COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 18, 2006, 9:00 A.M.–1:00 P.M.

STATISTICS 175 QUESTIONS

1. **Island Size and Bird Extinctions**

   Scientists agree that preserving certain habitats in their natural states is necessary to slow the accelerating rate of species extinctions. Given a finite amount of available land, is it better to have many small reserves or a few large ones?

   To help answering the question, researchers presented results of extensive bird surveys of the Krunnit Islands taken over four decades. They visited each island several times, cataloging species. If a species was found on a specific island in 1949, it was considered to be at risk of extinction for the next survey of the island in 1959. If it was not found in 1959, it was counted as an "extinction". The data (**island**) on island size (**area**), number of species at risk of extinction (**atrisk**), and number of extinctions (**extinct**) are shown below. (Data from Vaisanen and Jarvinen, "Dynamics of Protected Bird Communities in a Finnish Archipelago," *Journal of Animal Ecology* 46 (1977): 891-908.)

```
> island
     area atrisk extinc
1  185.80     75      5
2  105.80     67      3
3   30.70     66     10
4    8.50     51      6
5    4.80     28      3
6    4.50     20      4
7    4.30     43      8
8    3.60     31      3
9    2.60     28      5
10   1.70     32      6
11   1.20     30      8
12   0.70     20      2
13   0.70     31      9
14   0.60     16      5
15   0.40     15      7
16   0.30     33      8
17   0.20     40     13
18   0.07      6      3
```

Exploratory data analysis suggests that there is a linear relationship between the logit of extinction and the **log(area)**. The following generalized linear model is fitted in R.

```
> is.glm <- glm(cbind(extinc, atrisk-extinc)~log(area), family=binomial)
> summary(is.glm)

Call:
glm(formula = cbind(extinc, atrisk - extinc) ~ log(area), family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.71726  -0.67722   0.09726   0.48365   1.49545

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
log(area)   -0.29710    0.05485  -5.416 6.08e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.338  on 17  degrees of freedom
Residual deviance: 12.062  on 16  degrees of freedom
AIC: 75.394

Number of Fisher Scoring iterations: 4
```

(a) [**10 points**] Write down the mathematical model and interpret the model parameters of the fitted model. Clearly explain your notation.

(b) [**10 points**] Calculate the change in the odds of extinction resulting from a 50% reduction in island area. Express the change in terms of percentage. Report the 95% confidence interval.

(c) [**10 points**] Perform a deviance goodness-of-fit test to investigate whether the model is adequate. Use a deviance related test to see whether **log(area)** is significant. Which test is more reliable and why?

(d) [**5 points**] Write down the estimated equation for the logit of survival and explain why. If you think there is not enough information, state what extra information you would need.

(e) **[5 points]** List what possible problems exist for the observational study, and briefly justify.

2. **Learning Curve**

Consider two groups of workers manufacturing a particular product. Learning curve models try to characterize the reduction in manufacturing times as workers gain experience, i.e. manufacture more units of the product.

Define Z_{ij} as the manufacturing time for the jth unit by the ith worker, and $S_i = 1$ if the ith worker belongs to the experienced group, $S_i = 0$ if she belongs to the naive group, where $i = 1, \ldots, n; j = 1, \ldots, m$.

One traditional learning curve model is the following

$$Y_{ij} = \log(Z_{ij}) = \alpha_i + \log(j)\beta_i + \epsilon_{ij},$$

where the model parameter α_i is the initial manufacturing time and the parameter β_i is the learning rate of the ith worker, $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

(a) **[10 points]** Suppose one decides to model both α_i and β_i as independent random effects with means depending on the worker group, denoted as α_E, α_N, β_E and β_N, and the same variances as σ^2_α and σ^2_β.

Write down the mathematical model using the above notations.

(b) **[10 points]** Let $S = (S_1, \ldots, S_n)'$, $L = (\log 1, \ldots, \log m)'$,
$Y = (Y_{11}, \ldots, Y_{1m}, Y_{21}, \ldots, Y_{2m}, \ldots, Y_{n1}, \ldots, Y_{nm})'$
and
$\epsilon = (\epsilon_{11}, \ldots, \epsilon_{1m}, \epsilon_{21}, \ldots, \epsilon_{2m}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nm})'$

Express the above model using matrix expression.

(c) **[10 points]** Calculate the variance-covariance matrix of Y, denoted as V.

(d) **[10 points]** Now suppose one decides not to distinguish between the two groups of workers, and also treat their learning rates as fixed effects.

Write down the new model in matrix form, and calculate $var(Y)$ and its inverse.

(e) **[20 points]** Derive the BLUP of the initial manufacturing times of the workers. (*Carry out the calculation as far as you can.*)

2006 CWE STATISTICS 175 SOLUTIONS

1. **Island Size and Bird Extinctions**

   (a) Let y_i denote the number of extinction of island i and x_i is the area of island i, assume $y_i \sim Binomial(m_i, \pi_i)$ where m_i is the number at risk of island i and π_i is the probability of extinction. Then the logistic model is

   $$logit(\pi_i) = \beta_0 + \beta_1 \log(x_i).$$

   (b) Each halving of island area leads to a multiplicative change in the odds of extinction of $0.5^{-0.297}$, or 1.23, which means a 23% increase. The 95% confidence interval for β_1 is $[-0.403, -0.191]$. Thus, the confidence interval for the change is $[0.5^{-0.403} - 1, 0.5^{-0.191} - 1]$, i.e. $[13.6\%, 31.5\%]$.

   (c) The residual deviance approximately has a Chi-squared distribution with a degree of freedom of 16, which suggests a p-value of 0.74. This means that there is no evidence that the model is inadequate.

   The deviance reduction is 45.338-12.062=33.276, which should have an approximate Chi-squared distribution with a d.f. of 1, and the corresponding p-value is really small. This suggests that the negative linear relationship is very significant.

   The Chi-squared approximation of the deviance reduction is more accurate.

   (d) $logit(\hat{\pi}) = 1.196 + 0.297 log(area)$.

   (e) Possible problems include

   - Each species found on an island is assumed to have the same chance of becoming extinct.
   - Extinctions are assumed to occur independently.
   - A species found extinct may come back in a later year.

2. **Learning Curve**

   (a) The model is

   $$Y_{ij} = \alpha_E S_i + \alpha_N(1 - S_i) + S_i \log(j)\beta_E + (1 - S_i)\log(j)\beta_N + \alpha_i + \log(j)\beta_i + \epsilon_{ij},$$

   where $\alpha_i \overset{i.i.d.}{\sim} N(0, \sigma_\alpha^2)$, $\beta_i \overset{i.i.d.}{\sim} N(0, \sigma_\beta^2)$, and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

   (b) The expression is

   $$\begin{aligned} Y &= (I_n \bigotimes 1_m) S\alpha_E + (I_n \bigotimes 1_m)(1_n - S)\alpha_N \\ &+ (I_n \bigotimes L) S\beta_E + (I_n \bigotimes L)(1_n - S)\beta_N \\ &+ (I_n \bigotimes 1_m)\alpha + (I_n \bigotimes L)\beta + \epsilon. \end{aligned}$$

(c) The variance-covariance matrix is

$$\begin{aligned} V &= (I_n \bigotimes J_m)\sigma_\alpha^2 + (I_n \bigotimes LL')\sigma_\beta^2 + I_n \bigotimes I_m\sigma^2 \\ &= I_n \bigotimes (J_m\sigma_\alpha^2 + LL'\sigma_\beta^2 + I_m\sigma^2). \end{aligned}$$

(d) Let μ denote the mean of the random effects, and α_i be the zero-mean random effects. Then the model becomes

$$Y = 1_{nm}\mu + (I_n \bigotimes 1_m)\alpha + (I_n \bigotimes L)\beta + \epsilon.$$

The variance matrix is

$$V = I_n \bigotimes (J_m\sigma_\alpha^2 + I_m\sigma^2).$$

Its inverse is

$$V^{-1} = \frac{1}{\sigma^2} I_n \bigotimes (I_m - \frac{\sigma_\alpha^2}{\sigma^2 + m\sigma_\alpha^2} J_m).$$

(e) Basically, one needs to find the BLUP of $\mu + \alpha$.

Rewrite the model as

$$\begin{aligned} Y &= 1_{nm}\mu + (I_n \bigotimes 1_m)\alpha + (I_n \bigotimes L)\beta + \epsilon \\ &= [1_{nm}, I_n \bigotimes L](\mu, \beta)' + (I_n \bigotimes 1_m)\alpha + \epsilon \\ &\equiv Xu + (I_n \bigotimes 1_m)\alpha + \epsilon. \end{aligned}$$

We have the following result,

$$BLUP(\alpha) = CV^{-1}(Y - Xu^0),$$

where

$$C = cov(\alpha, Y) = I_n \bigotimes 1_m'\sigma_\alpha^2,$$

and

$$Xu^0 = X(X'V^{-1}X)^{-1}X'V^{-1}Y.$$

Then,

$$CV^{-1} = \frac{\sigma_\alpha^2}{\sigma^2 + m\sigma_\alpha^2} I_n \bigotimes 1_m',$$

and

$$BLUP(\mu + \alpha) = \mu^0 + BLUP(\alpha).$$