QUALIFYING EXAM SOLUTIONS
Statistical Methods
Saturday, Jan 2, 2008, 8:00 am -12:00 pm

1. (a) Plot is omitted. We can find strong evidence of interaction effect from the plot.

   (b) The ANOVA model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk};\ i = 1, 2, 3,\ j = 1, 2,\ k = 1, \cdots, 11.$$

   Note that

$$\begin{aligned}SST =& \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{11}(y_{ijk} - \bar{y}_{\cdots})^2 \\ =& 22\sum_{i=1}^{3}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 + 33\sum_{j=1}^{2}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2 + 11\sum_{i=1}^{3}\sum_{j=1}^{2}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{j\cdot\cdot} + \bar{y}_{\cdots})^2 \\ &+ \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{11}(y_{ijk} - \bar{y}_{ij\cdot})^2 \\ =& SSA + SSG + SSAG + SSE.\end{aligned}$$

   Thus, we can compute SSA, SSG and $SSAG$ from the table, but we can not compute SSE. In this table, we only need

$$SSG = 33[(8.94 - 11.70)^2 + (14.45 - 11.70)^2] = 66(8.94 - 11.70)^2 = 502.76.$$

   Then, we have the table

Source	df	SS	MS	F
G	1	502.76	502.76	14.69
A	2	3286.39	1643.2	48.02
G*A	2	491.30	245.65	7.18
Error	60	2053.49	34.22	
Total	65	6333.94		

   (c) Since Age is a quantity variable, we may need to look at whether the effect is linear for either interaction of main effect.

   (d) The plot show that the equal variance assumption is violated. We need a Box-cox transformation. This plot shows that it is very likely the squared root transformation.

2. Suppose we consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

   where $\epsilon_i \sim^{iid} N(0, \sigma^2)$, which gives estimators as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X_1.$$

In order to obtain those, we need the following quantities

$$\sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

and

$$\sum_{i=1}^{n} (Y_i - \bar{Y}).$$

However, $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ cannot be recovered from the condensed data. Therefore, we have to use some other models. One option is the joint modeling method as fitted a weighted regression as

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma_i^2)$ and i indicates the different values of X_i. We can model σ_i^2 by Gamma GLM and derived the estimated of σ^2 and then fit the model. The limitation of the condensed data is that this will give a larger variance estimates.

3. The study design is OK as it provides two counties with observations before and after the ban respectively, and the data have been summarized into a 2 × 2 table. The best way to analyze the 2 × 2 table is the use of the Odds Ratio. Let θ be the odds ratio. For this particular data, we have

$$\hat{\theta} = \frac{17 \times 16}{5 \times 18} = 3.02$$

and

$$\sigma_{\log \hat{\theta}} = [\frac{1}{17} + \frac{1}{5} + \frac{1}{18} + \frac{1}{16}]^{1/2} = 0.6139.$$

The 95% confidence interval is

$$3.02 e^{\pm 1.96 \times 0.6139} = [0.9067, 10.06],$$

which is insignificant. In addition, we can also use two-sample binomial methods, in which we assumes $X \sim bin(22, p_1)$ for Monreo County and $Y \sim bin(34, p_2)$ for Delaware County. Then, we have $\hat{p}_1 = 17/22 = 0.7727$, $\hat{V}(\hat{p}_1) = 0.007983$, $\hat{p}_2 = 18/34 = 0.5294$ and $\hat{V}(\hat{p}_2) = 0.007375$. When $p_1 = p_2 = p$, we have $\hat{p} = 0.625$. Then, the z-score is

$$\frac{0.7727 - 0.5294}{\sqrt{0.625(1 - 0.625)(1/22 + 1/34)}} = 1.839,$$

which implies insignificant at level 0.05. Thus, we conclude insignificance of the test.

The method used by this problem is try to combined the conclusion of the two test together. Since each of them may be a mistake with some probability, the total error rate could be higher than 0.05. This is caused by the multiple testing problem. Thus, we suspect the conclusion of this problem. In addition, our standard method gives a different answer.

4. (a) Assume $Y_{ij} \sim Poisson(\lambda_{ij})$, where Y_{ij} is the count at the (i,j)-th unit of the table, with $i,j = 1,2,3$. The fitted model is

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j.$$

This is the independent model. The fitted model gives estimates $\hat{\lambda}_{11} = 102.05$, $\lambda_{12} = 161.37$, $\lambda_{13} = 135.58$, $\lambda_{21} = 120.21$, $\lambda_{22} = 190.08$, $\lambda_{23} = 159.71$, $\lambda_{31} = 54.73$, $\lambda_{32} = 86.55$, and $\lambda_{33} = 72.72$. The independence assumption is rejected since the deviance goodness is $G^2 = 105.66$ based on 4 degrees of freedom ($\chi^2_{0.05,4} = 9.488$, but $G^2 > 9.488$).

(b) In the first part, let $\pi_1(x) = P(Democrat|x)$, $\pi_2(x) = P(Indepedent|x)$ and $\pi_3(x) = P(Republican|x)$. The fitted multinomial model is

$$\log(\frac{\pi_2}{\pi_1}) = -0.3694 + 0.2707x$$

and

$$\log(\frac{\pi_3}{\pi_1}) = -3.323 + 1.213x,$$

where x is $1,2,3$ respectively to Liberal, Moderate and Conservative.

In the second part, let $\pi_1(x) = P(Liberal|x)$, $\pi_2(x) = P(Moderate|x)$ and $\pi_3(x) = P(Conservative|x)$. The fitted multinomial model is

$$\log(\frac{\pi_2}{\pi_1}) = -0.5024 + 0.5758x$$

and

$$\log(\frac{\pi_3}{\pi_1}) = -1.6483 + 1.0767x,$$

where x is $1,2,3$ respectively to Democrat, Independent, and Republican.

(c) The models in (b) is equivalent to the row effect models for Poisson data. The best answer I think should be try the proportional odds model as

$$\log\frac{\pi_1 + \cdots + \pi_j}{\pi_{i+1} + \cdots + \pi_J} = \beta_{0j} + \beta_1 x; j = 1, \cdots, J-1.$$

Here $J = 3$. In addition, we may also consider to fit a row-column effect model, but row-column effect model is not a generalized linear model.

5. (a) The Conway-Maxwell-Poisson distribution satisfies $V(Y) \geq E(Y)$ with $V(Y) = E(Y)$ when $v = 1$. Thus, it can be used to model overdispersion.

(b) Here, we only need to look at λ^y term. Thus, the GLM model expression is

$$P(Y = y) = \exp\{y\log(\lambda) - \log Z(\lambda, v) - v\log(y!)\}$$

which implies the log-link is the canonical link.

6. This means that given SES, Boy Scout and Delinquent behavior is independent. However, marginal they are model Exactly, those are caused by the following Poisson model as

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

where i, j and k represent Boy Scout, Delinquent Behavior and SES respectively. This is the conditional independent model, but not independent model. It may cause marginally independent and this phenomenon is called Simpson's Paradox. For example, this could happen in the following data:

	SES			
	Yes		No	
Delinquent	Boy Scout		Boy Scout	
Behavior	Yes	No	Yes	No
Yes	10	20	90	30
No	20	40	30	10