QUALIFYING EXAM SOLUTIONS

Statistical Methods

January, 2009

1. (a) The 95% confidence interval for μ is

$$\bar{y} \pm t_{0.025,99} s_y/\sqrt{100} = 1.36 \pm 1.98 \times 1.2/10 = [1.1224, 1.5976].$$

(b) The 95% prediction intervale for Y_{new} is

$$\hat{\mu} \pm t_{0.025,99}\sqrt{s_y^2(1+1/100)} = 1.36 \pm 1.98 \times 1.2 \times 1.005 = [-1.0279, 3.7479].$$

(c) The 95% confidence interval is

$$[e^{-1.0279}, e^{3.7479}] = [0.3578, 42.4319].$$

(d) Because $E(X_i) \neq e^{E(Y_i)}$.

(e) We can use the continuous mapping theorem if the sample size is large.

(f) We can used (1) the normal probability plot, and (2) a Pearson χ^2 testing method.

2. (a) The model is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 XZ + \beta_5 X^2 Z + \epsilon,$$

where $\beta_0, \cdots, \beta_5$ are unknown parameters, and $\epsilon \overset{iid}{} N(0, \sigma^2)$ is the error term.

(b) In i, we test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ versus H_1 : one of β_3, β_4 and β_5 is not zero. Then, the model in (a) reduces to

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon.$$

Let SSE_F be the SSE of the Model in (a) and SSE_R be the SSE of the model above. Then, we can define a F-statistic

$$F^* = \frac{(SSE_R - SSE_F)/3}{MSE_F}$$

which can used used. It follows F_{3,df_F} under H_0, where df_F is the residual degree of freedom of the model in (a). We reject H_0 if $F^* > F_{0.05,3,df_F}$.

In ii, we can consider two models: (1)

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 XZ + \beta_5 X^2 Z + \epsilon.$$

Then, under $H_0 : \beta_4 = \beta_5 = 0$ versus H_1 : one of β_4 or β_5 is not zero. Then under H_0, (1) becomes (2) as

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_4 XZ + \beta_5 X^2 Z + \epsilon.$$

Let SSE_1 be the SSE of (1) and SSE_2 be the SSE of (2). We can use

$$F^* = \frac{(SSE_1 - SSE_2)/2}{MSE_1} \sim^{H_0} F_{2,df_1}$$

where df_1 is the residual degree of freedom of Model (1). We reject H_0 if $F^* > F_{0.05,2,df_F}$.

(c) The assumption is error term must be iid normally distributed. If n_1 and n_2 are large, then normal assumption can be released.

3. (a) The model is

$$Y = \mu + \alpha_i + \gamma_j + \epsilon_{ij}, i = 1, 2, 3, j = 1, 2, 3, 4,$$

where α_i is the fixed effect with $\alpha_1 = 0$, $\gamma_j \sim^{iid} N(0, \sigma_\gamma^2)$ is a random effect, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the error term. We assume γ_j and ϵ_{ij} are independent. The estimates are $\hat{\mu} = 349.00$, $\hat{\alpha}_1 = 0$, $\hat{\alpha}_2 = -6.25$, $\hat{\alpha}_3 = -42.5$, $\hat{\sigma}_\gamma = 11.39932$, and $\hat{\sigma} = 19.67020$.

(b) REML method is equivalent to the ANOVA method. Using it to fit model, the fixed effect is removed in the estimation procedre of the random effect. Therefore, fixed effect and random effect can be estimated separately.

(c) The predicted value of a new block is 0. Thus, the predicted value of the new pen is $349 - 6.25 = 342.75$.

(d) We need to consider a twice loglikelihood ratio statistic based on the "ML" method. Then, the value is $-52.44424 - (-56.65651) = 4.21$. Based on the χ_2^2 distribution with $\chi_{0.05,2}^2 = 5.99$, we accept the null hypothesis and conclude $\alpha_1 = \alpha_2 = \alpha_3 = 0$. We can use a bootstrap method to imporve the p-value.

(e) This must be a bootstrap method. The method is: (i) fit the model without block and estimate model parameters; (ii) generate data from the estimated model: (iii) compute the twice loglikelihood ratio statistic bewteen the model with and without block as a random effect. Repeat (ii) and (iii) many times and derive the simulated distribution of the test statistic, then the p-value is derived by the upper quantile value.

4. (a) The OLS estimator of b is

$$\hat{b} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n bX_i^2 + \sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2} = b^2 + \frac{\sum_{i=1}^n X_i \epsilon_i}{\sum_{i=1}^n X_i^2}.$$

Assume X_i is also a random variable and correlated with ϵ_i. Suppose n is large, then $\sum_{i=1}^n X_i^2 \to E(X_i^2)$ and $\sum_{i=1}^n X_i \epsilon_i / n \to E(X_i \epsilon_i)$ which is not zero. Thus, $\hat{b}$ is a biased estimator of b.

(b) In this case,

$$\hat{X}_i = Z_i(\frac{\sum_{i=1}^n X_j Z_j}{\sum_{j=1}^n Z_j^2})$$

and

$$\begin{aligned}\hat{c} =& \frac{\sum_{i=1}^n \hat{X}_i Y_i}{\sum_{i=1}^n \hat{X}_i^2} = \frac{\sum_{i=1}^n Z_i Y_i (\sum_{i=1}^n X_j Z_j / \sum_{j=1}^n Z_j^2)}{\sum_{i=1}^n Z_i^2 (\sum_{j=1}^n X_j Z_j / \sum_{j=1}^n Z_j^2)^2} = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n X_j Z_j} \\ =& \frac{\sum_{i=1}^n Z_i (bX_i + \epsilon_i)}{\sum_{i=1}^n X_j Z_j} \\ =& b + \frac{\sum_{i=1}^n Z_i \epsilon_i}{\sum_{i=1}^n X_j Z_j}.\end{aligned}$$

CLearly, we have $E(\hat{c}) = b$.

5. (a) The fitted model is

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x = -8.27957 + 2.75977x.$$

It means when LBM increases one unit, MUSCLE increases 2.75977 units. When LBM is 30, MUSCLE is 74.51353. This is an approximate relationship for LBM between 30 and 60.

(b) The ANOVA table is

Source	DF	SS	MS	F-Value	P-value
Model	1	26519	29619	64.31	
Error	60	27635	460.58		
Total	61	54154			

(c) It tests

$$H_0 : \beta_1 = 3.75 \leftrightarrow H_1 : \beta_1 \neq 3.75.$$

The test statistic is

$$T = \frac{2.75955 - 3.75}{0.36373} = -2.72$$

which is greater than $t_{0.025,60} = 2.003$ in absolute value. Therefore, we rejecte H_0 and conclude $\beta_1 \neq 3.75$. If we use 0.01, level, then $t_{0.005,60} = 2.66$. We have the same conclusion.

(d) The model assumes

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

where $\epsilon_i \sim^{iid} N(0, \sigma^2)$. We can use the residual plot (residual versus fitted value) to diagnose the assumption.

(e) The 95% confidence interval is

$$119.1197 \pm t_{0.025,60} \times \sqrt{3.1085^2 + 460.58} = [75.68, 162.55].$$

(f) This reason is because gender is not considered in the model. The next step is to fit gender in the model.

(g) Let H_0 be the model without gender and H_1 be the model with gender. We use an F-statistic to assess H_0, where the test statistic is

$$F^* = \frac{(30186 - 26519)/2}{413.22} \sim^{H_0} F_{2,58}.$$

We have $F^* = 4.437 > F_{0.05,2,58} = 3.16$. Thus, we reject H_0 and conclude the model without gender is not reasonable.

(h) The correlation between gender and gender:LBM inflates the estimates of their variances, which increases their p-value. But the test in (5g) does not have such a problem. Therefore, we trust the test in (5g).

6. (a) We can fit a model

$$brun = \beta_0 + \beta_1 \log(area) + \text{all the rest main effects} + N(0, \sigma^2).$$

(b) We can use a sin function of month such as

$$\sin \frac{(month - 4)\pi}{6}.$$

This function has a 12-month cycle with zero attained at 4 and 10, where the two numbers may be changed a little bit but usually not a lot.

(c) The model uses counts instead of the original burn areas. Therefore, the area burned has probabilities as $\pi_1 = P(area = 0)$, $\pi_2 = P(area \in (0, 1])$, $\pi_3 = P(area \in (1, 10])$, $\pi_4 = P(area \in (10, 100])$, and $\pi_5 = P(area > 100)$, with $\sum_{i=1}^{5} \pi_i = 0$. A multinomial model then can be used.

(d) The first model excludes small fires, bt the second considers it. However, the second order ignores the difference of moderate fires by the aggregation.