QUALIFYING EXAM SOLUTIONS
Statistical Methods
Spring, 2010

1. (a) The model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where β_0, β_1 are unknown parameters, $\epsilon_i \sim^{iid} N(0, \sigma^2)$ is the error term.

(b) He should equally choose the temperature values, such as $t_1 = -1$, $t_{10} = 1$, with $t_i = t_1 + (i-1)(t_{10} - t_1)/9$.

(c) In this model, we have

$$S_{yy} = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 93.429$$

$$S_{xy} = \sum_{i=1}^{10} (y_i - \bar{y})(x_i - \bar{x}) = 21.39$$

and

$$S_{xx} = \sum_{i=1}^{10} (x_i - \bar{x})^2 = 4.9$$

Therefore

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 4.365$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14.274,$$

and

$$MSE = \frac{1}{8}(S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}) = 0.05510.$$

Note that

$$\hat{V}(\hat{\beta}_1) = \frac{MSE}{S_{xx}} = \frac{0.05510}{5} = 0.01102.$$

Thus, the 95% confidence interval for β_1 is

$$4.365 \pm t_{0.025,8} \times \sqrt{0.01102} = [4.1229, 4.6071].$$

It is enough to see the summary values because those are sufficient statistics. This will also provide the same confidence interval for β_1.

2. Let Y be WebsiteUse, X_1 be Income, X_2 be OffFarm, X_3 be CashFlow, X_4 be NumberOrders, and X_5 be NumberDistributions.

(a) The model is

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 I_{X_2=1} + \beta_3 I_{X_2=2} + \beta_4 X_1 I_{X_2=1} + \beta_5 X_1 I_{X_2=2} + \epsilon_i,$$

where $\epsilon_i \sim^{iid} N(0, \sigma^2)$. To test whether the slope depend on X_2, we can compute the SS of $X_1 : X_2$ and define as F-statistic

$$F^* = \frac{SS(X_1 : X_2)/2}{MSE}$$

which follows the $F_{2,24}$ under $H_0 : \beta_4 = \beta_5 = 0$. We reject H_0 if $F^* > F_{0.05,2,24}$.

(b) Multicollinearity can cause (1) the estimate of parameters not precise, and (2) the inflation of the standard error of the estimates. Because the ridge regression can reduce the influence of multicollinearity (make both estimate and standard error stable), we may consider to use it to adjust the least square equation. The ridge regression estimate β by using

$$\hat{\beta} = (\lambda I + X'X)^{-1}X'Y$$

instead of the LSE as $\hat{\beta} = (X'X)^{-1}X'Y$. Then

$$E(\hat{\beta}) = (\lambda I + X'X)^{-1}X'X\beta$$

which is a biased estimate of β.

(c) Sppose the fitted model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$. If $\hat{y} = 200$, then $\hat{x}_1(200) = (200 - \hat{\beta}_0)/\hat{\beta}_1$. A delta-method can be used to derive the confidence interval for $x_1(200)$. Because the true distribution of $\hat{x}_1(200)$ is not normal and the data is not large, we need to consider a simulation method. Then the bootstrap method is used. In this method, we generated m dataset from the fitted model and computed $\hat{x}_1(200)$ for each. Then, we have m values of $\hat{x}_1(200)$, which roughly represent the true distribution of $\hat{x}_1(200)$. Then, a confidence interval is derived based on the simulated distribution.

3. (a) The model is

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

where $\epsilon_{ijk} \sim^{iid} N(0, \sigma^2)$. The constraints are

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{4} \beta_j = \sum_{i=1}^{3} (\alpha\beta)_{ij} = \sum_{j=1}^{4} (\alpha\beta)_{ij} = 0.$$

We need to test the significance of the interaction effect as

$$H_0 : (\alpha\beta)_{ij} = 0.$$

We use

$$F^* = \frac{0.1554}{0.0652} = 2.388$$

which is less than $F_{0.05,6,24} = 2.50$. Thus, we conclude the interaction effect is not significant. We claim the use of different fertilizers does not depend on the seedling rate. To test the difference in yield between the fertilizers, we test

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0.$$

We use

$$F^* = \frac{12.155}{0.0652} = 186.$$

Therefore, we conclude there is significant difference between fertilizers.

(b) We need to test $H_0 : \alpha_1 = \alpha_2$. Note that

$$\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot} = 16.755 - 17.9175 = -1.1625$$

and

$$V(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}) = \sigma^2/6 = 0.0109.$$

Then

$$|T| = |-\frac{1.1625}{\sqrt{0.0109}}| = 11.13 \sim^{H_0} t_{24}.$$

We conclude they are not the same.

(c) In this model, we need to assume $\beta_i \sim^{iid} N(0, \sigma^2_\beta)$ and $(\alpha\beta)_{ij} \sim^{iid} N(0, \sigma^2_{\alpha\beta})$. We need to test $H_0 : \sigma^2_\beta = 0$ versus $H_1 : \sigma^2_\beta > 0$. It can also be assessed by an F-statistic

$$F^* = \frac{MSB}{MSE} = 70.64 \sim^{H_0} F_{3,24}.$$

Thus, H_0 is rejected and we conclude $\sigma^2_\beta > 0$. It says the seeding rate are different among these four levels.

4. (a) The model assumes $O_i \sim Poisson(\lambda)$ with $\hat{\lambda} = 27.7$.

(b) The null is H_0 : the distribution of O_i is Poisson versus H_1 : the distribution of O_i is not Poisson. Under H_0 the Pearson X^2 follows χ^2_5 distribution. Because $X^2 > \chi^2_{0.05,5} = 11.07$, we reject H_0 and conclude the distribution is not Poisson. The test is adequate.

(c) We may use a quasi-Poisson model in which $V(Y) = \sigma^2 E(Y)$, and $\hat{\sigma}^2 = X^2/5 = 7.0776$. Thus, $\hat{V}(\bar{Y}) = \hat{\sigma}^2 \hat{E}(\bar{Y})/200 = 0.9802$. The 95% confidence interval for λ is

$$27.7 \pm 1.96 \times \sqrt{0.9802} = [25.76, 29.64].$$

5. (a) The assumption of the model is multinomial. Because the proportion increases as i increases, the slope must be the same.

(b) The odds ratio is

$$\hat{\theta} = \frac{1640 \times 2544}{3040 \times 2054} = 0.6681.$$

It means the risk of the academic performance for male to be low is about 33.19% lower than that for female.

(c) Based on the model, we have the expected values for female are 1626.44, 3063.76.54, 3170.78, respectively, and for male are 2057.00, 2521.28, 1628.72 (I think in the model expression, the sign of β should be $-$).

(d) The null hypothesis is the proportional odds assumption is valid. We used

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{3} n_{ij} \log(n_{ij}/\hat{n}_{ij}) = 0.3891.$$

Based on 1 degree of freedom, we accept H_0 because $G^2 < \chi^2_{0.05,1} = 3.84$.

6. (a) The survival function is

$$S(t) = e^{-\int_0^t \lambda(u)du} = e^{-\int_0^t 2u/\beta^2 du} = e^{-t^2/\beta^2}.$$

(b) The PDF is

$$f(t) = h(t)S(t) = \frac{2te^{-t^2/\beta^2}}{\beta^2}.$$

Thus, the loglikelihood function is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} \log f^{\delta_i}(t_i) S^{1-\delta_i}(t_i) \\ &= \sum_{i=1}^{n} \delta_i (\log 2t_i - 2\log\beta) - \frac{1}{\beta^2}\sum_{i=1}^{n} t_i^2\end{aligned}$$

Then,

$$\ell'(\beta) = -\frac{2}{\beta}\sum_{i=1}^{n}\delta_i + \frac{2}{\beta^3}\sum_{i=1}^{n} t_i^2$$

which implies

$$\hat{\beta} = \sqrt{\frac{\sum_{i=1}^{n} t_i^2}{\sum_{i=1}^{n}\delta_i}}$$

Note that $E(T^2) = \beta^2$ and

$$\ell''(\beta) = \frac{2}{\beta^2}\sum_{i=1}^{n}\delta_i - \frac{6}{\beta^4}\sum_{i=1}^{n} t_i^2.$$

Then, the Fisher Information is

$$I(\beta) = -E[\ell''(\beta)] = \frac{6 - 2\bar{\delta}}{\beta^2},$$

which implies

$$\hat{V}(\hat{\beta}) = \frac{\hat{\beta}^2}{n(6 - 2\bar{\delta})}.$$

Therefore, we have

$$\hat{\beta} = 236.04$$

and

$$s(\hat{\beta}) = 39.34.$$