PhD Qualifying Examination–Part II

Department of Statistics
University of South Carolina
May 25, 2021 - 9:00AM–11:00AM

READ FIRST THESE INSTRUCTIONS

1. DO NOT write your name on any of your answer sheets. Instead, write your pre-assigned codename.
2. There are two (2) problems on this examination.
3. You are **not allowed** to use search engines during the examination. But you may use the HELP manuals of the statistical packages that you use. Please adhere to the HONOR CODE in this instance. Any violation of the HONOR CODE (such as using search engines) will lead to a zero for the exam.
4. SAS OnDemand can be accessed through the following link https://odamid.oda.sas.com.
5. You have two hours for this examination. Both problems will be graded and are of equal weight.

The Problems

1. An experiment was conducted to determine the effects of various factors on the breaking strength of chromium alloys. Three factors were considered in the experiment: The type of alloy (the levels of which were the three major chromium alloys: Chromium hydride (Level 1); Nichrome (2); and Ferrochrome (3)); the temperature at which the breaking point was measured (the levels of which were High (1), Medium (2), and Low(3)); and the technician who performed operation of breaking the alloy (there were three technicians who were chosen from those working at the company that funded the study). At each combination of the three factor levels, four alloy specimens were pressured until they broke, and the breaking strength of the specimen was measured (in pounds per square inch, or psi). The model equation and part of the ANOVA table, including the Expected Mean Squares, are given below for this model, assuming the usual distributional assumptions are valid. The data set, `breakingstrength.txt`, is on Blackboard. Use $\alpha = 0.05$ for all hypothesis tests in this question.

$$Y_{ijkl} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

where $i = 1, \ldots, 3, j = 1, \ldots, 3, k = 1, \ldots, 3, l = 1, \ldots, 4$.

Source	**E(MS)**
A	$\frac{(3)(3)(4)}{3-1}\sum_{i=1}^{3}\alpha_i^2 + (3)(4)\sigma^2_{\alpha\gamma} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
B	$\frac{(3)(3)(4)}{3-1}\sum_{j=1}^{3}\beta_j^2 + (3)(4)\sigma^2_{\beta\gamma} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
C	$(3)(3)(4)\sigma^2_{\gamma} + (3)(4)\sigma^2_{\alpha\gamma} + (3)(4)\sigma^2_{\beta\gamma} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
$A \times B$	$\frac{(4)(3)}{(3-1)(3-1)}\sum_{i=1}^{3}\sum_{j=1}^{3}(\alpha\beta)^2_{ij} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
$A \times C$	$(3)(4)\sigma^2_{\alpha\gamma} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
$B \times C$	$(3)(4)\sigma^2_{\beta\gamma} + 4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
$A \times B \times C$	$4\sigma^2_{\alpha\beta\gamma} + \sigma^2$
Error	σ^2

(a) The effects of factor A are the α_i's; the effects of B are the β_j's; and the effects of C are the γ_k's. Based on the ANOVA table information and the description of the three factors, explain why the technician factor should be designated as factor C.

(b) Answer this BEFORE analyzing any of the data. If it happened that there was interaction between alloy type and temperature, interpret exactly what this would imply, in the context of the variables in this problem.

(c) Fit this ANOVA model using the data given. Perform all appropriate model diagnostics and take any remedial actions needed in the model fitting. Report your conclusions about the adequacy of the final ANOVA model.

(d) Once you are satisfied with the goodness of your model, perform the test for significant three-way interaction effects. Report the test statistic value (and its associated degrees of freedom) and your conclusion, providing numerical evidence to justify your conclusion.

(e) Perform the tests for each of the two-way interaction effects. Report each test statistic value (and its associated degrees of freedom) and your conclusions, providing numerical evidence to justify your conclusions.

(f) Perform the test for the effect of alloy type on breaking strength. Report your test statistic value (and its associated degrees of freedom) and your conclusion, providing numerical evidence to justify your conclusion. If you find a significant effect of alloy type, report in detail on which types are significantly different from each other, in terms of mean breaking strength.

(g) Perform the test for the effect of temperature on breaking strength. Report your test statistic value and your conclusion, providing numerical evidence to justify your conclusion. Of particular interest to the researchers was to determine whether the Medium Temperature yielded a different mean breaking strength than the other temperatures. Formally answer this question and report your results.

(h) The test for the effects of Technician is more complicated. To derive a test statistic for this test, use the following result about Satterthwaite's approximation: Under the usual ANOVA model assumptions, if $L = \sum_i c_i MS_i$ is a linear combination of mean squares associated with random effects in a balanced mixed model, then L can be used as the denominator of an F-test statistic, if it has the appropriate expected value $E(L) = \sum_i c_i E(MS_i)$. The approximate degrees of freedom associated with this linear combination L is

$$\nu = \frac{(\sum_i c_i MS_i)^2}{\sum_i c_i^2 MS_i^2 / df(MS_i)}.$$

Suggest a test statistic to test for the effects of Technician, give the approximate degrees of freedom for this statistic, and report the test statistic value for this data set.

(i) Give a formula for an unbiased estimator of $\sigma^2_{\beta\gamma}$.

2. Abalones are pearl-producing sea snails with colorful shells commonly used in jewelry. To study their growth pattern, physical measuremnts were collected on 958 **independent** female abalones from **the same population**. We consider two physical measurements in the analysis:

   - Diameter: measured perpendicular to length (mm)
   - LongestShell: Longest shell measurement (μm)

   The data set, `abalone2.txt`, is on Blackboard. The first 5 observations of the data are:

```
dat<-read.table("abalone2.txt", header=T, stringsAsFactor=F)
kable(dat[1:5,])
```

Diameter	LongestShell
0.4140	530
0.4262	545
0.4416	550
0.4047	535
0.4387	565

   (a) (i) Write down a simple linear regression model using **Diameter** to predict the length of the longest shell (**LongestShell**). Describe the least squares method to obtain the estimates for the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$. Write down the normal equations and the least squares estimates.

   (ii) Report the values of the estimates and the corresponding 95% confidence intervals for $\boldsymbol{\beta}$ using the data provided in the exam. Interpret the 95% confidence interval for the slope in the context of this problem.

   (b) Randomly split the data into two equal-sized training set (x_i, y_i) and testing set $(x_i^*, y_i^*), i = 1, ..., n$. Use the model described in (a), and let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^{\mathrm{T}}$ be the least square estimates using the training set. Define the training set error for $\widehat{\boldsymbol{\beta}}$ as

$$R_{train}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum \left\{ y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right\}^2,$$

   and define the test error as

$$R_{test}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum \left\{ y_i^* - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i^*) \right\}^2.$$

   Calculate $R_{train}(\widehat{\boldsymbol{\beta}})$ and $R_{test}(\widehat{\boldsymbol{\beta}})$. Repeat (b) 1,000 times and report the mean of $R_{train}(\widehat{\boldsymbol{\beta}})$ and $R_{test}(\widehat{\boldsymbol{\beta}})$ from 1,000 iterations.

   (c) Prove that, on average, the training-set error is less than or equal to the test-set error, using the $\widehat{\boldsymbol{\beta}}$ estimated from the training set. That is, show that

$$E\{R_{train}(\widehat{\boldsymbol{\beta}})\} \leq E\{R_{train}(\boldsymbol{\beta})\} = E\{R_{test}(\boldsymbol{\beta})\} \leq E\{R_{test}(\widehat{\boldsymbol{\beta}})\},$$

   where $\boldsymbol{\beta} = E(\widehat{\boldsymbol{\beta}})$.