Stat 505 - Environmental Statistics, Fall 2013 - HW 2

Due: Friday, November 1, by 5pm.

Remember, you are to do 3 of the following 4 problems. You can choose. My intent was to write 1 easier theory problem, 1 easier data analysis problem, 1 harder theory problem, and 1 harder data analysis problem. Your perception may differ; even so, choose 3 problems. If you do all 4, I'll grade them all and drop the lowest. Further information and suggestions for organizing answers to open ended data analysis problems (like problem 4) are on the 'homework information' part of the class web site.

I believe you can do everything in R, but a combination of SAS and R may get answers more easily.

1. The data in deguelin.txt are from a study of the effect of the toxicity of the "organic" insecticide deguelin on the Chrysanthemum aphid, *Macrosiphoniella sanborni*. A photo:

   `http://culturesheet.org/_media/photographs:pests:chrysanthemum_aphid.jpg`

   The toxicity study included 6 doses from 5.12 to 50.12 (units unknown). Between 48 and 50 aphids were exposed to each dose; the response is the number of dead aphids.

   (a) Consider a linear logistic dose response model for these data. The model is: logit $\pi_i = \beta_0 + \beta_1 D_i$. Estimate the intercept, slope and LD_{50}. Calculate the standard errors for each.

   (b) Historically, there was a major disagreement between advocates of the logit model and advocates of the probit model. The probit model is defined as

   $$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 D_i,$$

   where $\Phi^{-1}(\pi_i)$ is the inverse cumulative distribution function for a standard normal distribution. Use the probit model to estimate the intercept, slope, and LD_{50}.

   (c) Consider a total of eight alternative models by combining all combinations of (logit or probit link function) $\times$ (dose or log dose) $\times$ (no control mortality or non-zero control mortality). The models in parts a and b are two of these 8. Which model is the most reasonable choice for these data? Support your choice.

   (d) Using the most appropriate model from part b, construct a reasonable 95% confidence interval for the LD_{50}.

   (e) Using the most appropriate model from part c, estimate the Benchmark Dose for a Benchmark Risk = 5% and the lower 95% confidence bound. Because of the substantial apparent control mortality, the Benchmark Risk is defined in terms of the excess risk, that is 5% of the risk that is attributable to deguelin.

   (f) Find the NOEL (No observed effect level) and LOEL (Lowest observed adverse effect level). For the purpose of this part (and this part only), assume that the mortality for a dose of 0 is 14 aphids out of 48 tested.

2. Two related theory questions.

(a) In class, I said that Fieller's method gives a $1-\alpha$ coverage confidence interval for the LC_p by considering the pivotal quantity $\theta = \log\left(\frac{p}{1-p}\right) - \beta_0 - \beta_1 D$ and inverting tests of $\theta = 0$. If you assume that the estimates have a bivariate normal distribution, this means computing

$$z = \frac{\log\left(\frac{p}{1-p}\right) - \hat{\beta}_0 - \hat{\beta}_1 D}{\sqrt{\text{Var }(\hat{\beta}_0 + \hat{\beta}_1 D)}}$$

and finding the values of D that reject H0: $\log\left(\frac{p}{1-p}\right) - \beta_0 - \beta_1 D = 0$ at exactly $p = 1 - \alpha/2$. This part explores some issues with that confidence interval.

i. When the logistic, linear dose, model: logit $\pi_i = \beta_0 + \beta_1 X_i$, is fit a particular data set, the estimates are $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 0.25$ with an estimated variance-covariance matrix with elements: Var $\beta_0 = 0.2$, Var $\beta_1 = 0.005$ and Cov β_0, β_1 = -0.0005. Calculate a 95% confidence interval for the LC_{50}.

ii. Now imagine a different data set for which the estimates are exactly the same, $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 0.25$, but the variance-covariance matrix has elements: Var $\beta_0 = 4$, Var $\beta_1 = 1$ and Cov β_0, β_1 = -1.2. Calculate a 95% confidence interval for this LC_{50}.

iii. To understand what is happening in part ii, calculate the pivotal quantity, θ, and its standard error, $\sqrt{\text{Var }\theta}$, across a range of doses, e.g. 0 to 50. Use this information to explain why the confidence interval in part ii is the way it is.

(b) In class, our discussion has focused on the linear logistic dose-response model, logit $\pi_i = \beta_0 + \beta_1 D_i$. In some combinations of species and chemicals, there appears to be a threshold below which lower amounts of the chemical have similar effects,

$$\text{logit } \pi_i = \begin{cases} \beta_0 + \beta_1\theta & D_i \leq \theta \\ \beta_0 + \beta_1 D_i & D_i > \theta \end{cases}$$

The LC_{50} is still the concentration at which P[event] = 0.5. This is one difference between the threshold model and the 'background mortality' model discussed in class, in which the LC_{50} is usually defined in terms of excess risk, ignoring the background mortality. All three parameters (β_0, β_1, and θ) and their asymptotic variance-covariance matrix $\begin{bmatrix} V_0 & C & C_{0t} \\ C & V_1 & C_{1t} \\ C_{0t} & C_1 t & V_t \end{bmatrix}$ can be estimated by maximum likelihood and the inverse observed information.

i. For the threshold model, derive the LC50, in terms of the model pararameters, (β_0, β_1, and θ).

ii. Does the addition of a threshold parameter to the model change how you would derive a confidence interval for the LC_{50}? In other words, could you use the pivotal quantity, θ, defined at the very beginning of the problem to construct a confidence interval for LC_{50} from estimates from the threshold model? Explain why or why not.

3. In class, I said that overdispersion models can be interpreted as a consequence of correlation between individuals in a litter. This is explored in this problem. Consider a beta-binomial model for responses Y_{ij} of individuals j in litter i. For this problem, there are no treatment effects (i.e. consider data from a single dose). The model is:

$$\begin{aligned} Y_{ij} \mid p_i &\sim \text{independent Bernoulli}(p_i) \\ p_i &\sim \beta(\alpha,\, \beta) \end{aligned}$$

A Bernoulli distribution is a Binomial distribution with 1 event, i.e. Bernoulli(p_i) is the same as Binomial(1, p_i).

(a) Derive the mean and variance of Y_{ij}.
Note: I am interested in the marginal distribution of Y_{ij}, not the conditional distribution of $Y_{ij} \mid p_i$. It may help to know that if $p_i \sim \beta(\alpha,\, beta)$, E $p_i = \frac{\alpha}{\alpha+\beta}$ and Var $p_i = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}$.

(b) Derive the covariance of Y_{ij} and Y_{ik}, that is the covariance of responses from two individuals in the same litter.
Hint: The conditional variance formula used a couple of times in class has a conditional covariance extension:
Cov $(Y_{ij},\, Y_{ik})$ = Cov (E $Y_{ij} \mid p_i$, E $Y_{ik} \mid p_i$) + E Cov $(Y_{ij},\, Y_{ik} \mid p_i)$

(c) Derive the correlation between responses from two individuals in the same litter

(d) Derive the correlation between responses from two individuals in different litters.

4. The data in chlorpyrifos.txt are from a sediment toxicity test measuring the effect of the pesticide chlorpyrifos on the harpacticoid copepod *Amphiascus tenuiremis*. Sediment tends to accumulate chemicals present in the overlying water because many chemicals bind to clay particles. *Amphiascus* is a benthic invertebrate commonly used in sediment toxicity tests because they live in the top 1-2 cm of the sediment (where chemicals from the water tend to accumulate) and they are an important food source for fish and other aquatic life. A picture of a closely related critter is at

`http://entomology.tfrec.wsu.edu/pearent/images/misc/Crustacea/harpact.jpg`

These copepods have life stages, rather like frogs. Frogs have 2 stages, tadpoles and adults that are very different. Copepods have more, but I'm focusing on two: nauplii and adults. The life stages of copepods differ in their sensitivity to chemicals. I am providing data from nauplii (stage = N), which are very young individuals, and adults (stage = A). These tests are done by putting 25 adults or 20 nauplii in a small beaker, adding sediment with different concentrations of the contaminant, and recording mortality after 96 hours. There are replicate beakers at each dose. The doses used for adults are not the same as those used for the nauplii. I don't know why, but my guess is because it is hard to dissolve most pesticides in water, so it is difficult to prepare a solution at a specific concentration. Once a solution is prepared, the concentration of chlorpyrifos can be measured accurately. The data file has one row for each beaker, with the stage, the dose, the number of live animals (count) and the number of individuals put in that beaker (n). Note that the % mortality is 100(1-count/n).

The investigators want to know:
For the adults: The LC_{50} and a 95% confidence interval
For the adults: The Benchmark Dose for a 5% excess risk (i.e. 5% mortality above background), using a 95% confidence bound.
Whether the LC_{50}'s for the adults and nauplii are the same. They really need a formal test (e.g. Wald or likelihood ratio test). Comparing confidence intervals is not sufficient.