Solutions to Methods in Fall 2006, Aug 10 2006, 8:00-12:00

1. (a) The loglikelihood function is

$$\begin{aligned}\ell(\beta,\lambda) =& \sum_{i=1}^{5}\log\{[(\lambda e^{-\lambda Y_i})^{\delta_i}(e^{-\lambda Y_i})^{1-\delta_i}]^{1-Z_i}[(\beta\lambda e^{-\beta\lambda Y_i})^{\delta_i}(e^{-\beta\lambda Y_i})^{1-\delta_i}]^{Z_i}\}\\ =& \sum_{i=1}^{5}\log\{\beta^{\delta_i Z_i}\lambda^{\delta_i}e^{-\lambda Y_i[(1-Z_i)+\beta Z_i]}\}\\ =& \log(\beta)\sum_{i=1}^{5}\delta_i Z_i + \log(\lambda)\sum_{i=1}^{5}\delta_i - \lambda\sum_{i=1}^{5}Y_i(1-Z_i+\beta Z_i).\end{aligned}$$

Then, we have

$$\begin{aligned}\frac{\partial\ell}{\partial\beta} =& \frac{1}{\beta}\sum_{i=1}^{5}\delta_i Z_i - \lambda\sum_{i=1}^{5}Y_i Z_i\\ \frac{\partial\ell}{\partial\lambda} =& \frac{1}{\lambda}\sum_{i=1}^{5}\delta_i - \sum_{i=1}^{5}Y_i(1-Z_i) - \beta\sum_{i=1}^{5}Y_i\ Z_i.\end{aligned}$$

Then, we have

$$\begin{aligned}\hat{\beta} =& \frac{1}{\hat{\lambda}}\frac{\sum_{i=1}^{5}\delta_i Z_i}{\sum_{i=1}^{5}Y_i Z_i}\\ \hat{\lambda} =& \frac{\sum_{i=1}^{5}\delta_i(1-Z_i)}{\sum_{i=1}^{5}Y_i(1-Z_i)}\end{aligned}$$

Put the observations in. We have $\hat{\lambda} = 0.0952$ and $\hat{\beta} = 0.8754$.

(b) The partial likelihood is the product of the conditional probability of individuals with covariate Z_i fails at Y_i on some subject failed at time Y_i. In fact, it can be expressed as the product of

$$L_i = \frac{\sum_{\text{death at }t_i} h(t_i|x_j)}{\sum_{\text{at risk at }t_i} h(t_i|x_j)}.$$

We should calculate 3 terms separately since there are 3 failures and then calculate their product. The steps is

$$L_1 = \frac{h(3;z=1)}{\sum_{i=1}^{5}\beta^{1-Z_i}h(3;z=1)} = \frac{\beta}{(3+2\beta)}.$$

Similarly we have $L_3 = 1/(2+\beta)$ and $L_5 = 1$. Thus, their product is

$$L = \frac{\beta}{(3+2\beta)(2+\beta)}.$$

By maximizing L, we have $\hat{\beta} = \sqrt{3} = 1.732$.

(c) First, we need to get the variance of $\hat{\beta}$: for part (a), we can use the Fisher Information; for part (b), we can also use the expected value of the negative of the second order derivative. Then, we can construct a z confidence interval.

2. (a) The regression lines for supplier A is

$$\begin{aligned} Y =& 37.19367 - 3.83362 - 0.07452time + 0.00622time \\ =& 33.36005 - 0.0683time, \end{aligned}$$

for supplier B is

$$\begin{aligned} Y =& 37.19367 - 1.98755 - 0.07452time + 0.01823time \\ =& 35.20612 - 0.05629time \end{aligned}$$

and for supplier C is

$$Y = 37.19367 - 0.07452time.$$

(b) The remaining for supplier A is $33.36005 - 0.0683(250) = 16.28505$ and for supplier C is $37.19367 - 0.07452(250) = 18.56367$. The estimated variance of their difference is

$$(1, 250) \begin{pmatrix} 3.7369 & -0.02598 \\ -0.02598 & 0.0002152 \end{pmatrix} \begin{pmatrix} 1 \\ 250 \end{pmatrix} = 4.1969.$$

The t-value is

$$\frac{18.56367 - 16.28505}{\sqrt{4.1969}} = 1.112.$$

Thus, there are not significantly different.

(c) We need to calculate the MSE, and the residual degrees of freedom. Note that

$$R^2 = \frac{SSR}{SST} = 0.8688 \Rightarrow SST = \frac{SSR}{0.8688} = \frac{936.536}{0.8688} = 1077.965.$$

Therefore, for model 3, we have

$$MSE = \frac{1077.965 - 1018.656}{23} = 2.58.$$

We can use an F test for their difference as

$$F = \frac{(1018.656 - 936.536)/2}{2.58} = 15.91$$

indicating the three regression lines are significantly different.

3. (a) The correlation matrix is

$$R = R(Y, Y) = \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

(b) If ρ is known, we can get $R^{-1/2}$. Let $Y' = R^{-1/2}Y$. and $X = R^{-1/2}X$, where X is the design matrix as

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}.$$

Let $\beta = (\beta_0, \beta_1)^t$. Then, we have

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon},$$

where $\tilde{\epsilon}$ is identically independently distributed with mean 0 and common variance σ^2. This gives

$$\hat{\beta} = (\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t\tilde{Y} = (X^tR^{-1}X)^{-1}X^tR^{-1}Y$$

as the minimum variance unbiased estimators.

(c) We prefer the second one since the first one does not consider the correlation but the second one does.

(d) First, let us look at the covariance in the first design. We have

$$\begin{aligned} & Cov(\bar{Y}_1, \bar{Y}_2) \\ =& \frac{1}{n^2}(Cov(Y_1, Y_2) + Cov(Y_3, Y_2) + Cov(Y_3, Y_4) + \cdots + Cov(Y_{n-1}, Y_n)) \\ =& \frac{(2n-2)\rho\sigma^2}{n^2}, \end{aligned}$$

and

$$V(\bar{Y}_1) = V(\bar{Y}_2) = \frac{1}{n}\sigma^2.$$

Then, we have

$$V(\bar{Y}_1 - \bar{Y}_2) = \frac{2\sigma^2}{n} - \frac{4(n-1)\rho\sigma^2}{n^2} = \frac{\sigma^2}{n^2}[2n - 4(n-1)\rho].$$

In the second design, we have

$$Cov(\bar{Y}_1, \bar{Y}_2) = \frac{1}{n^2}Cov(Y_{n/2}, Y_{n/2+1}) = \rho\sigma^2/n^2$$

and

$$V(\bar{Y}_1) = V(\bar{Y}_2) = \frac{\sigma^2}{n^2}[n + 2(n-1)\rho].$$

Thus,

$$V(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma^2}{n^2}[2n + 4(n-1)\rho] + \frac{2\rho\sigma^2}{n^2}.$$

Then, if $\rho > 0$ the first design leads a smaller variance and if $\rho < 0$ the second leads a smaller variance.

4. (a) Since sex only has two levels, it does not matter whether it is a linear term or another form of term.

   (b) The p-value of $log(income)$ is very small. Thus it has the least linear form. This p-value tells the significant of the nonlinearity.

   (c) Residual degree of freedom is $df = 47 - (3 + 1.797 + 3.748) = 38.355$.

   (d) The gamma-GLM is more appropriate.

   (e) Male with $income > 3.75$.

   (f) The output tree only changes by changing 3.75 by log(3.75) as the cut-point value.

5. (a) The intercept only increases from 5.979 to 6.244 and the t-value of the increases is about $(6.244 - 5.979)/1.5 = 0.17$ indicating it is very insignificant. Thus it is reasonable to combine these two categories. There is another reason. Usually, if counts are low in two categories and these two categories are next to each ordder in ordinal response classification, we like to combine them so that the result is more stable.

   (b) It is more like to be musical since the t-value is small and the baseline is musical revue.

   (c) Revival is not significant. Thus the revival "yes" does not significantly last longer than revival "no". Exactly, its sign is negative and so it is shorter.

   (d) Suppose we still fit the main effect model. Then the response has 5 categories. Thus, there are 4 estimated equations. Each of then has 5 parameters (one for intercept, 2 for type, 1 for revival and 1 for week1). Thus, there are $4 \times 5 = 20$ parameters.

   (e) The logistic regression line is

$$\log \frac{p}{1-p} = 3.618 - (0.340005 - 0.718951 + 0.049819(0.92)) = 3.951 \Rightarrow p = 0.9811.$$