		Definition
predictor	trial	trial number
predictor	location	location on plate
predictor	treatment	survival or purchased honey bees
predictor	colony_id	colony identification number
predictor	time	amount of time on
predictor	larval_age	age of the larva
response	num_mite	number of mites

1. Prepare the data by factoring, remove NAs or zero cells, and possible collapsing of categories. We collapse treatment to zeros and ones. We also collapse trial, location, treatment, and colony_id. We will keep **treatment** as it is a primary interest of the investigator.

2. We then conduct uni-variable analysis of each predictor with the response variable. We check if the p-values are less than 0.25. These are considered candidates for the preliminary model as well as clinically meaningful variables.

predictor	p-value	p-value < 0.25	Candidate
trial	trial number	2.00E-16	Yes
location	location on plate	0.593	No
treatment	survival or purchased honey bees	0.423	
colony_id	colony identification number	0.105	Yes
time	amount of time on	0.0937	Yes
larval_age	age of the larva	7.51E-10	Yes

Figure 1: Variable candidates for the model

3.The candidates for the model are trial, treatment, colony_id, time, and larval_age. Due to the zero inflated data we will use the Poisson model.

```
glm(formula = num_mite ~ trial + treatment + colony_id + time +
    larval_age, family = "poisson", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2746  -0.8952  -0.6642  -0.4425   6.6753

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.225337   0.238948 -13.498  < 2e-16 ***
trial        0.333105   0.026584  12.530  < 2e-16 ***
treatment    0.857788   0.157072   5.461 4.73e-08 ***
colony_id   -0.134011   0.023494  -5.704 1.17e-08 ***
time         0.001631   0.000973   1.676   0.0937 .
larval_age   0.186639   0.030323   6.155 7.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Unadjusted model with candidates

4. We will remove predictors that have a high p-value that pose no change in the delta beta percent value. We remove the time variable due to its high p-value. After removing time, we see that there is no change in the dbp from Table 1 and there are no p-value over 0.05 in Figure 3.

```
glm(formula = num_mite ~ trial + treatment + colony_id + larval_age,
    family = "poisson", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2221  -0.9185  -0.6501  -0.4295   6.7842

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.11368    0.22920 -13.585  < 2e-16 ***
trial        0.33310    0.02658  12.530  < 2e-16 ***
treatment    0.85779    0.15707   5.461 4.73e-08 ***
colony_id   -0.13401    0.02349  -5.704 1.17e-08 ***
larval_age   0.18664    0.03032   6.155 7.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Adjusted model after removing time variable

Predictor	unadjusted model est	removed time variable	dbp
trial	0.333105	0.33310	0
treatment	0.857788	0.85779	0
colony_id	-0.134011	-0.13401	0
larval_age	0.186639	0.18664	0

Table 1: Delta Beta Percent after removing time

5. We check the continuous variable to see if it is linear in the logit. We show below in figure 4.

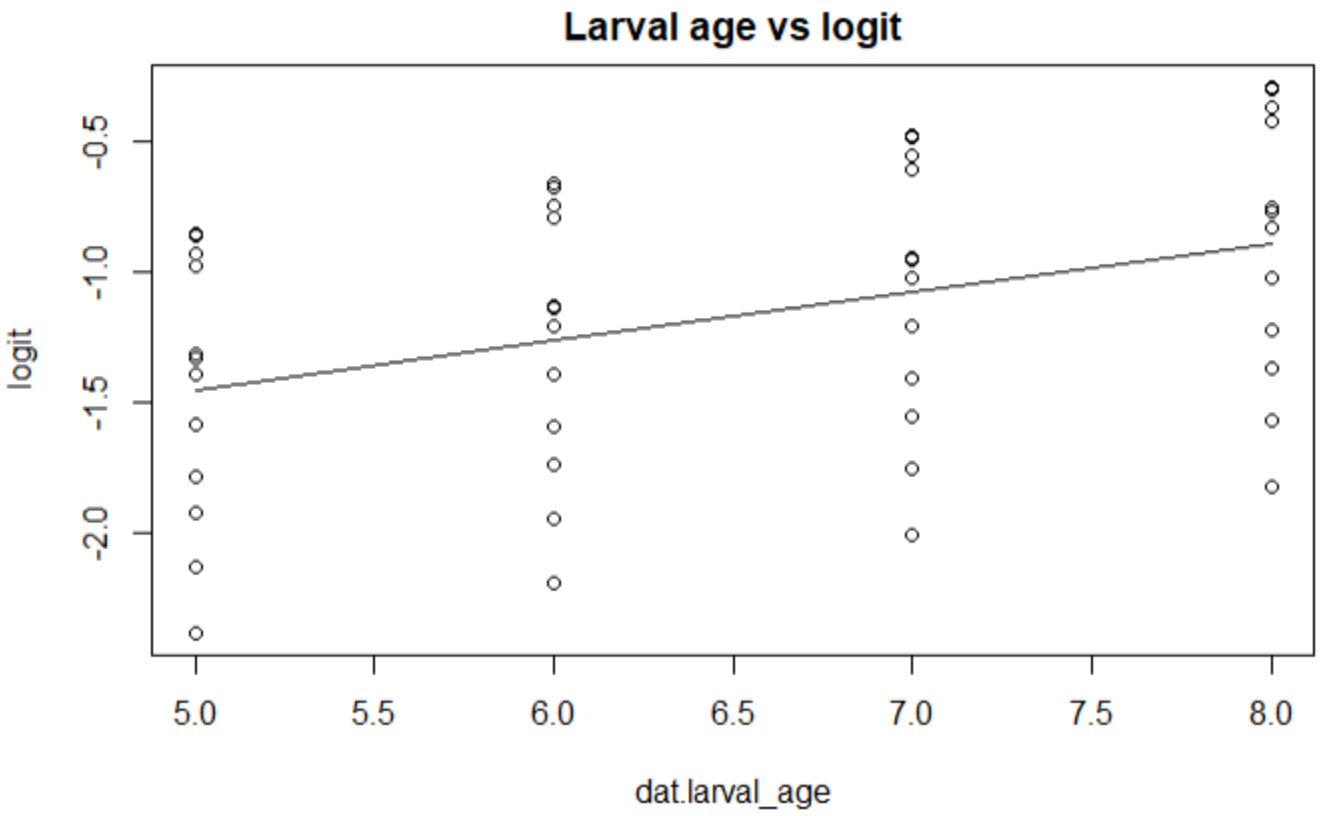

Figure 4: Larval age vs logit

6. We check for interactions by looking at plausible pairs and check their p-values. If the p-value is less than 0.05 then it is significant enough to be added to the main effects model.

Interaction		p-value	significant
trial	treatment	0.984877	No
trial	colony_id	0.52533	No
trial	larval_age	0.193384	No
treatment	colony_id	0.864	No
treatment	larval_age	0.000916	Yes
colony_id	larval_age	0.0405	Yes

7. The preliminary final model is

num_mite ~ trial + treatment + colony_id + larval_age + colony_id:larval_age + treatment:larval_age

```
Call:
glm(formula = num_mite ~ trial + treatment + colony_id + larval_age +
    colony_id:larval_age + treatment:larval_age, family = "poisson",
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3070  -0.8627  -0.6679  -0.4596   6.9943

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.13540    0.52540  -5.968 2.41e-09 ***
trial                 0.33345    0.02660  12.534  < 2e-16 ***
treatment            -1.66450    0.82717  -2.012  0.04419 *
colony_id             0.05815    0.11854   0.491  0.62377
larval_age            0.18885    0.07653   2.468  0.01360 *
colony_id:larval_age -0.02866    0.01732  -1.655  0.09795 .
treatment:larval_age  0.37556    0.12117   3.099  0.00194 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2932.9  on 2559  degrees of freedom
Residual deviance: 2698.6  on 2553  degrees of freedom
AIC: 4053.3

Number of Fisher Scoring iterations: 6
```

8. Assess the fit of the model with pearson test, chisquare test, and goodness of fit.
The p-value is for chi square test is 0 which is significant. The low p-value means that the model fit is not good.

The p-value for HL test is <2.2e-16 meaning that this is significant. Therefore, we reject the model fit. Meaning the model is not a good fit.

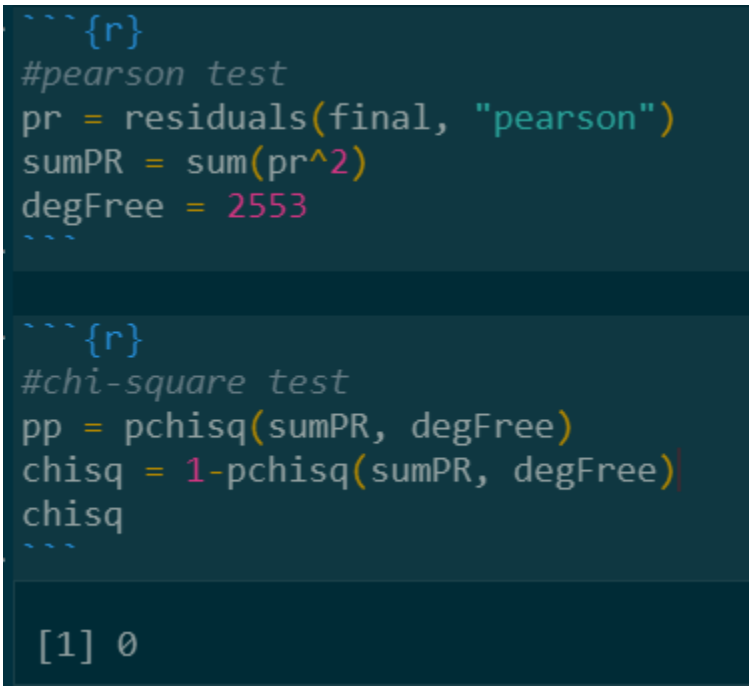

Figure 5: Pearson and ChiSquare Test

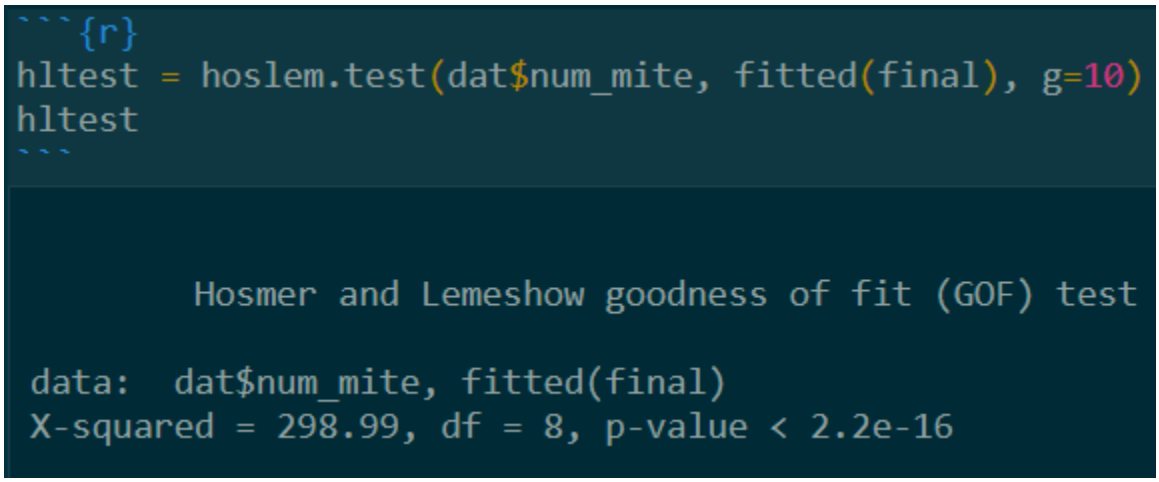

Figure 6: HL Test

```{r}
cbind(hltest$observed, hltest$expected)
```

	y0	y1	yhat0	yhat1
[0.106,0.135]	251	5	226.8876	29.11243
(0.135,0.184]	264	24	242.4407	45.55930
(0.184,0.224]	183	41	176.6076	47.39236
(0.224,0.27]	181	75	193.3487	62.65126
(0.27,0.331]	182	74	179.8769	76.12307
(0.331,0.368]	114	142	164.8239	91.17613
(0.368,0.411]	91	165	159.0963	96.90368
(0.411,0.524]	191	65	132.6772	123.32282
(0.524,0.619]	87	201	123.6222	164.37783
(0.619,0.854]	120	104	64.6189	159.38110

Returning to no interactions.