QUALIFYING EXAM SOLUTIONS
Statistical Methods
Saturday, Aug 11, 2007, 8:00 am -12:00 pm

1. (a) The assumptions are: (i) $E(Y) = \beta_0 + \beta_1 X_j$; and (ii) $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$.

   (b) The complete table is

Source	df	SS	MS
Regression	1	160	160
Error	38	17.60	0.4632
LOF	2	1.0933	0.5467
PE	36	16.6667	0.4930
Total	39	177.60	

   (c) The first model is the simple linear regression model as given before

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

   with $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$. The second is the ANOVA model as given by

$$y_{ij} = \mu_j + \epsilon_{ij}$$

   with $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$.

   (d) If μ_j is a linear function of X_j as $\mu_j = \beta_0 + \beta_1 X_j$, then the second model reduces to the first model. To test the adequacy of the first model, we need to look at the F-statistic

$$F^* = \frac{LOF/df_{LOF}}{PE/df_{PE}} = \frac{0.5467}{0.4930} = 1.1089 < F_{0.05,2,36} = 3.26.$$

   Therefore, the lack of fit is not significant.

2. (a) The statistical model is

$$Y_{ij} = \mu_{ij} + \epsilon_{ijk}$$

   with $\epsilon_{ijj} \sim^{iid} N(0, \sigma^2)$, $i = 1, 2, 3$, $j = 1, 2$ and $k = 1, \cdots, n$, where n is the replicate number. In additional, the model can also be written as an ANOVA model as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

   by a zero sume constraint given by

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{2} \beta_j = \sum_{i=1}^{3} (\alpha\beta)_{ij} + \sum_{j=1}^{2} (\alpha\beta)_{ij} = 0.$$

   In terms of the second model, we have

$$\mu_{11} = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}.$$

(b) Since the sum of the coefficients are 0, we have the contrast $(0.5, 0.5, -1)$ giving

$$\frac{\mu_{11}+\mu_{21}}{2} - \mu_{31} = 0.5\mu_{11} + 0.5\mu_{21} - 1\mu_{31}.$$

(c) Note that

$$\mu_{21} = \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$$

and

$$\mu_{31} = \mu + \alpha_3 + \beta_1 + (\alpha\beta)_{31}.$$

We have

$$\begin{aligned}\frac{\mu_{11}+\mu_{21}}{2} - \mu_{31} =& \mu + \frac{\alpha_1+\alpha_2}{2} + \beta_1 + \frac{(\alpha\beta)_{11}+(\alpha\beta)_{21}}{2} \\ & - \mu - \alpha_3 - \beta_1 - (\alpha\beta)_{31}. \\ =& \frac{\alpha_1+\alpha_2}{2} - \alpha_3 + \frac{(\alpha\beta)_{11}+(\alpha\beta)_{21}}{2} - (\alpha\beta)_{31}.\end{aligned}$$

The last line in SAS should be

```
A*B 0.5 0 0.5 0 -1 0;
```

3. (a) i. Note that Y_i follows a binomial distribution. Let $p_i = p_i(x_i)$ be the probability of a foul to a black player. Then, we have

$$\begin{aligned}P(Y_i = y_i|x_i) =& \binom{K}{y_i} p_i^{y_i}[1-p_i]^{K-y_i} \\ =& \exp\{y_i \log\frac{p_i}{1-p_i} - K\log[1-p_i] + \log\binom{K}{y_i}\}.\end{aligned}$$

Thus, we have

$$\theta_i = \log\frac{p_i}{1-p_i}; \phi = 1; b(\theta_i) = K\log(1+\frac{e^{\theta_i}}{1+e^{\theta_2}}); c(y_i) = \log\binom{K}{y_i}.$$

ii. The canonical link function is the logistic link.

iii. The variance function is

$$V(Y_i) = Kp_i(1-p_i) = K\frac{e^{\theta_i}}{(1+e^{\theta_i})^2}.$$

(b) Note that e^{θ_i} is the odds ratio. The odds increases $e^{-0.5} - 1 = -0.3935$. The odds for a black referee calling a foul to black player is 39.35% lower than calling a foul to white player.

(c) It is possible since we can summarize the data into a 2×2 table as

	Referee	
Play	Black	White
Black	y_1	y_2
While	$n - y_1$	$n - y_2$

where y_1 is the total number of fouls called to black players by black referees, and y_2 is the total number of fouls called to black player by white referees. We can use methods of 2×2 table to analyze the data.

4. (a)

$$\hat{S}_1(t) = \begin{cases} 1, & \text{when } t < 46 \\ (1 - 1/8) = 0.875, & \text{when } 46 \leq t < 78 \\ 0.875(1 - 1/5) = 0.700 & \text{when } t \leq 78 < 124 \\ 0.700(1 - 1/4) = 0.525 & \text{when } t \geq 124 \end{cases}$$

and

$$\hat{S}_2(t) = \begin{cases} 1, & \text{when } t < 9 \\ (1 - 1/9) = 0.889, & \text{when } 9 \leq t < 26 \\ 0.889(1 - 1/8) = 0.778, & \text{when } 26 \leq t < 46 \\ 0.778(1 - 1/6) = 0.648, & \text{when } 46 \leq t < 64 \\ 0.648(1 - 1/5) = 0.519, & \text{when } 64 \leq t < 75 \\ 0.519(1 - 1/4) = 0.389, & \text{when } 75 \leq t < 100 \\ 0.389(1 - 1/3) = 0.259, & \text{when } t \geq 100 \end{cases}$$

(b) The testing method is called the logrank test. The death happened at time equals to 9, 26, 46, 64, 75, 78, 100, 124 with counts 1, 1, 2, 1, 1, 1, 1, 1, and the at risk from both groups as $(8, 9)$, $(8, 8)$, $(8, 6)$, $(6, 5)$, $(5, 4)$, $(5, 3)$, $(4, 3)$ and $(4, 2)$. Then, we have the expected values for group 1 as

$$e_1 = 1(\frac{8}{8+9}) + 1(\frac{8}{8+8}) + 2(\frac{8}{8+6}) + \cdots + 1(\frac{4}{4+2}) = 5.0776$$

and similarly $e_2 = 3.9224$. Thus, the test statistic is

$$L = \frac{(3 - 5.0776)^2}{5.0776} + \frac{(6 - 3.9224)^2}{3.9224} = 1.95 < 3.84$$

Thus, the logrank test accepts the null hypothesis that the two survival functions are the same.

(c) For group 1, the expected survival time is

$$\frac{\sum_{i=1}^{8} t_i}{\sum_{i=1}^{8} \delta_i} = \frac{46 + 46 + 64 + 78 + 124 + 130 + 150 + 150}{3} = 262.54$$

and for group 2, the expected survival time is

$$\frac{9 + 26 + 43 + 46 + 64 + 75 + 100 + 130 + 150}{6} = 107.01.$$

(d) Plot $\log[-\log(\hat{S}(t))]$ versus $\log(t)$, where $\hat{S}(t)$ is the Kaplan-Meier estimators for the two groups. If we find two parallel curves, then the proportional hazard proporty is OK. In addition, we can also use the same plot to check issues that have not been asked by this part. For example, if we find straight lines with 45 degree, then we can say that the survival functions follow exponential distribution. If we find straight lines with other degrees, then we can say they follow Weibull distribution.

5. (a) Let $i, j, k = 1, 2$ indicate A, B C as "yes" or "no" respectively. Then, the, the mutual independence model is

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k,$$

the joint independence model is

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$$

and the conditional independence model is

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}.$$

(b) The estimated variables are below according to (joint independence, conditional independence)

	Drug A "Yes"		Drug A "No"	
	Drug B "Yes"	Drug B "No"	Drug B "Yes"	Drug B "No"
Drug C "Yes"	(29.66,28.50)	(9.89,9.50)	(9.89,10.44)	(17.57,18.56)
Drug C "No"	(51.34,52.50)	(17.11,17.50)	(17.11,16.56)	(30.43,29.44)

(c) The Pearson Chi-square statistic is

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(n_{ijk} - \hat{n}_{ijk})^2}{\hat{n}_{ijk}}$$

and the loglikehood statistic is

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} n_{ijk} \log \frac{n_{ijk}}{\hat{n}_{ijk}}.$$

The loglikelihood statistic is 0.7953 or 0.5646 respectively. The Pearson statistic is 0.8014 or 0.5601 respectively. Both of them claim insignificance of the interaction between B and C. Thus, the joint independence model is preferred.