Data analysis 3: Removal of caterpillars from salad

B030831

June 28, 2021

1 Introduction

In this analysis, we look at data collected during an experiment. A series of tests (24 tests in total) was conducted, under a specified shaking force (ranging from 1 to 5 on an arbitrary scale, but only 3 and 4 were used here), exposure time (in seconds) and in a modified atmosphere where a gas acting as a muscle relaxant was added (measured as a concentration percentage) in order to remove caterpillars from salads. Each salad was placed with 12 caterpillars on its leaves (one caterpillar per leave) and was then put under different atmospheres, shaking forces and exposure time. Each caterpillar was only used once. The number of caterpillars dislodged at the end of the shaking period was then recorded. Table 1 below summarizes the data collected during the experiment.

		Number dislodged	
Gas concentration (%)	Exposure time (s)	Force 3	Force 4
25	10	10	11
	20	11	12
	60	12	12
12.5	10	10	9
	20	12	10
	60	11	10
5	10	8	5
	20	9	7
	60	11	9
2.5	10	8	4
	20	8	6
	60	10	5

Table 1: Number of caterpillars dislodged under different conditions.

We are interested in how the number of caterpillars dislodged depend on the shaking force, gas concentration and exposure time. This is clearly a binary count data where the total number of observations is known (12). The probability of success corresponds to the probability of a caterpillar being dislodged from the salad, during the experiment. We will thus fit a generalized linear model with a logit link function in order to calculate the probability of success. Note that a simple standard linear regression would not be appropriate as we would get a negative count. Firstly, we will start with the fullest model and proceed by eliminating the parameters which are not significant to arrive at the best fit for the data. Then, we will make a transformation on the exposure time variable in order to get a more realistic probability prediction when the exposure time is very small.

2 Model Selection

We want to fit a logistic regression model in order to estimate the probabilities of the caterpillars to be dislodged. We have three variables to fit: two continuous variables and one factor variable. Exposure time and gas concentration are treated as continuous variable while the shaking force is treated as a factor (since the shaking scale is arbitrary). We start with the fullest model that includes all the interactions; three 2 levels interaction and one 3 levels interaction. We can fit this model using the *glm* built in function in R. We then proceed to a backwards elimination, eliminating the least significant parameters step by step until we arrive at the best model. We can compare the fits between models by comparing their residual deviances. For two general models A and B with residuals deviances $Deviance(A)$ and $Deviance(B)$ with a and b ($a > b$) degrees of freedom respectively, we compare $Deviance(A) - Deviance(B)$ to a Chi-Squared with $a - b$ degrees of freedom. If we find the value to be significant, then model B offers a significantly better fit at a chosen level of significance. Here, we use the built in R function *step*(Full model). This built in function automatically proceed to the backwards elimination described above and selects the model with lowest Aikaike's criterion (AIC). AIC is given by

$$AIC = -2log(L) + 2k,$$

where L is the likelihood function of the model and k is the number of estimated parameters in the model. The model with lowest AIC will be the one that fits the data the best. However, as opposed to using residual deviances, one cannot proceed to a test statistics when using the AIC as a model selection criteria. We get from R that the model with the lowest AIC of 72.402 is one that contains the two continuous variables (exposure time and gas concentration), the shaking force factor and the interaction between the force factor and the gas concentration. The two other 2 levels interactions and the 3 levels interaction were deemed insignificant. Therefore our model is

$$log\left(\frac{\pi}{1-\pi}\right) = \eta + \alpha I + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 I, \tag{1}$$

where $I = 1$ when the shaking force is 4 and $I = 0$ when the shaking force is 3. The variable x_1 corresponds to gas concentration and the variable x_2 to the exposure time. The last term of equation (1) represents the interaction between the force factor and gas concentration. The probability that a caterpillar gets dislodged is π.

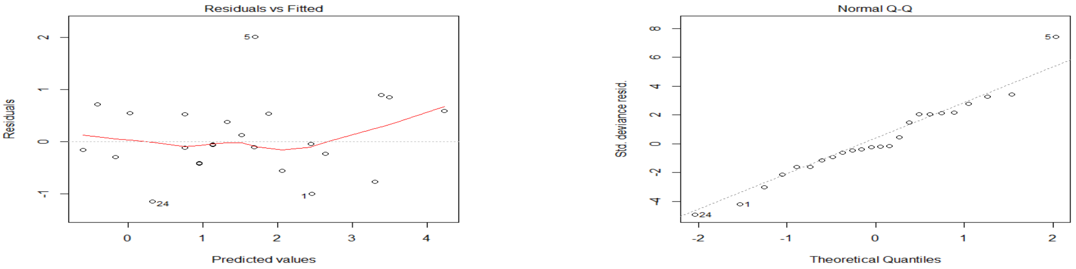

Figure 1: Residuals and Normality plots, respectively.

Let us look at the residual and the normality plots for this generalized linear model. We see from Figure 1 above that the variance looks homogeneous and that the observations are normally distributed. Observation number 5 seems to be an outlier. Table 2 below summarizes the parameter estimates and their p-value.

Parameters	Estimates	P-values
η	0.391	0.329
α	-1.61	0.000832
β_1	0.0753	0.0223
β_2	0.0185	0.0153
β_3	0.0983	0.0538

Table 2: Estimate of parameters and p-values.

We see from table 2 and equation (1) that the probability of a caterpillar being dislodged is greater at force 3 than force 4. This may be because force 3 is actually stronger than force 4, as the scaling was arbitrarily chosen. Furthermore, we see that both continuous variables have a positive impact on the probability of being dislodged, as expected. The interaction term for the shaking force of 4 also contributes positively to the probability of a caterpillar being dislodged. Now, let us inverse the logit function in (1) in order to get an expression for π. We obtain

$$\pi = \frac{exp(\eta + \alpha I + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 I)}{1 + exp(\eta + \alpha I + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 I)}. \tag{2}$$

And so an estimate of the probability π is given by

$$\hat{\pi} = \frac{exp(0.391 - 1.61I + 0.0753x_1 + 0.0185x_2 + 0.0983x_1 I)}{1 + exp(0.391 - 1.61I + 0.0753x_1 + 0.0185x_2 + 0.0983x_1 I)}. \tag{3}$$

Let us plot the function in (3) for given values of x_2 (fixed exposure time) for both a shaking force of 3 and a shaking force of 4.

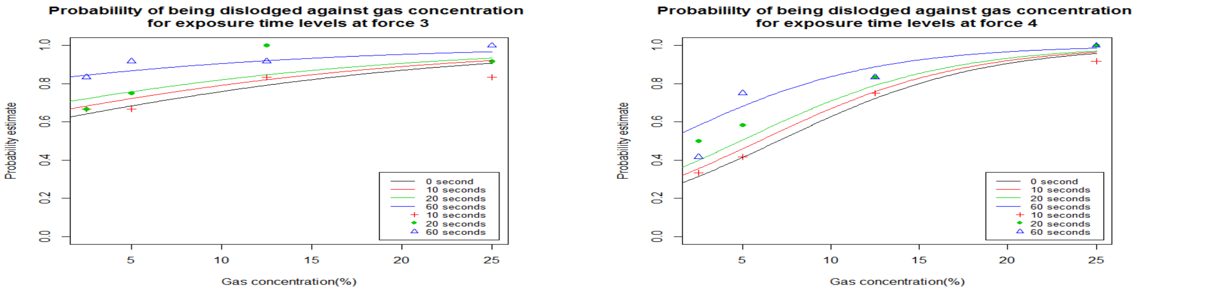

Figure 2: Fitted lines of equation (3) for some fixed exposure time levels. Shaking force 3 is on the left and force 4 on the right.

We see that, as expected, the gas concentration has less impact when the shaking force is 3, since $I = 0$ in equation (3). Thus the coefficient multiplying gas concentration is smaller. Figure 1 is simply to have a good overview of the the evolution of $\hat{\pi}$ as one could easily work out a particular probability estimate by just plugging the required exposure time and gas concentration in equation (3). However one can see from equation (3), and perhaps more easily from Figure 1, that when the exposure time is 0 (represented by the grey line in Figure 1), we still get a positive probability. Certainly, when the exposure time is 0, nothing has happened and the probability of being 0 should thus be 0 as well. Let us consider a data transformation that takes this feature into account.

3 Data transformation

As noted above, we require than the estimated probability of a caterpillar being dislodged tends to 0 as the exposure time also tends to 0. Thus looking at equation (3), we want that $\hat{\pi} \to 0$ as $x_2 \to 0$. We thus want the numerator in (3) to go to zero as x_2 goes to 0. This is easily done by applying a *log* transformation to x_2, since the exponential term goes to 0 as the exponent goes to $-\infty$. After the transformation, the model in (1) has an AIC of 71.232 which is slightly lower than the same model without the transformation. Once again, let us look at the residuals and normality plot of this model.

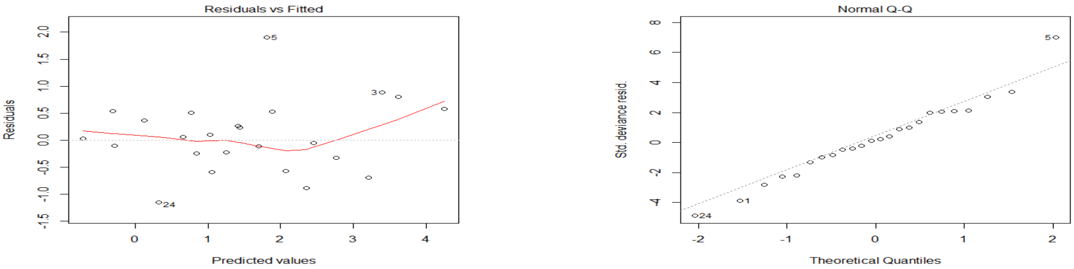

Figure 3: Residuals and Normality plots, respectively.

We see that the variance after the residuals plots are very similar for both models and there does not seem to be any noticeable pattern. Observation 5 again seems to be an outlier. However, the normality looks slightly better, which explains the lower AIC for this model. The new parameter estimates and their associated p-values are summarized in the table below.

Parameters	Estimates	P-values
η	-0.871	0.240
α	-1.62	0.000806
β_1	0.0756	0.0224
β_2	0.582	0.00780
β_3	0.0988	0.0530

Table 3: Estimate of parameters and p-values.

And so, after the *log* transformation, equation (3) becomes

$$\hat{\pi} = \frac{exp(-0.871 - 1.62I + 0.0756x_1 + 0.582 \cdot log(x_2) + 0.0988x_1I)}{1 + exp(-0.871 - 1.62I + 0.0756x_1 + 0.582 \cdot log(x_2) + 0.0988x_1I)}. \tag{4}$$

One could also wonder whether the estimated probability of a caterpillar being dislodged should also be 0 when the gas concentration is 0. When looking at the collected data, we see that a large proportion of caterpillars are dislodged at a 2.5% gas concentration, provided that the exposure time is large. Therefore, we have no reason to expect a similar result as the exposure time when the gas concentration is 0. Note that when $x_2 = 0$, the transformation $log(x_2)$ is not defined. Therefore, the model in (4) can only be used for a positive exposure time. This is not an issue as when the exposure time is 0, the experiment has not actually began, and so the estimated probability is 0, by definition. Let us now plot the equivalent graphs of Figure 1, this time with the log transformation on x_2.

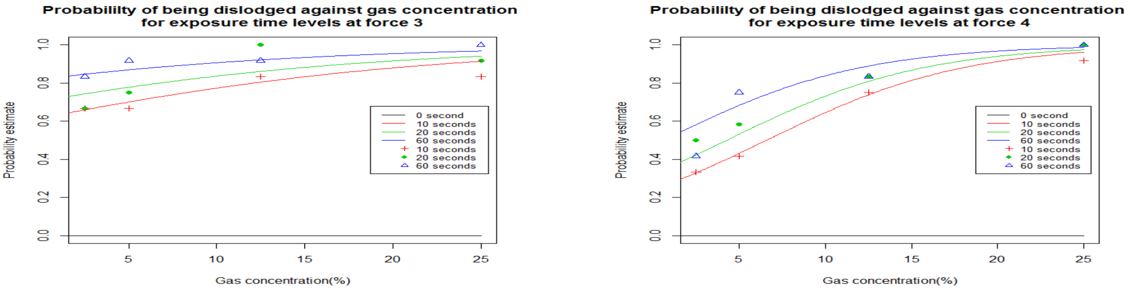

Figure 4: Fitted lines of equation (4) for some fixed exposure time levels. Shaking force 3 is on the left and force 4 on the right.

We see in Figure 4 that when the exposure time is 0, the estimated probability of a caterpillar being dislodged is 0, as we wanted. It is of interest to compute a few values of the estimated probability for both models and at some fixed exposure time and gas concentration is order to compare and see which prediction seems more reasonable. For a simpler readibility, we only compare estimated probabilities at a shaking force of 4. The values are summarized in the table 4 below.

		Estimated probabilities	
Gas concentration (%)	Exposure time (s)	Model (3)	Model (4)
25	1	0.96	0.87
	20	0.97	0.97
	100	1	0.99
12.5	1	0.72	0.43
	20	0.79	0.81
	100	0.94	0.91
5	1	0.42	0.17
	20	0.50	0.53
	100	0.82	0.73
0	1	0.23	0.076
	20	0.30	0.32
	100	0.65	0.55

Table 4: Estimated probabilities for fixed gas concentration and exposure time.

We see in Table 4 that the probabilities for an exposure time of 1 second seem to be too high and not realistic. When we calculates the proportion from the observations for a

25% gas concentration and an exposure time of 10 seconds, we get a probability of 0.91, while the estimated probability from model (3) is 0.96 for an exposure time of 1 second. Similarly, from the observations, for a 5% gas concentration and an exposure time of 10 seconds, the proportion is 0.416 while we get 0.42 for a 1 second exposure time for model (3). However, we would expect significantly smaller probabilities for an exposure time of 1 second compared to an exposure time of 10 seconds. The predictions from model (3) generally overestimate the probabilities, especially for short exposure time. On the other hand, we see that model (4) gives more sensible probability estimates (i.e smaller estimates) when the exposure time is short.

4 Conclusion

After fitting the most general model including all the interaction terms and proceeding to a backward elimination, we concluded that model (1) was the one which fitted the data best with an AIC of 72.402. However there is an issue with this model as the estimated probability does not go to 0 as exposure time goes 0, which should be the case in experimentally. Generally, for short exposure time, the model was overestimating the probability of a caterpillar being dislodged, as was shown in Table 4. We thus came up with a *log* transformation on the exposure time in order to make the estimated probability go to 0 as exposure time goes to 0. This transformed model offered a better fit to the data with an AIC of 71.232 and gave more sensible predictions. Indeed, as shown in Table 4, at an exposure time of 1 second, the predicted probabilities are considerably lower than the ones estimated by the first model. Therefore, model (4) should be used in order to estimate the probabilities of a caterpillar being dislodged, given a gas concentration, a shaking force and an exposure time. It would be of interest to carry out the same experiment with a gas concentration of 0% in order to know whether a *log* transformation on gas concentration makes sense. With the current observations, we see that caterpillars are still getting dislodged at a gas concentration of 2.5%. There is thus little evidence for us to think that they would not fall with a gas concentration of 0%. With the current information we have from the experiment, taking both the log transformation of gas concentration and exposure time gives a worse fit with an AIC of 71.745.