QUALIFYING EXAM SOLUTIONS
Statistical Methods
Fall, 2010

1. (a) The model is

$$y_i = \mu + \beta_1 dd + \beta_2 I_{season=1} + \beta_3 dd I_{season=1} + \epsilon_i,$$

where $\epsilon_i \sim^{iid} N(0, \sigma^2)$.

(b) The ANOVA table is

Source	DF	SS	MS	F-value	P-value
Model	4	10155748264	2538937066	65.29	
Error	32	1244379077	38886846		
Total	36	11400127341			

We use the F-statistic F^* to test H_0: all the independent variables do not have any effect. We reject H_0 and conclude H_1 that at least one independent variable is significant.

(c) This is given by

$$R^2 = \frac{10155748264}{11400127341} = 0.8908.$$

(d) The fitted model for season two is

$$Y = (90213.88 - 3539.67) + (60.51 - 39.93)dd = 86674.21 + 20.58dd$$

and for season two is

$$Y = 90213.88 - 60.51dd.$$

Because one the season main and the season:dd interaction effect are significant, we conclude there is a significantly season effect.

(e) We have two methods. In the first, we can define a dummy variable which is zero before the half way and one after the half way, and fit it in the mode. In the second, we can use the residual to test whether the residuals before the half way and after the half way have the same expected value. The model can be written down accordingly.

2. (a) The model is

$$y_{ijk} = \mu + \alpha_i + \gamma_{j(i)} + \epsilon_{ijk}, i = 1, 2, 3, j = 1, 2, 3, k = 1, \cdots, 9$$

where $\gamma_{j(i)}$ is nested in locations, and $\epsilon_i \sim N(0, \sigma^2)$ is the error term. We may take α_i as either a fixed effect or a random effect. If it is fixed, then $\alpha_1 + \alpha_2 + \alpha_3 = 0$. If it is random, then $\alpha_i \sim N(0, \sigma_\alpha^2)$. For $\gamma_{j(i)}$, we assume $\gamma_{j(i)} \sim N(0, \sigma_\gamma^2)$ as a random effect.

(b) We can use the ANOVA values to estimate the effect of locations, and site nested in locations. Suppose we have SSA for locations, SSB for sites, SSAB for location-site interaction, and SSE for the error term. Then SSA can represent the location main effect, SSB+SSAB can represent the nested effect. Their degrees of freedom are 2 and 6 respectively.

(c) Let $MSE = SSE/72$, $MSB(A) = (SSB + SSAB)/6$ and $MSA = SSA/2$. The estimates are $\hat{\sigma}^2 = MSE$, $\hat{\sigma}_\gamma^2 = MSB(A) - MSE)/9$, and $\hat{\sigma}_\alpha^2 = (MSA - MSB(A))/27$. The advantage is the method is simple. The disadvantage is that the estimates of variances may be negative.

(d) We can test $H_0 : \sigma_\alpha^2 = 0$ and use

$$F^* = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}^2}.$$

However, it does not follow an F- distribution. We need to consider a bootstrap method to compute the p-value.

(e) In this case, we must treat α_i as a fixed effect. Then, we use

$$T = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{s(\hat{\alpha}_1 - \hat{\alpha}_2)}$$

which follows t_{72} under $H_0 : \alpha_1 = \alpha_2$.

3. (a) These two are equivalent. Please read the solution of Problem 2 in 2012.

(b) The estimates of parameters are the same but the residual deviance and the residual degree of freedom are different. The assumption from the binary data to the group data is: the proportion are the same if the independent variable are the same. Therefore, the saturated models in these two cases are different.

(c) In i, the model in (1) is

$$\log \lambda_{ijkl} = \mu + \alpha_i + (\beta\gamma\delta)_{jkl}$$

where α_i is the price range, and $(\beta\gamma\delta)_{jkl}$ is the combination of the rest independent variables. In (2) the model is

$$\log \lambda_{ijkl} = \mu + \alpha_i + (\beta\gamma\delta)_{jkl} + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il}.$$

In ii, we can use the goodness of fit statistic G^2 defined by the difference of the residual deviances between the two models. The degree of freedom is 12.

4. (a) The PDF can be written into

$$f(y) = \lambda e^{-\lambda y} = e^{-\lambda y + \log \lambda}.$$

Thus, $\theta = -\lambda$, $b(\theta) = -\log \lambda = -\log(-\theta)$, $\phi = 1$, $a(\phi) = 1$, and $c(y, \phi) = 0$.

(b) Note that $\mu = 1/\lambda$. The canonical link is

$$g(\mu) = \theta \Rightarrow \theta = -\frac{1}{\mu} \Rightarrow g(\mu) = -\frac{1}{\mu}.$$

Note that $b'(\theta) = -1/\theta$ and $b''(\theta) = 1/\theta^2$. We have $V(Y) = 1/\lambda^2 = \mu^2$.

(c) If we fit

$$-\frac{1}{\mu} = x'\beta$$

then the right side can be any real numbers but the left must be negative.

(d) We need to use a χ^2 distribution because ϕ is not a real parameter.

(e) Let y_i and $\hat{y}_i$ be the observations and fitted value for $i = 1, \cdots, n$. Then, $\hat{\mu}_i = \hat{y}_i$ which implies $\hat{\lambda}_i = 1/\hat{y}_i$. Thus, the deviance is defined by

$$\begin{aligned} G^2 =& -2\sum_{i=1}^{n} \log(\hat{\lambda}_i e^{-\hat{\lambda}_i y_i}) \\ =& -2\sum_{i=1}^{n} \log(\frac{1}{\hat{y}_i} e^{-y_i/\hat{y}_i}) \\ =& 2\sum_{i=1}^{n} (\frac{y_i}{\hat{y}_i} + \log \hat{y}_i). \end{aligned}$$

5. (a) The independence model is

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

with $\sum_{i=1}^4 \alpha_i = \sum_{j=1}^4 \beta_j = 0$, and residual *df* $= 9$. The linear-by-linear association model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma u_i v_j,$$

where $u_i = 1, 2, 3, 4$ for $i = 1, 2, 3, 4$, respectively, and $v_j = 1, 2, 3, 4$ for $j = 1, 2, 3, 4$, respectively. The residual *df* $= 8$. The row-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_i v_j,$$

where γ_i is unknown and $v_j = 1, 2, 3, 4$ for $j = 1, 2, 3, 4$. We may assume $\sum_{i=1}^4 \gamma_i = 0$. The residual *df* $= 4$. The column effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_j u_i.$$

The residual *df* $= 4$. The saturated model is

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

with $\sum_{i=1}^4 \alpha_i = \sum_{j=1}^4 \beta_j = \sum_{i=1}^4 (\alpha\beta)_{ij} = \sum_{j=1}^4 (\alpha\beta)_{ij} = 0$. The residual *df* $= 0$.

(b) The G^2 of the independence model is

$$G^2 = 2y_{ij} \log(y_{ij}/\hat{y}_{ij})$$

where $y_{ij} = y_{i+}\bar{y}_{+j}/y_{++}$ with y_{i+}, y_{+j}, and y_{++} be the row sum, column sum, and total sum, respectively. For this data, we have 20.518 based on 9 degree of freedom. Because $\chi_{0.05,9} = 16.92$, we conclude the independence model does not fit the data.

(c) The difference of the residual deviances is $20.518 - 7.979 = 12.539 > \chi^2_{0.05,1} = 3.84$. Therefore, the reduction of the deviance is significant, which implies the linear-by-linear model is more appropriate. In addition, the residual deviance of the linear-by-linear model is $7.979 < \chi_{0.05,8} = 15.51$. Thus, the linear-by-linear model is good enough for the data.

6. (a) For treatment group, $\hat{S}(10) = (1 - 1/16) = 15/16 = 0.9375$ and $\hat{S}(12) = 0.9375 \times (1 - 1/13) = 0.8705$. For the placebo group, $\hat{S}(7) = 15/16 = 0.9375$, $\hat{S}(9) = 0.9375(1 - 1/15) = 0.875$, $\hat{S}(10) = 0.875(1 - 1/13) = 0.8077$, and $\hat{S}(12) = \hat{S}(11) = 0.8077(1 - 1/12) = 0.7404$.

(b) We can use Greenwood's formula for variance

$$\hat{V}[\hat{S}(t)] = \hat{S}^2(t) \sum_{t_i \le 12} \frac{d_i}{Y_i(Y_i - d_i)}.$$

For the treatment group, we have $\hat{V}[\hat{S}(12)] = 0.0073$. The 95% confidence interval for $S(12)$ is

$$0.8705 \pm 1.96 \times \sqrt{0.0073} = [0.7030, 1]$$

For the placebo group, we have $\hat{V}[\hat{S}(12)] = 0.01256$. The 95\$ confidence interval for $S(12)$ is

$$0.7404 \pm 1.96 \times \sqrt{0.01256} = [0.5207, 0.9601].$$

(c) We need to consider the log-rank test. It is not easily to judge the conclusion exactly.

(d) The MLE of λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^n d_i}{\sum_{i=1}^n t_i}.$$

We have $\hat{\lambda}_{trt} = 11/397 = 0.0277$ and $\hat{\lambda}_{pla} = 11/235 = 0.0468$.

(e) Part a) does not have any assumption for the survival function. Therefore, it can be used to diagnose the exponential distribution assumption.