Solutions to Methods in Fall 2003

1. (a) Let y_i and n_i be the total number of murder and the total number of crimes, x be the year index. Then the random component is y_i are independent from $B(p_i, n_i)$ distribution for $i = 0, 1, \cdots, 41$; the linear compoent is $\eta = \alpha + x\beta$; there are three possible link function we can use, *probit link* by $\eta_i = \Phi^{-1}(p_i)$, *logistic link* bu $\eta_i = \log[p_i/(1-p_i)]$, and *complementary loglog link* by $\eta_i = \log[-\log(p_i)]$.

   (b) The predicted probability of murder in 1960 is

   $$p_0 = \frac{e^{-0.9814}}{1 + e^{-0.9814}} = 0.2726,$$

   and the predicted probability of murder in 2000 is

   $$p_{40} = \frac{e^{-0.9814-40\times 0.0175}}{1 + e^{-0.9814-40\times 0.0175}} = 0.1569.$$

   (c) The odds ratio is

   $$\hat{\theta} = e^{-40\times 0.0175} = 0.4966.$$

   The p-value of the odds ratio is less than 2×10^{-16}.

   (d) The 95% confidence interval of the mean rate in 1960 is

   $$\begin{aligned}&[0.00004413 - 2.0337 \times 0.000003765, 0.00004413 + 2.0227 \times 0.000003765]\\ =&[0.000003647, 0.00005175].\end{aligned}$$

   The variance in 2000 is

   $$\begin{aligned}&0.000003765^2 - 2 \times 40 \times 0.8608 \times 0.000003765 \times 0.0000001581\\ +&40^2 \times 0.0000001581^2 = 0.00000363^2.\end{aligned}$$

   The mean in 2000 is

   $$0.00004413 + 40 \times 0.0000008432 = 0.00007786.$$

   Thus, the 95% confidence interval is

   $$\begin{aligned}&[0.00007786 - 1.96 \times 0.00000363, 0.00007786 + 1.96 \times 0.00000363]\\ =&[0.00007075, 0.00008497].\end{aligned}$$

   (e) From 1960 to 2000, the precentage of murder in a crime decreases from 27.26% to 15.69%, but the rate of murder is still increasing from 0.00004413 to 0.0000786. This happpens because the rate of other types of crimes increases much faster than that of murder.

2. (a) The loglikelihood is

$$L = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^t\beta)^2$$

and

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = (N-d)\sigma^2.$$

Note that the MLE are $\hat{y}_i = x_i^t\hat{\beta}$ and $\hat{\sigma}^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2/N$. We have

$$\begin{aligned} AIC =& N\log(2\pi) + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + N\log(\hat{\sigma}^2) + 2d \\ =& Constant + 2d + N\log(\hat{\sigma}^2) \\ =& Constant + 2d + N\log\frac{RSS}{N}, \end{aligned}$$

where RSS is the residual sum of squares.

(b) (Omitted).

(c) The AIC for the full model is -7.77, and for the reduce models are 13.30, -8.85 and 2.54 respectively. Thus, we choose the second reduce model. It gives RSS equal to 2.764712 and this is the σ^2 the full model be preferred.

3. (a) The design matrix, an 80×4 matrxi is

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 40 & 1 & 40 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 40 & 0 & 0 \end{pmatrix}$$

Let $\beta = (\mu, \alpha, \gamma, (\alpha\gamma))$. The constraint is the values of Where and interaction are 0 if it is from New Nexico. Then, the expression is

$$Y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I_{80})$$

The fit is good because the $R^2 = 0.9245$ is close to one.

(b) The estimated trend for Nex Mexico is

$$\mu = 13.15385 - 0.23616 year$$

and the predicted trend for US is

$$\mu = 13.15385 - 4.41692 - (0.23616 - 0.07778)year = 8.73693 - 0.15838year.$$

Solve the equation

$$13.15385 - 0.23616year = 8.73693 - 0.15838year \Rightarrow year = 56.79 \approx 57.$$

Thus, in the year $56 + 1945 = 2001$, the rate is expected close.

(c) The null hypothesis is $H_0 : \alpha = \gamma = (\alpha\gamma) = 0$.

(d) The predicted values are

$$13.15385 - 0.23616 \times 20 = 8.43065$$

for New Mexico and

$$8.73693 - 0.15838 \times 20 = 5.56933.$$

(e) No, becuase the interaction effect is included in the model.

4. (a) We first find the region of θ_0, $C(\theta_0)$, based on the fact given by

$$\frac{p}{n-p}F_{1-\frac{\alpha}{2},p,n-p} \le \frac{S(\theta) - S(\hat{\theta})}{S(\hat{\theta})} \le \frac{p}{n-p}F_{\frac{\alpha}{2},p,n-p}.$$

We get

$$C(\theta) = \{\theta : S(\hat{\theta})[1 + \frac{p}{n-p}F_{1-\frac{\alpha}{2},p,n-p}] \le S(\theta) \le S(\hat{\theta})[1 + \frac{p}{n-p}F_{\frac{\alpha}{2},p,n-p}]\}$$

Second, we get he predicted region of y by

$$C(y) = \{y : y = f(x_0, \theta), \theta \in C(\theta)\}.$$

Then, we have the confidence prediction region as

$$C_p(y) = \{y : y = f(x_0, \theta) \pm 1.96\hat{\sigma}, \theta \in C(\theta)\},$$

where

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum[y_i - f(x_i, \hat{\theta})]^2.$$

(b) The plot can not tell us which one is better since it is based on the log-scale of the survival function.

5. (a)

$$S(24) = (1 - \frac{1}{11})(1 - \frac{1}{10})(1 - \frac{6}{7}) = 0.7012.$$

We use the formula

$$V(\hat{S}(t)) = \hat{S}(t)^2 \sum \frac{d_i}{r_i(r_i - d_i)}$$

where d_i is the death and r_i is the at risk population at time i. Thus, we have

$$V(\hat{S}(24)) = 0.7012^2[\frac{1}{11(11-1)} + \frac{1}{10(10-9)} + \frac{1}{7(7-1)}] = 0.02164.$$

The standard deviation is $0.02164^{0.5} = 0.1470$.

(b) For this case, $h(t) = 2t/\theta^2$ and $S(t) = e^{-t^2/\theta^2}$. The loglikelihood function is

$$\begin{aligned} \ell &= \log\{\prod_{i=1}^{n}[h(t_i)]^{\delta_i}[S(t_i)]\} \\ &= \sum_{i=1}^{n} \delta_i \log(2t_i) - \sum_{i=1}^{n} 2\delta_i \log(\theta) - \sum_{i=1}^{n} \frac{t_i^2}{\theta^2}. \end{aligned}$$

It gives

$$\hat{\theta} = \left[\frac{\sum_{i=1}^{n} t_i^2}{\sum_{i=1}^{n} \delta_i}\right]^{1/2} = 50.86.$$

The Fisher Information is

$$I(\theta) = \frac{2\sum_{i=1}^{n} \delta_i}{\theta^2} + \frac{6\sum_{i=1}^{n} t_i^2}{\theta^4} \Rightarrow \frac{1}{I(\hat{\theta})} = 46.19$$

indicating $\sigma(\hat{\theta}) = 46.19^{1/2} = 6.796$.

6. (a) For S_1, the counts are $676 + 569 = 1245$, $656 + 557 = 1213$, $93 + 153 = 246$ and $151 + 127 = 278$ respectively. The odds ratio is

$$\hat{\theta}_1 = \frac{1245}{1213} = 1.1599$$

and the standard error of $\log(\theta)$ is

$$\sigma(\log(\hat{\theta}_1)) = \frac{1}{1245} + \frac{1}{1213} + \frac{1}{246} + \frac{1}{278} = 0.00929.$$

Since

$$|\frac{\log(1.1599)}{\sqrt{0.00929}}| = 1.54 < 1.96.$$

The IQ and S_1 are marginal independent. Similarly, for S_2, we have

$$\hat{\theta}_2 = \frac{(769)(684)}{(807)(722)} = 0.9028.$$

and

$$V(\log(\hat{\theta}_2) = \frac{1}{769} + \frac{1}{807} + \frac{1}{722} + \frac{1}{684} = 0.00539.$$

Note that

$$|\frac{\log(0.9028)}{\sqrt{0.00539}}| = 1.39 < 1.96.$$

We still conclude S_2 and IQ are marginal independent.

(b) For IQ low, we have

$$\hat{\theta}_1 = \frac{(676)(153)}{(569)(94)} = 1.934$$

and

$$\sigma(\log(\hat{\theta}_1)) = \frac{1}{676} + \frac{1}{153} + \frac{1}{569} + \frac{1}{94} = 0.01235.$$

Since

$$|\frac{\log(1.934)}{\sqrt{0.01235}}| = 5.94,$$

we conclude that S_1 and S_2 are significantly not independent for Low IQ. Similarly, we have

$$\hat{\theta}_2 = \frac{(656)(127)}{(557)(151)} = 0.9905.$$

and

$$\sigma(\log(\hat{\theta}_1)) = \frac{1}{656} + \frac{1}{557} + \frac{1}{151} + \frac{1}{127} = 0.0178.$$

Since

$$|\frac{\log(0.9905)}{\sqrt{0.0178}}| = 0.07,$$

we conclude that in hig IQ group, S_1 and S_2 are almost independent.

(c) The data indicates that S_1 and S_2 are both high and both low are more likely that one is high and one is low. We can also say that the risk for S_1 to be low when S_2 is low is about 30% righ than when S_2 is high marginal. For IQ low, it becomes about 90%.

7. (a) The ANOVA table is

	df	SS	MS	F
Drug A	1	15	15	7.69
Drug B	1	10	10	5.13
A*B	1	5.5	5.5	2.82
Gender	1	10	10	5.13
Day	7	36	5.14	2.63
Error	20	39	1.95	
Total	31	115.5		

(b) If Day is a random effect, then the F-value of day is still 2.63. Since $F_{0.05,7,20} = 2.51$. The day effect issignificant. For Gender, drug interaction effect, the p-values are 1.41 and 0.77 respectively. Thus, Gender and drug interaction are not significant.

(c) Let σ^2 be the variance of the error term and let σ_d^2 be the variance of the day effect. Note that

$$V(\bar{Y}) = \frac{\hat{\sigma}_d^2}{8} + \frac{\hat{\sigma}^2}{20}.$$

Thus, its 95% confidence interval is

$$[10 - t_{0.025,7}(\frac{5.14}{8} + \frac{1.95}{20})^{1/2}, 10 + t_{0.025,7}(\frac{5.14}{8} + \frac{1.95}{20})^{1/2}] = [7.97, 12.03].$$