Solution: Discussion # 6, Family and Viral Load Effects on Shrimp Survival

Summary

Method of analysis

The best method of analysis is logistic regression, using whether a shrimp survived or died as the response variable, and its family and viral load as explanatory variables. It would not be appropriate to assess the effect of viral load on survival ignoring family (i.e. lumping together all the families); analysis ignoring the grouping of observations into families would violate the assumption of independence. In contrast, it would be acceptable to compare families ignoring viral load, since the viral loads were naturally occurring (not manipulated), but it would be more informative to include both explanatory variables and their interaction in one analysis. Answering the third question posed in the assignment—does the effect of viral load vary among families—requires testing the interaction.

Because for several families there is complete separation between the ranges of viral loads of shrimp which died and of shrimp which survived, the Firth penalized-likelihood method must be used when running the full model containing the family x viral-load interaction. This is not necessary for models without the interaction.

Because the very skewed distributions of viral load likely would lead to some observations (those with the very high viral loads) having very high leverage, transformation of the viral load variable is advisable. I felt log transformation worked adequately, but other transformations would be reasonable.

Conclusions

There is little evidence of an interaction of family with viral load, i.e. that the effect of viral load differed among families (likelihood ratio test: $X^2 = 10.95$, $df = 11$, $P = 0.4476$).

In an additive model, there is strong evidence for an effect of viral load on survival (LRT $X^2 = 64.577$, $df = 1$, $P < 0.0001$). The conclusion concerning differences among families is less clear: LRT $X^2 = 18.674$, $df = 11$, $P = 0.0672$, so while the test is not significant at the conventional $\alpha=0.05$, it is nearly so and therefore does not support actively concluding that there is not a family (i.e. genetic) effect. [For Dustin’s actual data, the family effect was stronger and the test gave a $P$ value of 0.02. The interaction also was nearly significant using his data: $P = 0.06$.]

Data Exploration

The distributions of viral load within the families all are strongly skewed (long right tails), as shown in the left panel of the graphs at the top of the next page. Log or similar transformation produces much more symmetric distributions with fewer extreme—and thus high leverage—observations. All further analyses shown here will use log-transformed viral loads.

The scatterplot on the next page shows the relationship of shrimp mortality (= 1 on the vertical axis) to log-viral load and family. (This plot was produced in JMP’s Graph Builder, with the smoother stiffened substantially from the default level.) The probability of mortality appears to increase with increasing viral load for all families. The rate of this increase (the slope) appears generally similar for most families, but the slope is considerably steeper for family 7 and less steep for families 1 and 5. Families also appear to differ somewhat in the probability of mortality (as suggested by the smoother in the plot) for a given viral load. One very notable observation
(#42) is a shrimp in family 5 that survived despite having one of the highest viral loads in the entire study.
Analysis

Full model

Analysis

The tests of the terms in the full model, with the family x viral load interaction, are shown to the right. This analysis used the Firth penalized likelihood method in order to deal with the complete separation shown in several of the families (families 7 and 9–12), in which the highest viral load of a surviving shrimp was lower than the lowest viral load of a dead shrimp.

This model fits different slopes for the different families, as shown in the plot below. The test indicates there is no evidence of an interaction: while there are differences in estimated slopes, similar to the differences in the smoothers in the plot on the previous page, the analysis indicates that these differences are no greater than would be expected due simply to the random variability in the data.

Model diagnostics

A plot of residuals vs. fits (next page) shows two exceptional points: observations 42 and 6. Observation 42, as noted above, was a shrimp in family 5 which survived despite having a very
high viral load; the much lower slope for that family was due primarily to this one observation. Observation 6 was a shrimp in family 1 which had a fairly high viral load and which died; it has a large residual, especially relative to its estimated probability of mortality, because most individuals in this family survived and the one other individual that died had a fairly low viral load, so that the model estimates the probability of mortality in this family to be fairly low and to not increase strongly with viral load.

This exploration of the residuals suggests that even what weak suggestion there is of an interaction is largely due to only a few unusual observations.

The goodness of fit tests for this model are quite non-significant.

I see no reason to doubt the validity of the results from this model.

**Conclusion**

The effect of viral load on survival—the rate of increase in mortality with increasing viral load—does not appear to differ among the families.
Additive model

Analysis

The tests of the terms in the additive model produced by removing the family x viral load interaction are shown at the right. This analysis used the regular maximum likelihood method rather than Firth’s penalized likelihood method, since when all families are fit with the same slope the complete separation no longer exists.

This model fits a single slope (in the ‘linear predictor’) for all families, but different intercepts for the families; the different intercepts shift the sigmoid curve for the predicted probabilities left or right. The estimated model is roughly as shown in the plot below (this plot uses a smoother, rather than the fitted function, to interpolate between the estimates for the observations, which is why it is not exactly correct, with some crossing of lines).

The analysis indicates that there is compelling evidence of an effect of viral load on survival: $\chi^2 = 64.577$, df = 1, $P < 0.0001$. The conclusion concerning differences among families is less clear: $\chi^2 = 18.674$, df = 11, $P = 0.0672$. This indicates that the differences among the families in the plot below are small enough that they would not be very surprising as results purely of random variation under the null hypothesis, but this somewhat low $P$-value is suggestive that there might truly be differences among the families.
The estimate for the slope in the linear predictor is $-1.4163$ (see the estimates below). This implies that the odds of survival decrease by a factor of $\exp(-1.4163) = 0.2426$ for each 1-unit increase in log(viral load), with a 95% CI for this odds ratio of $(0.133, 0.3835)$. (This CI is from JMP, which uses the likelihood function to calculate the CI; using the standard error of the slope, as Minitab does, results in a slightly different CI.)

The parameter estimates for the families in the table above all are relative to family 9. They show that the most different families are 2, 7 and 12; 2 and 7 had high mortality at low viral loads while 12 had low mortality even at moderately high viral loads. In other words, family 12 had high tolerance for the virus, while families 2 and 7 had low tolerance.

Model Diagnostics

The plot of residuals vs. fits for the additive model (next page) shows fewer aberrant observations than did the similar plot for the interaction model. In particular, although observation 42 still has a large residual, it is more in keeping with the pattern of the other observations. The goodness of fit tests again are not significant.

I again see no reason to worry about the validity of this analysis.
Conclusions II and III

There is clear evidence that viral load does affect survival, and more specifically that it decreases survival. Quantification of this effect is given above in terms of the odds ratio.

There arguably is a suggestion that survival at any given viral load varied among families, but the differences were not statistically significant at the conventional $\alpha = 0.05$ level.

Conclusions

Does viral load affect survival?

Yes, very clearly and strongly.

Does survival differ among families?

After accounting for differences in viral load, the answer to this question is uncertain, given the small-ish $P$-value of the test. As stated above, there is arguably a suggestion of differences, but not what would be considered persuasive evidence.

An alternative way to answer this question would be to ignore viral load. This would be legitimate since viral load was not manipulated. Differences in overall survival among the families therefore represent how these families fare in the conditions of this experiment. Interestingly, a simple contingency table analysis comparing survival rates across families gives a very large $P$-value. This is because the families that were most sensitive to the virus (had lowest survival at a given viral load) tended to have lower viral loads. Overall survival therefore varied less among families than did survival at a given viral load.

Does the effect of viral load on survival differ among families?

There is no evidence of this.