Renel Chesak

April 2014

Bio 548

Statistical Summary

**Connect Model to Biology**

When an oil spill occurs, the immediate effects on wildlife and the environment are devastating. Some oil companies claim that through the process of ‘weathering,’ the oil is dispersed, evaporated, and otherwise made harmless to wildlife. In this experiment, the researchers look at the effects of indirect oil contact on crickets in comparison to no contact and direct contact, to find out whether exposure to ‘weathered’ oil is harmless or not. In other words, does the treatment (direct/indirect exposure to oil and the control) affect cricket survivorship? Previous research has shown that direct exposure to oil has a significant effect, so the experimenter is really most interested in the following central biological question: Does indirect exposure to oil have a significantly greater effect on cricket mortality than no exposure to oil (control)?

Toy model:

Mortality <=? Treatment

Using a generalized linear mixed model to test these relationships will allow us to answer our central biological question. This test allows us to compare differences in natural log odds (e.g. odds of death with control treatment vs. odds of death with indirect treatment), which is similar to looking at differences in means. This means it will allow us to determine whether the effects of treatments are significantly different.

This is a hypothesis testing situation. The experimenter had three reps for each treatment, each with 6 crickets (total of 18 crickets for each treatment within one plot). One plot describes a temporal block where each individual treatment was given (each treatment having three reps within a plot; total of 54 crickets in one plot). An experimental unit is one cricket. Based on this, the expanded toy model is:

Mortality <=? Treatment + plot + rep(plot) + trt\*plot + trt\*rep(plot)

Mortality is a random, binomial response variable where 0 = alive and 1 = dead. Treatment is a fixed, categorical predictor variable with three levels: control (no oil exposure), indirect (oil in pan below cricket, but no direct contact), and direct (direct oil exposure). Plot is a random, categorical predictor variable that is included because there could be variation between temporal plots. The last three variables are random, categorical predictor variables. Rep nested within plot is included because there may be variation associated with each rep within a plot. Treatment by plot is included because there may be variation specific to a treatment in different plots. Treatment by rep nested within plot is included because there could be variation specific to treatment in reps within plots.

Sample dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cricket | plot | rep | time | treatment | Alive/Dead |
| 1 | 1 | 1 | 2 | control | 1 |
| 2 | 1 | 1 | 2 | control | 0 |
| 3 | 1 | 1 | 2 | control | 0 |
| 4 | 1 | 1 | 2 | control | 0 |
| 5 | 1 | 1 | 2 | control | 0 |
| 6 | 1 | 1 | 2 | control | 0 |

**Accommodating Realities**

The Likelihood Ratio (LR) test was used to test whether the random variables could be removed from the model. It consists of backward stepwise regression, and detects variables that do not significantly change the fit of the model so that they can be thrown out. Specifically, it determines the contribution of a specific random variable by comparing the fit of the model with and without the variable ([Bolker, Brooks et al. 2009](#_ENREF_2)). The fit is measured as deviance, in this case, -2 times the log-likelihood ratio. The likelihood ratio consists of the likelihood of the current model divided by the likelihood of the saturated model, in which there are as many parameters as there are data points. The reason it gets transformed by -2 times its log is so that the distribution can be known, and therefore can be used for hypothesis testing purposes. The deviance (D) is calculated for the model with and without the variable in question, and the change in D is used to assess whether the variable is important. The larger the change in D, the more likely it is that the variable was important. The following table uses data from running PROC GLIMMIX using the method=LAPLACE option:



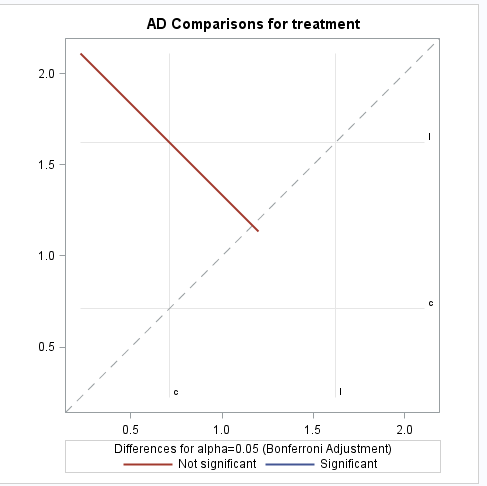
The lower the -2 log-likelihood (-2 LL), including negative numbers, the better the fit of the model. You can see that the best fit model includes all of the random variables we are testing. But the question remains as to whether the difference in fittness is signigicant when comparing the -2 LL of models with and without particular variables. For this, we look at the table below, which looks at the change in D for two models being compared, the change in degrees of freedom (DF), and the 2-tailed P value associated with that Chi Squared value. Since some of the models only contain one DF, that P value can be corrected for a 1-tailed test by divinding by two. From these P values, we can see that only two comparisons are close to being significant. Comparing models A and B gives us a nearly significant P value (P=0.08748). Comparing models D and E gives us a significant P value (P=0.04743). Thus, we can see that the only important random effect is plot, which is the random factor that was being tested in the comparison between D and E. Based on this, it is clear that rep(plot), trt\*plot, and trt\*rep(plot) could be removed without significantly changing the model. Thus we are left with the following revised model:

Mortality <=? Treatment + plot



LR tests are preferred over other tests, such as Wald tests, because it works well with random as well as nested factors and requires fewer assumptions.

In the original dataset, there were three time points at which cricket mortality was recorded for each treatment. However, for the purposes of the specific biological question being asked, it was deemed that one time point was sufficient to answer the question. Running the statistics on all three time points would require the SAS PROC GLIMMIX procedure to be run three times, once for each time point, and thus there would be three times as much data to analyze. It would also require a repeated measures analysis of variance because more than one observation was taken on a single cricket. This would control for any variance seen between the time points. Such an inquiry would be redundant because the same pattern appears in all of the time points. Upon running the third time point in PROC GLIMMIX, the indirect treatment was not significantly different from the control. The direct treatment was dropped from the model because all crickets were dead, and thus all data points were zero. Because one cannot take the log of zero, this treatment cannot be run through a GLM. Because the central biological question is most interested in the comparison between the control and indirect treatments, and also because the significance of the direct treatment is intuitively large (given that all crickets are dead), this seems a fair option. The results of the model at the third time point are summarized in the following diffogram:



This plot shows whether the odds of death are significantly different between treatments. The axes are the natural log odds (a transformation of “how likely death is / how likely life is”) for each treatment; this is replicated on the x- and y-axis. The dotted diagonal line represents where the natural log odds are the same on both axes (and thus the same for the treatments being compared). The plot is a visual representation of how different the treatments’ natural log odds are. In other words, it compares the natural log odds of each treatment to see if they are significantly different. If a point falls along the dotted diagonal line, the natural log odds are not significantly different. The center of each solid line represents the natural log odds comparison, and the solid line itself represents the 95% confidence interval around the natural log odds. The further the center of each solid line is away from the dotted diagonal line, the more significantly different the treatments being compared are. Here we can see that indirect exposure to oil *does not* have a significantly greater effect on cricket mortality than no exposure to oil (control) (this is seen as the red solid line). This model was run using the same code as for the first time point, except that the direct treatment was dropped from the raw data, and the “where time=3” option is used to designate the third time point.

**Looking at data**

This way of displaying the data is designed to assess the relative effects of treatment on mortality. We can see here that the direct treatment has the biggest effect on mortality, followed by the indirect treatment. We can see that the indirect exposure leads to a slightly higher percent mortality than the control, but we cannot say whether it is significant without running a generalized linear mixed model.

This way of displaying the data is designed to assess how mortality varies within plots across treatments. We can see that for the control, there is some variation. For the indirect treatment, there is no variation. For the direct treatment, there is large variation across plots. Given this variation, it seems that plot is an important variable in our model.

**Analyze data**

To analyze this data, I used a generalized linear mixed model (GLMM) because the data is binomial, and also includes random effects. Because the data is binomial, it cannot fit a normal distribution nor can it have homogeneity of variance, so it does not fit the assumptions of parametric tests. GLMMs combine linear mixed models (which include random effects) and generalized linear models (which deal with binomial data). In this model, I specify the distribution as binary (because of the binomial nature of the data); the link function as logarithmic (gives the odds; odds = how likely death is / how likely life is), which is typical for count data ([Bolker, Brooks et al. 2009](#_ENREF_1)); and that the random effect, plot, could vary randomly. This test allows us to compare differences in natural log odds (e.g. odds of death with control treatment vs. odds of death with indirect treatment), which is similar to looking at differences in means. This means it will allow us to determine whether the differences in the effects of treatments are significant, which will allow us to answer our central biological question. The following is the SAS (9.3) code used:

**proc** **glimmix** data=work.hello plots=diffogram;

where time=**1**;

class treatment plot;

model ad (event='1')=treatment / dist=binary link = logit ddfm=kr;

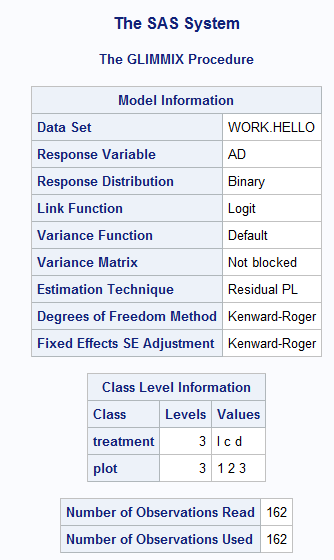
random plot;

lsmeans treatment/ilink cl pdiff adjust=bon oddsratio;

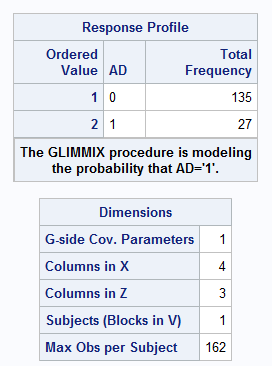
**run**;

The plots=diffogram function gives you a diffogram plot, which is a visual representation of a measure of the confidence interval around the odds ratio. “where time=1” is included because the actual dataset includes three time points when alive/dead was assessed, but in this analysis we are only interested in the first. The (event=’1’) specifies that we are interested in looking at mortality (where 0=alive and 1=dead). The Kenward –Roger (ddfm=kr) option dictates that denominator degrees of freedom for t and F tests are approximated according to Kendward in Roger, as well as adjusting the standard error. This method is cited as being the best option for LMMs ([Bolker, Brooks et al. 2009](#_ENREF_1)). The lsmeans treatment function asks for the natural log odds. The ilink (inverse link) function asks for the probability of death for each treatment (the number of deaths / the number of trials). The cl function asks for the confidence limit. The pdiff function asks for the probability that two things are different. The adjust=bon adjusts the acceptable level for a Type I error, because we conducted three individual tests (three plots). The odds ratio function asks for the natural log odds ratio (e.g. ln(odds when x= indirect treatment) – ln(odds when x=direct treatment)).

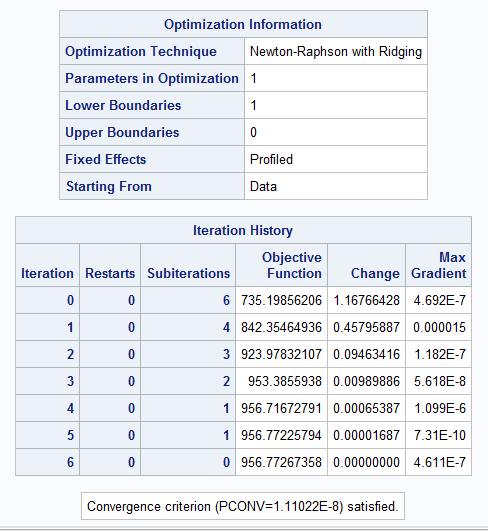
The following is the analysis given by running the above code:

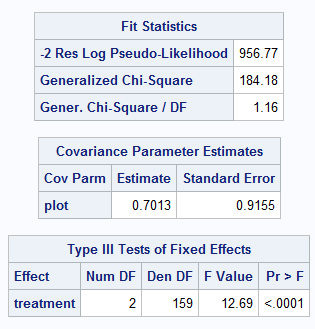


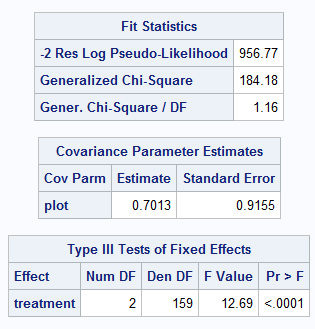
Everything appears to be in order. Treatment and plot both have 3 levels. There are 162 observations (54 per plot X 3 plots = 162).



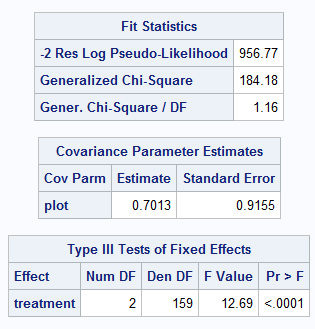
This confirms that we are looking at the probability that a cricket will be dead.



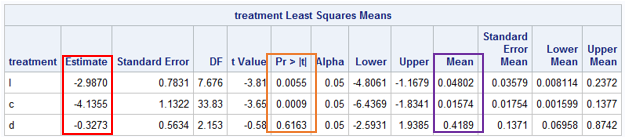
This first value is important when comparing how well the data fit different models, as described in the accomidating realities section. The lower the number (including negative numbers), the better the fit. This can be helpful when attempting to simplify the model by dropping interaction terms. For example, if the fit of the model differs by a large value, the term that was dropped was important, and must be included.



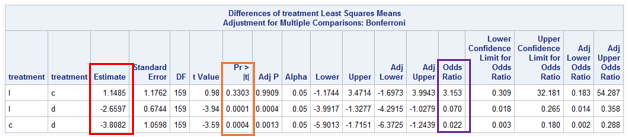
This tells us how much plot covaries with treatment. This high value indicates that it is important to include in the model, because it accounts for a significant portion of the variance.



Here we can see that the fixed effect (treatment) significantly (p<0.001) effects the response variable (mortality). The degrees of freedom (2) are as expected, since there are 3 levels (c, i, d).



Odds = how likely death is / how likely life is. The estimate (outlined in red) is the natural log odds (NLO), which is simply a transformation of the odds that allows the data to fall between negative infinity and positive infinity (so it will have great distribution qualities that are good for GLMs). We can see that in the control treatment, the odds of death are the lowest (NLO=-4.1355), followed by the indirect treatment (NLO=-2.9870), and lastly by the direct treatment (NLO=-0.3273). This illustrates that the indirect treatment has an intermediate effect that falls between no oil exposure and direct oil exposure. p (outlined in orange) tells us whether these results are significantly different from zero. In our model, an odd of zero is meaningless, so this particular p value is meaningless as well. This value tells us that the odds are significant for the control and indirect treatments only (p<0.05 for both), and not the direct treatment (p=0.6163). Given this, it seems intuitive that the p values are meaningless, because the odds of death are the highest in the direct treatment, so one would expect that if any odds were to be significant, it would be with the direct treatment. The mean (outlined in purple) tells us P (P = the number of deaths / the number of trials); these values will fall between zero and one. From this we can see that the control treatment has the fewest deaths/trials (P=0.01574), followed by the indirect treatment (P=0.04802), then followed by the direct treatment (P=0.4189). Again, we see an intermediate effect of the indirect treatment.



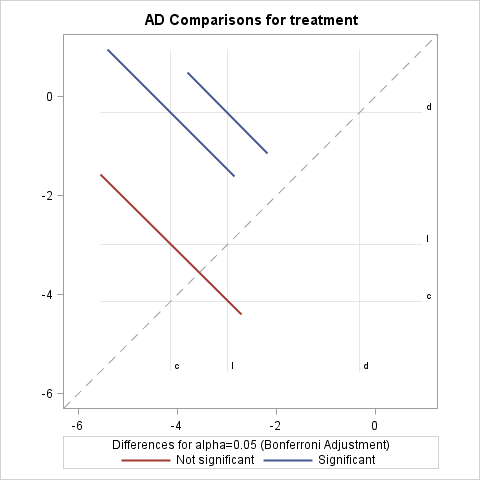
Here we can see a comparison of the treatments. The estimate (outlined in red) is the natural log odds ratio (e.g. ln(odds when x=indirect treatment) – ln(odds when x=direct treatment)); these values will fall between negative infinity and positive infinity. This looks at the differences in the log odds, which is similar to looking at differences in means. When comparing indirect and control treatments, we see the the log odds ratio (LOR) is 1.1485; this indicates that odds of death are greater in the indirect treatment than in the control. The odds of death are smaller in the indirect treatment than in the direct (LOR=-2.6597). The odds of death are smaller in the control treatment than in the direct (LOR=-3.8082), and this difference is larger than the difference between the indirect and direct treatments, which again shows that the indirect treatment has an intermediate effect. The odds ratio (OR) (outlined in purple) tells us the “odds of death with treatment A / odds of death with treatment B;” these values will fall between zero and positive infinity. Again we see that the odds of death are greater in the indirect treatment than in the control (OR=3.153). The odds of death are less in the indirect treatment than in the direct (OR=0.070). The odds of death are less in the control than in the direct (OR=0.022); again we see that the disparity is greatest between the control and direct treatments, suggesting that the indirect treatment has an intermediate effect. Most importantly, when we look at the p value (outlined in orange), we can see whether these differences in log odds are statistically significant. The only statistically significant differences are those between the indirect and direct treatments, and between the control and direct treatments (p<0.05). The difference between the indirect and control treatments are not statistically significant (p=0.3303). This answers our central biological question: indirect exposure to oil *does not* have a significantly (p=0.3303) greater effect on cricket mortality than no exposure to oil (control).

**Describe Statistics**

*Methods:* A generalized linear mixed model with mortality as the response variable, treatment (control, direct, or indirect) as the fixed factor, and plot as a random factor was performed using SAS PROC GLIMMIX (version 9.3).

*Results:* The generalized linear mixed model indicates that the only statistically significant differences in the natural log odds ratios are those between the indirect and direct treatments, and between the control and direct treatments (p<0.05). The difference between the indirect and control treatments are not statistically significant (p=0.3303). This answers our central biological question: indirect exposure to oil *does not* have a significantly (p=0.3303) greater effect on cricket mortality than no exposure to oil (control). The model also reveals that plot co-varies significantly with treatment. This indicates that there was likely some difference in treatment (due to human error) that was not planned for, as each plot was supposed to be the same.

**Visual Display**



This plot shows whether the odds of death are significantly different between treatments. The axes are the natural log odds (a transformation of “how likely death is / how likely life is”) for each treatment; this is replicated on the x- and y-axis. The dotted diagonal line represents where the natural log odds are the same on both axes (and thus the same for the treatments being compared). The plot is a visual representation of how different the treatments’ natural log odds are. In other words, it compares the natural log odds of each treatment to see if they are significantly different. If a point falls along the dotted diagonal line, the natural log odds are not significantly different. The center of each solid line represents the natural log odds comparison, and the solid line itself represents the 95% confidence interval around the natural log odds. The further the center of each solid line is away from the dotted diagonal line, the more significantly different the treatments being compared are. Here again we can see that indirect exposure to oil *does not* have a significantly greater effect on cricket mortality than no exposure to oil (control) (this is seen as the red solid line). The only statistically significant differences are those between the indirect and direct treatments (this is seen as the top blue solid line), and between the control and direct treatments (this is seen as the middle, blue solid line).

References:

Bolker, B. M., et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution." Trends in Ecology & Evolution **24**(3): 127-135.

Bolker, B. M., et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution." Trends in Ecology & Evolution **24**(3): 127-135.