Solutions to Methods in Winter 2005

1. (a) Note in the linear model, we assume the error terms are iid normality distributed, and the response variable is linearly dependent on predictor variables. Thus, we need to look at the residual plots: (i) plot residual versus each of three predictor variables for linearity and we expected to see no obvious pattern in these plots; (ii) plot residual versus predicted response for homogeneous variance; (iii) plot residuals versus order for independence; (iv) qq-plot and so on. Additionally, we also need to look at the outlier in the residual plots. Among those, the outlier detection plot is the most important and then is the plot for homogeneous variance.

   (b) The plot tells us what λ should be chosen for the transformation of $Y^\lambda/(\lambda-1)$. Since the best λ is 0, we need a log-transformation on the response.

   (c) The fitted model for rates in $100,000$ individuals is

   $$Y = 169.3797 + 1.0664X_1 - 1.2315X_2 + 0.9185X_3.$$

   The model says the rate is expected to increase 1.0664 in $100,000$ if the rate of minority increases 1%, the rate is expected to decrease -1.2315 in $100,000$ if the percentage of people ≥ 25 years old in college increases 1% and the rate is expected to increase 0.9185 in $100,000$ if the percentage of not single houses increase 1%.

   (d) The null hypothesis for the t test of housing is the coefficient of housing is 0 and the alternative hypothesis is it is not 0; the null hypothesis for the t test of college is the coefficient of housing is 0 and the alternative hypothesis is it is not 0; the null hypothesis for the t test of minority is the coefficient of housing is 0 and the alternative hypothesis is it is not 0. The overall F test states the null as all the coefficients of minority, college and housing are 0 versus the alternative that at least one is not 0.

   (e) It suggests that the cities used in the study have average rates higher than west Lafayette.

   (f) The ANOVA Table is

	DF	Sum Sq	Mean Sq	F value
minority	1	14895	14895	66.73
highschool	1	5057	5057	22.66
housing	1	4606	4606	20.64
Residuals	1	13838.4	223.2	

(g) $R^2 = 0.6563$ and Adjusted $R^2 = 0.6396$.

$$R^2 = \frac{17272.1 + 4549.0 + 4608.4}{17272.1 + 4549.0 + 4608.4 + 13842.2} = 0.6563$$

and

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{n-p} = 1 - \frac{(1 - 0.6563)65}{62} = 0.6396.$$

2. (a) We may either use loglikelihood statistic or Pearson statistic. When all the rates are the same, we have $\hat{p} = 0.1285$. Then, the predicted values for the cases are 740.04, 462.60, 712.54 and 89.82. Thus, we have

$$G^2 = 2\sum\sum n_{ij}\log(n_{ij} - \hat{n}_{ij}) = 84.83$$

and

$$X^2 = \sum\sum \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 83.01.$$

Both are significant based on χ^2_3 distribution and so the rates are significantly different.

(b) In natural group, $\hat{p}_n = 0.1285$, in the inoculated group, $\hat{p}_i = 0.01207$. Since

$$\log\frac{\hat{p}_n}{1 - \hat{p}_n} = \hat{\alpha}$$

and

$$\log\frac{\hat{p}_i}{1 - \hat{p}_i} = \hat{\alpha} + \hat{\beta}.$$

It gives $\hat{\alpha} = -1.91$ and $\hat{\beta} = -2.49$.

Note that $\hat{\beta}$ is the logarithm of odds ratio. Let C_1, C_2 be the numbers of cases and D_1 and D_2 be the numbers of death. Then, we have $C_1 = 15603$, $C_2 = 7788$, $D_1 = 2005$, $D_2 = 94$. Thus, we have

$$V(\hat{\beta}) = \frac{1}{C_1 - D_1} + \frac{1}{D_1} + \frac{1}{C_2 - D_2} + \frac{1}{D_2} = 0.01134$$

and $\sigma(\hat{\beta}) = 0.1065$.

(c) Let death be d_i and case be c_i. We put the results in the formulae of the Pearson residual for te binomial data as

$$r_{PR} = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}} = \frac{d_i/c_i - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/c_i}}$$

and the deviance residual as

$$r_{DR} = sign(y - \hat{\mu})\sqrt{d_i} = sign(\frac{d_i}{c_i} - p_i)[2(d_i\log\frac{(d_i/c_i)}{\hat{p}} + (c_i - d_i)\log\frac{1 - (d_i/c_i)}{1 - \hat{p}})]^{1/2}.$$

We have for the collapsed 2×2 table, they are 0. (You may use it as thye answer). In addition, I also compute the Pearson and deviance residual for the whole table. The Pearson residuals for the natural smallpox are $4.01, 1.86, -6.96, 3.86$; the Pearson residuals for the inoculated smallpox are $1.37, 3.28, 0.87, -1.83$; the deviance residuals for the natural smallpox are $3.94, 1.84, -7.24, -3.68$; the deviance residual residuals for the inoculated smallpox are $1.24, 2.76, 0.84, 1.91$.

3. The effect of students should be considered as an error term. There are total $n = 3(20) = 60$ observations.

   (a) School should be treated as fixed effect since all must be chosen. Skill effect should be treated as random effect since skills were selected form a long list.

   (b) The table is

Source	DF	SS	MS	F
School	2	220.0	110.0	11
Skill	4	96.0	24.0	2.4
School*Skill	8	176	22	2.2
Error	45	450	10	

In order to calculate the distribution of the F statistic, we should consider the model

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ijk},$$

where $\gamma_j \sim N(0, \sigma_1^2)$, $(\alpha\gamma)_{ij} \sim N(0, \sigma_2^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. We denote the first to the last in the above table as MSA, MSB, $MSAB$ and MSE respectively. Note that

$$\begin{aligned}
SST =& \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(y_{ijk} - \bar{y}_{...})^2 \\
=& \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(y_{ijk} - \bar{y}_{ij.})^2 + \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\
&+ \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(\bar{y}_{i..} - \bar{y}_{...})^2 \\
=& SSE + SSAB + SSB + SSA.
\end{aligned}$$

The real distribution of the statistics can be found in textbook. In the following, I display the inference. It is clear that $SSE \sim \sigma^2\chi^2_{45}$. To compute the distribution of $SSAB$, we note that

$$\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} = [(\alpha\gamma)_{ij} - (\bar{\alpha\gamma})_{i.} - (\bar{\alpha\gamma})_{.j} + (\bar{\alpha\gamma})_{..}] + (\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..} - \bar{\epsilon}_{.j.} + \bar{\epsilon}_{...}),$$

which follows follows normal distribution with mean 0. Further, we have $SSAB \sim (\sigma^2 + 4\sigma_2^2)\chi_8^2$. Note that

$$\bar{y}_{.j.} - \bar{y}_{...} = (\gamma_j - \bar{\gamma}) + [(\bar{\alpha\gamma})_{.j} - (\bar{\alpha\gamma})_{..}] + (\bar{\epsilon}_{.j.} - \bar{\epsilon}_{...})$$

which also follows a normal distribution with mean 0. We have $SSB \sim (\sigma^2 + 4\sigma_2^2 + 12\sigma_1^2)\chi_4^2$. Finally, we have

$$\bar{y}_{i..} - \bar{y} = (\alpha_i - \bar{\alpha}) + [(\bar{\alpha\gamma})_{i.} - (\bar{\alpha\gamma})_{..}] + (\bar{\epsilon}_{.j.} - \bar{\epsilon}_{...})$$

which folloes a normal distirbution with mean $(\alpha_i - \bar{\alpha})$, which indicates that SSA does not follow a χ^2 distribution, but we can calculate

$$E(SSA) = 20\sum_{i=1}^{3}(\alpha_i - \bar{\alpha})^2 + 4\sigma_2^2 + \sigma^2.$$

Then, we have $E(MSE) = \sigma^2$, $E(MSAB) = (\sigma^2 + 4\sigma_2^2)$, $E(MSB) = (\sigma^2 + 4\sigma_2^2 + 12\sigma_1^2)$, $E(MSA) = 20\sum_{i=1}^{3}(\alpha_i - \bar{\alpha})^2 + 4\sigma_2^2 + \sigma^2$, where

$$\begin{aligned}
MSA =& 10\sum_{i=1}^{3}(\bar{y}_{i..} - \bar{y}_{...})^2 \\
MSB =& 3\sum_{j=1}^{5}(\bar{y}_{.j.} - \bar{y}_{...})^2 \\
MSAB =& \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\
MSE =& \frac{1}{45}\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}(y_{ijk} - \bar{y}_{ij.})^2.
\end{aligned}$$

In order to test the school effect, we take null $\alpha_1 = \alpha_2 = \alpha_3 = 0$ which indicates that we need to consider $MSA/MSAB$ since under the null, MSA and $MSAB$ have the same expected value. The result is $110/22 = 5$ comparing it with $F_{0.05,2,45} = 3.27$, we reject the null.

(c) Let μ be the overall mean. Then we can construct the 95% confidence interval based on t_{45} distribution. The overall average is

$$\hat{\mu} = \frac{1}{60}\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4} y_{ijk} = \mu + \frac{1}{5}\sum_{j=1}^{5}\gamma_j + \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5}(\alpha\gamma)_{ij} + \frac{1}{60}\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{4}\epsilon_{ijk}$$

indicating that

$$\hat{\mu} \sim N(\mu, \frac{1}{5}\sigma_1^2 + \frac{1}{15}\sigma_2^2 + \frac{1}{60}\sigma^2).$$

To estimate those parameters we have $\hat{\sigma}^2 = 10$, $\hat{\sigma}_2^2 = (22 - 10)/4 = 12$ and $\hat{\sigma}_1^2 = (24 - 10)/12 = 1.17$. Then, we can estimated accordingly and so a 95% confidence interval for μ is derived based on 4 degree of freedom.

4. (a) Yes, the model fits the data since residual deviance is 6.51. Based on χ^2_6 distribution, it is small enough.

   (b) The object is to look at the ratio of numbers of satell and cases. Then, $\log(cases)$ must be an offset term.

   (c) This model considers the number of cases instead of the ratio as the response variable. It will give a large predicted value for number of Satell of crabs with a particular measurement of Width. Result can not be explained according to the size influences the number of satellite crabs.

5. (a) For normal weight, the survival function is: $\hat{S}(2) = 0.9$, $\hat{S}(4) = 0.8$, $\hat{S}(6) = 0.7$, $\hat{S}(8) = 0.6$, $\hat{S}(9) = 0.48$, $\hat{S}(10) = 0.36$ and $\hat{S}(12) = 0.36$. For overweight weight, the survival function is: $\hat{S}(1) = 0.9$, $\hat{S}(3) = 0.8$, $\hat{S}(4) = 0.7$, $\hat{S}(5) = 0.6$, $\hat{S}(6) = 0.5$, $\hat{S}(7) = 0.4$, $\hat{S}(9) = 0.2667$, $\hat{S}(11) = 0.1333$, $\hat{S}(12) = 0.1333$.

   We can test the equality based on the log-rank test. Let $r_k(t_j)$ be the at risk in group k, where $k = n$ or $k = o$. Let $r(t_j)$ be the overall at risk. Then, the expect death in group k is $D_k r_k(t_j)/r(t_j)$. For example, when $t = 1$, the at risk are all 10. It splits 0.5 to normal group and 0.5 to overweight group. When $t = 2$, the at risk in overweight group is 9 so it gives 10/19 to normal group and 9/19 to overweight group. Do it for all the numbers. We have the expected for normal group is

$$\frac{1}{2} + \frac{10}{19} + \frac{1}{2} + \frac{2(9)}{17} + \frac{8}{15} + \frac{2(8)}{14} + \frac{7}{12} + \frac{7}{10} + \frac{2(5)}{8} + \frac{4}{6} + \frac{3}{5} = 8.06.$$

   We observed 6 early-stops in the normal group and 8 early-stops in the overweight group. Thus, the expected in group overweight is $14 - 8.06 = 5.94$. Based on Pearson statistic, we have

$$X^2 = \frac{(6 - 8.06)^2}{8.06} + \frac{(8 - 5.94)^2}{5.94} = 1.24.$$

   Based on χ^2_1 distribution, it is not large and we accept that the survival functions are the same.

   (b) In the first group, the numbers of stopping early or not are $(6, 3)$ and in the second group, they are $(8, 1)$. Since we accept that the survival functions are the same. Thus, in this case, there are 1 stopped early and 4 not. The estimated of logarithm of the relative risk is $\log(\hat{r}) = \log(14/4) = 1.2528$. The estimated of the variance is

$$V(\log \hat{r}) = \frac{1}{14} + \frac{1}{4} = 0.3214.$$

Thus, the 95% confidence interval is

$$[e^{1.2528-1.96\sqrt{0.32}}, e^{1.2528-1.96\sqrt{0.32}}] = [1.1522, 10.63].$$