QUALIFYING EXAM SOLUTIONS
Statistical Methods
Spring, 2011

1. (a) The simple linear regression is

$$y_{ij} = \beta_0 + \beta_1 X_i + \epsilon_{ij}$$

$i = 1, 2, 3$, $j = 1, \cdots, n_i$, where $\epsilon \sim N(0, \sigma^2)$, and β_0 is the expected value and β_1 is the slope.

(b) In this problem, we have

$$\begin{aligned} S_{yy} =& \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ =& 0.23 + 0.14 + 0.35 + 3(4.0 - 3.575)^2 + 2(3.8 - 3.575)^2 + 3(3.0 - 3.575)^2 \\ =& 2.355 \end{aligned}$$

$$\begin{aligned} S_{xy} =& \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) \\ =& -3(4.0 - 3.575) + 3(3.0 - 3.575) \\ =& -3 \end{aligned}$$

and

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = 3 \times 1 + 3 \times 1 = 6.$$

Therefore, we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -\frac{3}{6} = -0.5$$

and

$$\hat{\beta}_0 = 3.575 + 0.5 \times 2.0 = 4.575.$$

(c) The ANOVA model is

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

where μ_i is the expected value of the i-th level. This model reduces to the linear regression model if (x_i, μ_i) are on a straight line. The estimates of parameters are $\hat{\mu}_1 = 4.0$, $\hat{\mu}_2 = 3.8$ and $\hat{\mu}_3 = 3.0$. They are unbiased if the simple linear regression hold because $\hat{\beta}_0 + \hat{\beta}_1 x$ are unbiased. The variance may be greater than the variance of the simple linear regression model because its residual degree of freedom is large.

We need to compute the SSE in the simple linear regression model. For the regression model, we have $\hat{Y}_{1j} = 4.075$, $\hat{Y}_{2j} = 3.575$, and $\hat{Y}_{3j} = 3.075$. Then, we have

$$SSM = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(\hat{Y}_{ij} - \bar{Y})^2 = 3(4.075 - 3.575)^2 + 3(3.075 - 3.575)^2 = 1.5.$$

Then, $SSE = 2.355 - 1.5 = 0.855$. The F-statistic is

$$F^* = \frac{(0.855 - 0.23 - 0.14 - 0.35)/1}{(0.23 + 0.14 + 0.35)/5} = 0.9375$$

which is less that $F_{0.05,1,5} = 6.608$. Thus we accept H_0 and conclude the linear regression model.

2. (a) The model is

$$y_i = \beta_0 + \beta_1 age + \beta_2 smk + \beta_3 quet + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$. The estimate of smk, $\hat{\beta}_2 = 10.20701$, states the change of the intercept from nonsmoker to smoker.

(b) The partial coefficient of determination of quet is defined by

$$\frac{SSR(quet|smk, age)}{SSE(quet)} = \frac{236.08086}{236.08086 + 1487.55817}.$$

It reflects the proportion of quet in the residual sum of squares. Because it is small, we may exclude it from the model.

(c) The plot shows that sbp and quet are almost linear, but we may not be able to include it in the model because of multicollinearity.

(d) Because Type I SS and Type II SS are very different, there is a concern of multicollinearity.

(e) Ridge regression can reduce the influence of the multicolinearity because it adds a penalty to the least equation.

3. (a) The (factor-effect) model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, i = 1, 2, 3, j = 1, 2, 3, 4, k = 1, 2,$$

where $\epsilon_{ijk} \sim^{iid} N(0, \sigma^2)$ is the error term, α_i is the drinks main effect, β_j is the brands main effect, and $(\alpha\beta)_{ij}$ is the interaction. We need to use the zero-sum constraint. The second part contains two contrast $L_1 = \mu_1 - \mu_2$ and $L_2 = \mu_1 - (\mu_2 + \mu_3)/2$. The null hypothesis is $H_0 : \mu_1 = \mu_2$ in (1) and is $H_0 : \mu_1 = (\mu_2 + \mu_3)/2$. We can use the test statistic

$$T_1 = \frac{\bar{Y}_1 - \bar{Y}_2}{s(\bar{Y}_1 - \bar{Y}_2)}$$

and

$$T_2 = \frac{\bar{Y}_1 - (\bar{Y}_1 + \bar{Y}_2)/2}{s(\bar{Y}_1 - (\bar{Y}_2 + \bar{Y}_3)/2)}.$$

The critical value needs to be adjusted for the Bonfferoni, Scheffé or Tukey method.

(b) The ANOVA model is

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, \cdots, 8, j = 1, 2, 3, 4,$$

where $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$ and $\sum_{i=1}^{8} \alpha_i = 0$. We can interpret the model as: the parameter μ is the average of the 8 brands, α_i is the different between the particular brand and the average, and ϵ_{ij} is the error. For the third part, we use the F-statistic

$$F^* = \frac{3.92/(8-1)}{0.72/24} = 18.67 > F_{0.05,7,24} = 2.42.$$

We conclude the levels of brand are significantly different. The percent is $3.92/(3.92 + 0.72) = 0.8448$. For part iv, we have

$$H_0 : \mu \leq 4 \leftrightarrow H_1 : \mu > 4.$$

The test statistic is

$$T = \frac{\hat{\mu} - 4}{s(\hat{\mu})}$$

which follows t_{24}. We reject H_0 if $T > 1.7109$.

4. (a) The odds ratio is

$$\hat{\theta} = \frac{16 \times 2897}{28 \times 256} = 6.47.$$

(b) The standard error of the $\log \hat{\theta}$ is

$$\sqrt{1/16 + 1/256 + 1/28 + 1/2897} = 0.3201.$$

Thus, the 95% confidence interval is

$$\theta e^{\pm 1.96 \times 0.3201} = [3.45, 12.12].$$

Interpret the result: the risks increases 6.47 times from Cholesterol low to high.

(c) The model is

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

where ϵ_{ijk} is the expected value, α_i is the heart disease main effect, β_j is the gender main effect, and γ_k is the Cholesterol main effect. Based on this model, we have

$$\hat{\lambda}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2}.$$

Therefore for the particular cell, we have

$$\hat{\lambda}_{111} = \frac{80 \times (272 + 2925) \times (272 + 332)}{6117^2} = 4.1285.$$

(d) The model is

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

Based on this model, we have

$$\hat{\lambda}_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}}.$$

Therefore for the particular cell, we have

$$\hat{\lambda}_{111} = \frac{272(16 + 13)}{(272 + 332)} = 13.06.$$

(e) Let $\hat{n}_{ijk}$ be the fitted value from model (c). Then, we have

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

If $G^2 > \chi^2_{0.05,4}$, then we reject model (c).

5. (a) The model is

$$\log\frac{\pi_1}{\pi_2+\pi_3} = \theta_1 - \beta x$$

and

$$\log\frac{\pi_1+\pi_2}{\pi_3} = \theta_2 - \beta x,$$

where x is the value of degree. The odds ratios between "Not at all" and the combination of "Sort of" and "Very" and between the combination of "Not at all" and "Sort of" are the same, which is $e^{0.4614} = 1.5863$. The probabilities are

$$\hat{\pi}_1 = \frac{e^{0.0264+4\times 0.4614}}{1+e^{0.0264+4\times 0.4614}} = 0.8667$$

and

$$\hat{\pi}_3 = \frac{1}{1+e^{2.3137+4\times 0.4614}} = 0.0154.$$

Then, $\hat{\pi}_2 = 1 - 0.8667 - 0.0154 = 0.1179$ and the predicted counts

$$173 \times (0.8667, 0.1179, 0.0154) = (149.94, 20.40, 2.66).$$

(b) The model is

$$\log\frac{\pi_1}{\pi_2+\pi_3} = \theta_1 - \beta_1 I_{x=1} - \beta_2 I_{x=2} - \beta_3 I_{x=3} - \beta_4 I_{x=4}$$

and

$$\log\frac{\pi_1+\pi_2}{\pi_3} = \theta_2 - \beta_1 I_{x=1} - \beta_2 I_{x=2} - \beta_3 I_{x=3} - \beta_4 I_{x=4}.$$

The goodness of fit statistic is

$$G^2 = 2658.161 - 2656.409 = 1.752 < \chi^2_{0.05,3} = 7.81.$$

Thus, we accept the model in (a).

(c) We may consider the baseline loglinear model

$$\log\frac{\pi_2}{\pi_1} = \mu_1 + \beta_1 x$$

and

$$\log\frac{\pi_3}{\pi_1} = \mu_2 + \beta_2 x.$$

6. (a) Here we have $\hat{S}(1) = 1 - 1/9 = 8/9$, and $\hat{S}(2) = 8/9(1-1/8) = 7/9$.

(b) The model assumes $f(t) = \lambda e^{-\lambda t}$ and $S(t) = e^{-\lambda t}$. We have the MLE as

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

Then, we have $\hat{\lambda}_{trt} = 7/27 = 0.2593$ and $\hat{\lambda}_{pla} = 6/16 = 0.375$.

(c) The model is

$$-\log \lambda = \mu + \beta I_{trt=1} \Rightarrow \hat{\lambda} = e^{-(\mu+\beta I_{trt=1})}.$$

We have $\hat{\lambda}_{pla} = e^{-0.981} = 0.3749$ and $\hat{\lambda}_2 = 0.2592$. The treatment is not effective because the z-value is $|0.369/0.556| = 0.664 < 1.95$.

(d) The model is

$$h(t) = h_0(t) e^{\beta I_{trt=1}}$$

where $h_0(t)$ is the hazard function at placebo and $h(t)$ is the hazard function for the particular treatment. We can test $H_0 : \beta = 0$. Because the z-value is $|-0.8188/0.7376| = 1.11 < 1.96$, we conclude treatment effect is not significant. To assess the proportional hazard assumption, we can use the plot $\log\{\log[-S(t)]\}$ versus $\log t$ because under the null hypothesis we have

$$S(t) = [S_0(t)]^{e^{\beta' x}}.$$