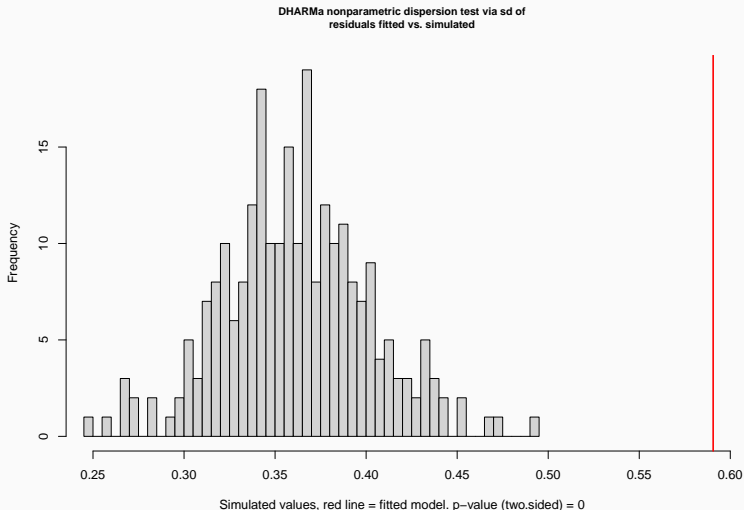
```
library("DHARMA")  
eggs.poi.res <- simulateResiduals(eggs.poi, plot = TRUE)
```

DHARMA residual diagnostics



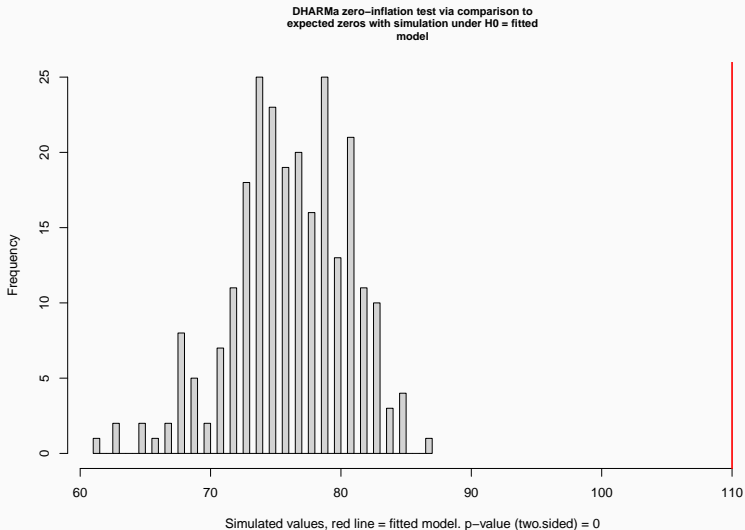
Checking overdispersion

```
testDispersion(eggs.poi.res)
```



Checking zero inflation

```
testZeroInflation(eggs.poi.res)
```



Accounting for zero-inflation

Mixture model:

1. Model probability of 0 (Binomial)

Mixture model:

1. Model probability of 0 (Binomial)
2. Model counts (including 0) (Poisson/Negative Binomial)

Modelling egg number as Zero-Inflated Poisson (ZIP)

Nests may be occupied or not:

Probability nest not occupied ~ old (Binomial)

For occupied nests:

Number of eggs ~ Nest diameter (Poisson)

```
library("glmmTMB")
eggs.zip <- glmmTMB(n.eggs ~ diameter,
                    family = "poisson",
                    ziformula = ~ old,
                    data = eggs)
```

Modelling egg number as Zero-Inflated Poisson

```
Family: poisson ( log )  
Formula:          n.eggs ~ diameter  
Zero inflation:    ~old  
Data: eggs
```

AIC	BIC	logLik	deviance	df.resid
993.8	1008.6	-492.9	985.8	296

Conditional model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.41622	0.13619	3.056	0.00224 **
diameter	0.11248	0.01155	9.737	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-inflation model:

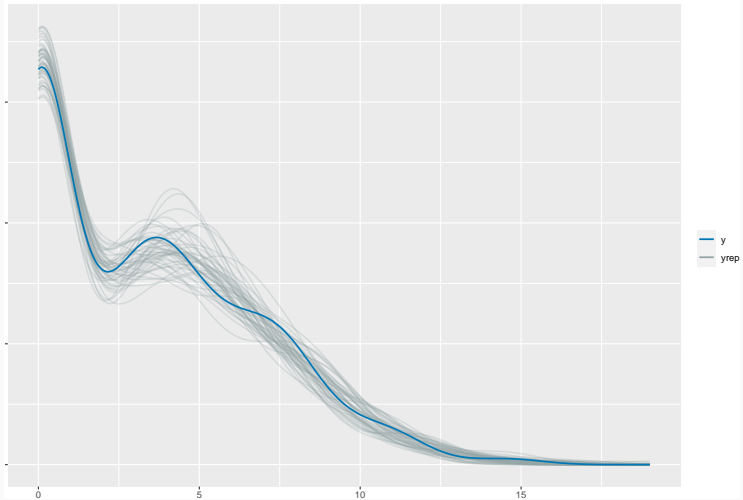
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4054	0.2803	-8.582	<2e-16 ***
oldyes	5.4897	0.5830	9.416	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Checking ZIP model

```
pp_check(eggs.zip)
```

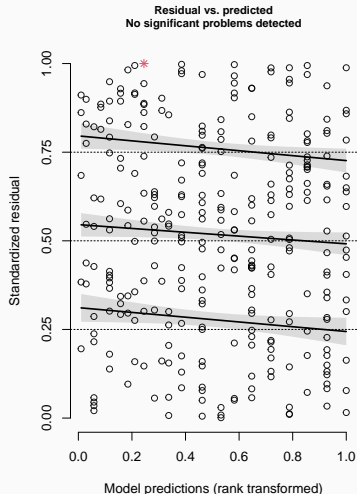
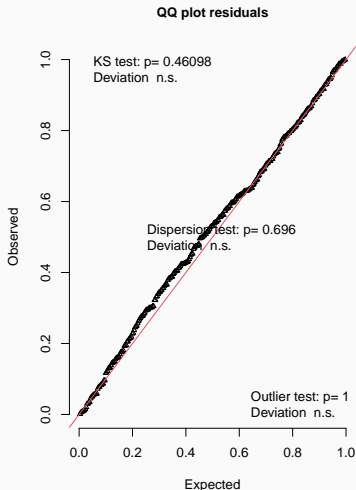
Posterior Predictive Check



Checking ZIP model with DHARMA

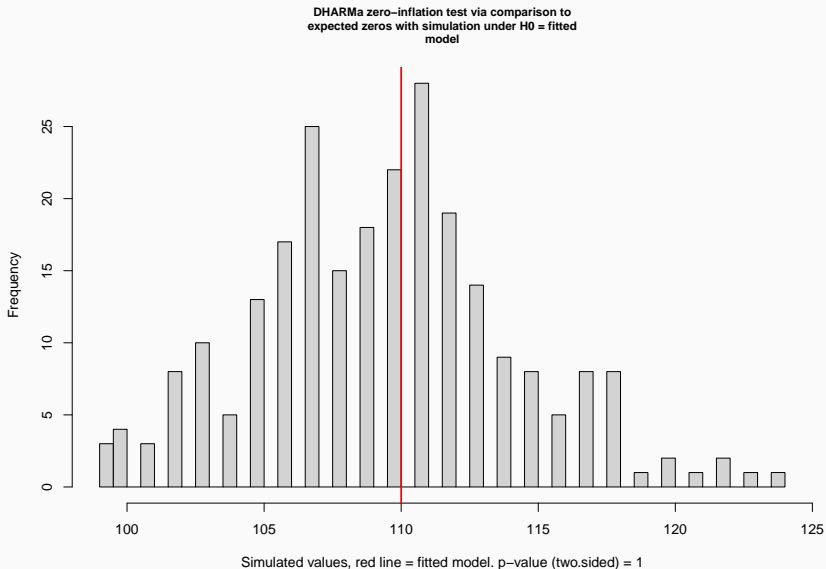
```
eggs.zip.res <- simulateResiduals(eggs.zip, plot = TRUE)
```

DHARMA residual diagnostics



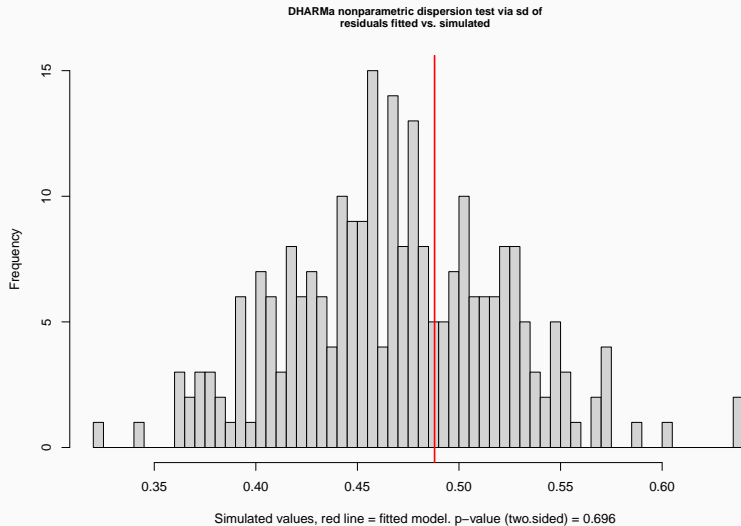
Checking ZIP model with DHARMA

```
testZeroInflation(eggs.zip.res)
```



Checking ZIP model with DHARMA

```
testDispersion(eggs.zip.res)
```



Modelling egg number as Zero-Inflated Negative Binomial (ZINB)

(If there were overdispersion with Poisson)

```
eggs.zinb <- glmmTMB(n.eggs ~ diameter,  
                     family = "nbinom2",  
                     ziformula = ~ old,  
                     data = eggs)
```

Modelling egg number as ZINB

```
Family: nbinom2 ( log )
Formula:          n.eggs ~ diameter
Zero inflation:    ~old
Data: eggs
```

AIC	BIC	logLik	deviance	df.resid
995.7	1014.2	-492.8	985.7	295

Dispersion parameter for nbinom2 family (): 143

Conditional model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.4118	0.1389	2.964	0.00304 **
diameter	0.1128	0.0118	9.561	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-inflation model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4160	0.2846	-8.489	<2e-16 ***
oldyes	5.4995	0.5850	9.401	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparing models

```
library("parameters")
compare_models(eggs.poi, eggs.zip, eggs.zinb)
```

Parameter	eggs.poi	eggs.zip	eggs.zinb
(Intercept)	0.31 (0.06, 0.56)	0.42 (0.15, 0.68)	0.41 (0.14, 0.68)
diameter	0.11 (0.09, 0.14)	0.11 (0.09, 0.14)	0.11 (0.09, 0.14)
old (yes)	-3.79 (-5.60, -1.98)		
old (yes) * diameter	0.09 (-0.06, 0.23)		
Observations	300	300	300

Comparing models

```
library("performance")
compare_performance(eggs.poi, eggs.zip, eggs.zinb)
```

```
# Comparison of Model Performance Indices
```

Name	Model	AIC	BIC	RMSE	Sigma	Score_log	Score_sph
eggs.poi	glm	1176.701	1191.516	2.324	1.334	-1.948	
eggs.zip	glmmTMB	993.790	1008.605	2.324	1.000	-1.643	
eggs.zinb	glmmTMB	995.666	1014.185	2.324	143.279		

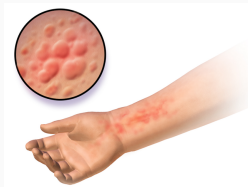
Accounting for zero-inflation with hurdle models

Tracking measles outbreak

Counting number of hives/person

Many people not sick (0 hives)

Those sick, have many hives (>1)



ZIP/ZINB:

1. Binomial model: probability of zero

Hurdle:

ZIP/ZINB:

1. Binomial model: probability of zero
2. Count model (Poisson/NegBin) includes zero

Hurdle:

ZIP/ZINB:

1. Binomial model: probability of zero
2. Count model (Poisson/NegBin) includes zero

Hurdle:

1. Binomial model: probability of non-zero

ZIP/ZINB:

1. Binomial model: probability of zero
2. Count model (Poisson/NegBin) includes zero

Hurdle:

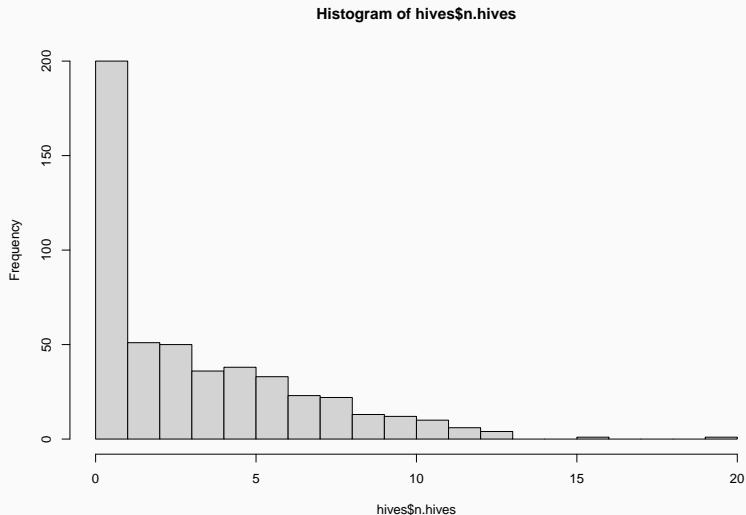
1. Binomial model: probability of non-zero
2. Count model truncated at 1

How many hives per skin area?

```
hives <- read.csv("data/hives.csv")
```

age	vaccinated	area.cm2	n.hives
Min. : 1.0	Min. :0.000	Min. : 5.000	Min. : 0.000
1st Qu.:23.0	1st Qu.:0.000	1st Qu.: 6.000	1st Qu.: 0.000
Median :45.0	Median :1.000	Median : 8.000	Median : 2.000
Mean :44.7	Mean :0.648	Mean : 7.482	Mean : 3.256
3rd Qu.:65.0	3rd Qu.:1.000	3rd Qu.: 9.000	3rd Qu.: 5.250
Max. :90.0	Max. :1.000	Max. :10.000	Max. :20.000

Many people with 0 hives

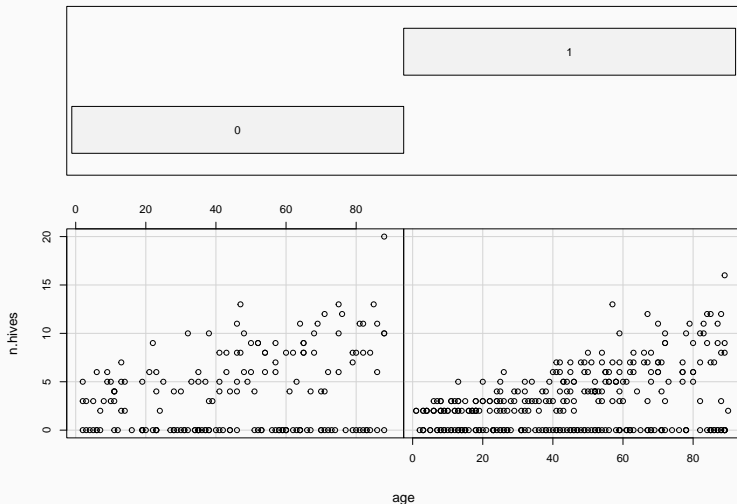


(that does not mean we need zero-inflated model!)

Number of hives ~ age * vaccinated

```
coplot(n.hives ~ age | as.factor(vaccinated), data = hives)
```

Given : as.factor(vaccinated)



```
hives.poi <- glm(n.hives ~ vaccinated * age,  
                 offset = log(area.cm2),  
                 data = hives,  
                 family = poisson)
```

Trying Poisson GLM

Call:

```
glm(formula = n.hives ~ vaccinated * age, family = poisson, data = hives,  
     offset = log(area.cm2))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.0081	-2.2235	0.1396	1.2155	4.2198

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.363696	0.095097	-14.340	< 2e-16 ***
vaccinated	-0.334184	0.122887	-2.719	0.00654 **
age	0.013626	0.001623	8.395	< 2e-16 ***
vaccinated:age	0.002034	0.002075	0.980	0.32708

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

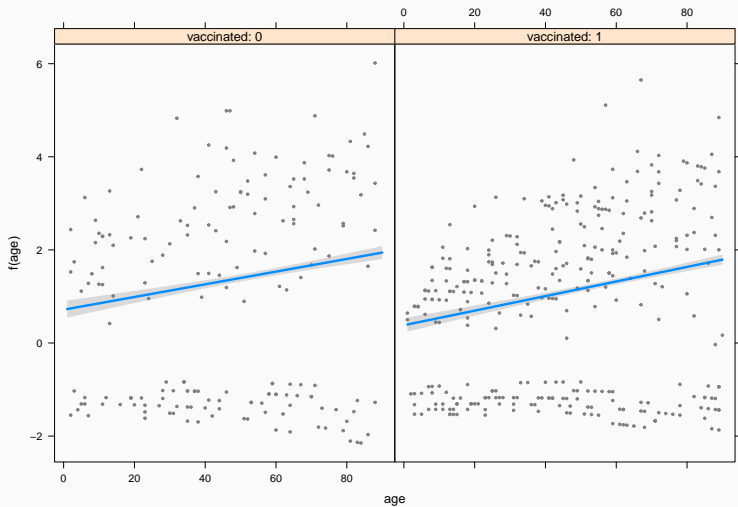
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2137.0 on 499 degrees of freedom

Residual deviance: 1891.7 on 496 degrees of freedom

AIC: 2925.6

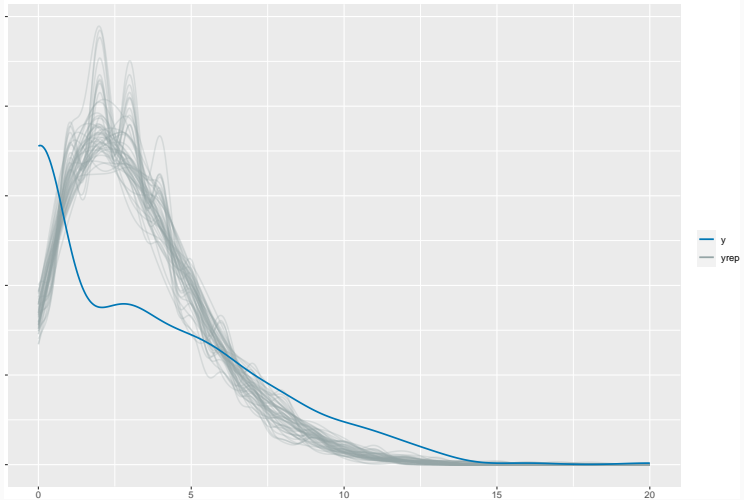
Visualising fitted Poisson GLM



Checking Poisson GLM

```
pp_check(hives.poi)
```

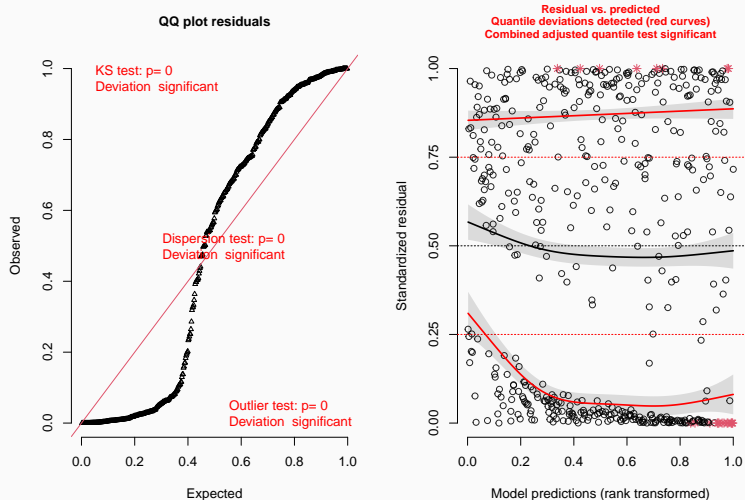
Posterior Predictive Check



Checking Poisson GLM

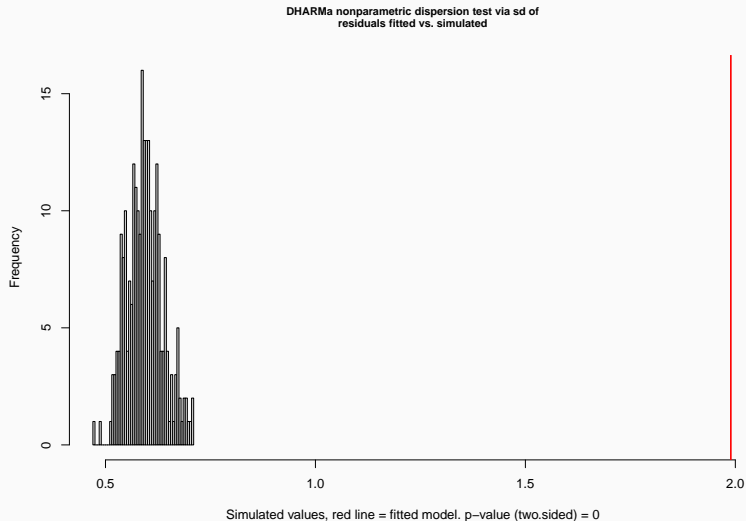
```
hives.poi.res <- simulateResiduals(hives.poi, plot = TRUE)
```

DHARMA residual diagnostics



Checking overdispersion

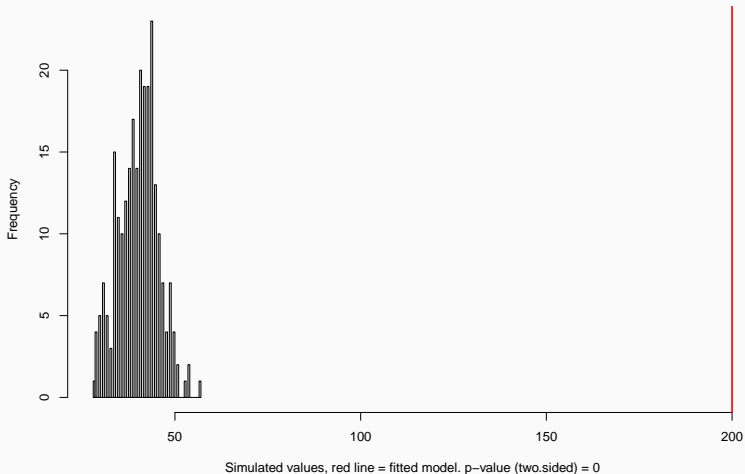
```
testDispersion(hives.poi.res)
```



Checking zero inflation

```
testZeroInflation(hives.poi.res)
```

DHARMA zero-inflation test via comparison to
expected zeros with simulation under H_0 = fitted
model



Accounting for zero-inflation with hurdle model

```
hives.hur <- glmmTMB(n.hives ~ vaccinated + age,  
                    family = truncated_poisson,  
                    ziformula = ~ 1,  
                    offset = log(area.cm2),  
                    data = hives)
```

Accounting for zero-inflation with hurdle model

```
Family: truncated_poisson ( log )
Formula:          n.hives ~ vaccinated + age
Zero inflation:    ~1
Data: hives
Offset: log(area.cm2)
```

AIC	BIC	logLik	deviance	df.resid
1932.1	1949.0	-962.1	1924.1	496

Conditional model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.853885	0.070755	-12.068	< 2e-16 ***
vaccinated	-0.365664	0.051532	-7.096	1.29e-12 ***
age	0.014860	0.001065	13.955	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Zero-inflation model:

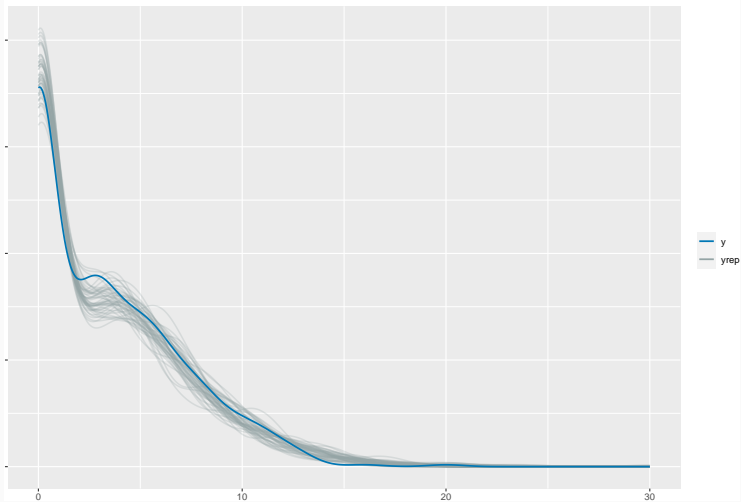
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.40547	0.09129	-4.442	8.93e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Checking hurdle model

```
pp_check(hives.hur)
```

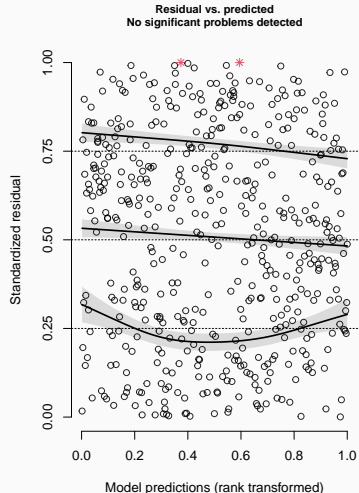
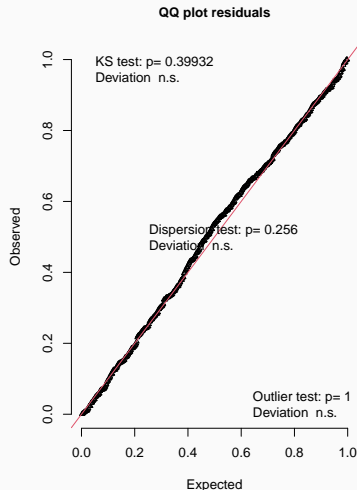
Posterior Predictive Check



Checking hurdle model with DHARMA

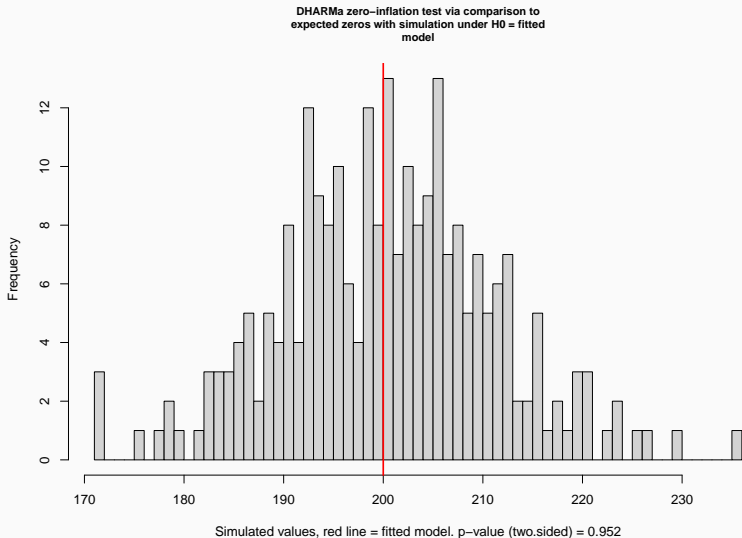
```
hives.hur.res <- simulateResiduals(hives.hur, plot = TRUE)
```

DHARMA residual diagnostics



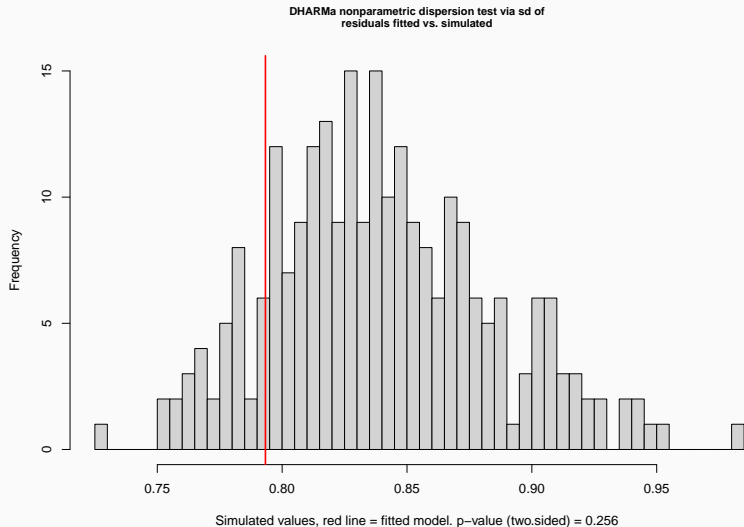
Checking zero inflation

```
testZeroInflation(hives.hur.res)
```



Checking overdispersion

```
testDispersion(hives.hur.res)
```



Comparing models

```
compare_models(hives.poi, hives.hur)
```

Parameter	hives.poi	hives.hur
(Intercept)	-1.36 (-1.55, -1.18)	-0.85 (-0.99, -0.72)
vaccinated	-0.33 (-0.58, -0.09)	-0.37 (-0.47, -0.26)
age	0.01 (0.01, 0.02)	0.01 (0.01, 0.02)
vaccinated * age	2.03e-03 (0.00, 0.01)	
Observations	500	500

Comparing models

```
compare_performance(hives.poi, hives.hur)
```

```
# Comparison of Model Performance Indices
```

Name	Model	AIC	BIC	RMSE	Sigma	Score_log	Score_spher
hives.poi	glm	2925.603	2942.462	3.299	1.953	-2.918	0
hives.hur	glmmTMB	1932.124	1948.982	3.313	1.000	-2.262	0