Biostats VHM 802 Course
Winter 2007, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Question 2 of final exam

Question 2.

We use the following notation,

> y_{ijkl} = amylase specific activity measured for the l'th replication of analysis temperature i, growth temperature j, and variety k,

where $i = 1, \ldots, 8$ ($\sim$ analysis temperatures $t_i = 10$, 13, 15, 20, 25, 30, 35, 40 °C); $j = 1, 2$ ($\sim$ growth temperature 13, 25 °C); $k = 1, 2$ ($\sim$ varieties B73, O43); and $l = 1, \ldots, 3$ ($\sim$ replications).

A)

The experiment was carried out as a completely randomised design, with 3 replicates for each of the 32 experimental conditions. The design can also be described as a $8 \times 2 \times 2$ factorial with 3 replicates. The design is balanced and complete.

The statistical model corresponds to a full factorial with all main effects of and interactions between the three factors. The model equation takes the form,

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl},$$

with $\varepsilon_{ijkl} \sim \mathrm{N}(0, \sigma^2)$.

The ANOVA table shows a non-significant three-factor interaction ($P = 0.22$), two non-significant two-factor interactions ($P = 0.15$ and $P = 0.97$) and one clearly significant ($P < 0.0005$) two-factor interaction, namely of `variety*temp_grw`. The main effects of the factors involved in the interaction are of less interest, but the main effect of `temp_anl` is clearly significant as well ($P < 0.0005$). It can therefore be said that all three factors have a clearly significant effect on the amylase specific activity (in the sequel abbreviated as ASA).

B)

The two effects of interest are the main effect of `temp_anl` and the interaction `variety*temp_grw`. Consider first the interaction. The interaction plot shows that variety B73 has higher ASA at growth temperature 25 °C than 13 °C, whereas variety O43 has the highest ASA at 13 °C. In addition, the ASA of variety B73 is higher than that of variety O43 at both temperatures. The table of means for the 4 combined levels of variety and temperature shows this pattern as well. In order to assess which differences among the levels are statistically significant, we compute LSD and BSD values (the latter corrected for a total of $4 \cdot 3/2 = 6$ comparisons).

$$\begin{aligned} \mathrm{LSD} &= t_{.975}(64)\sqrt{\mathrm{MSE}}\sqrt{2/24} \approx 14.63, \\ \mathrm{BSD} &= t_{1-0.025/6}(64)\sqrt{\mathrm{MSE}}\sqrt{2/24} \approx 19.46, \end{aligned}$$

where $t_{.975}(64) \approx t_{.975}(60) = 2.00$ and $t_{1-0.025/6}(64) = t_{.996}(64) \approx t_{.995}(60) = 2.66$ (best approximations possible from the textbook table). This shows that at an overall (or simultaneous) error level of 0.05 there is statistical evidence to declare all comparisons significantly different, except the one

between the two growth temperatures for variety O43. This comparison is not significantly different at an individual error level of 0.05 (because $320.8 - 306.6 = 14.2 < 14.63$) but close to significant.

Consider now the effect of analysis temperature. The interaction plot (and the table of means) shows that ASA increases with the temperature up till about 30 °C, and then drops again. We compute again the LSD and BSD values to assess statistical significance; here the BSD value accounts for $8 \cdot 7/2 = 28$ comparisons.

$$\begin{aligned} \text{LSD} &= t_{.975}(64)\sqrt{\text{MSE}}\sqrt{2/12} \approx 20.69, \\ \text{BSD} &= t_{1-0.025/28}(64)\sqrt{\text{MSE}}\sqrt{2/12} \approx 33.41, \end{aligned}$$

where $t_{.975}(64) \approx 2.00$ as before, and $t_{1-0.025/28}(64) = t_{.9991}(64) \approx t_{.999}(60) = 3.23$ (best approximation possible from the textbook table). At an individual error level of 0.05, all "adjacent" temperature comparisons are statistically significant, except for 25 and 30 °C. This is no longer true with the Bonferroni corrected comparisons, and the resulting pattern is most easily represented by a letter coding scheme,

$$10^{a} \quad 13^{ab} \quad 40^{b} \quad 15^{b} \quad 20^{c} \quad 35^{cd} \quad 25^{de} \quad 30^{e}\,.$$

Even at a simultaneous error level of 0.05, there is statistical evidence to declare all temperatures except 25 °C "inferior" to 30 °C.

C)

The statistical models analysed for part C) contain an interaction term between variety and growth temperature as well as different functions of analysis temperature,

$$y_{ijkl} = \mu + \alpha_j + \gamma_k + (\alpha\gamma)_{jk} + \beta_1 t_i + \ldots + \beta_r t_i^r + \varepsilon_{ijkl},$$

where r is the highest polynomial order for $t_i \sim$ `temp_anl`. The four models are nested within each other, and they are all nested within a model with the interaction and a categorical effect of `temp_anl`. No listing is shown for this model but its sum of squares and degrees of freedom for error can be obtained by adding the values for the removed terms in the full model. The following tables gives a summary of each of the five models.

Model for `temp_anl`	SSE	DFE	R^2	R^2 (adj.)	P for highest order term
linear	293339	91	36.1	33.3	< 0.0005
quadratic	67846	90	85.2	84.4	< 0.0005
cubic	57814	89	87.4	86.6	< 0.0005
quartic	56397	88	87.7	86.7	0.14
categorical	55673	85	87.9	86.4	n/a

The table shows that among the polynomial models, the linear model is very poor (not too surprising in view of the plots), the quadratic model is a great improvement, and that the cubic and quartic models add further predictive ability to the model. As the coefficient for the highest order term in the cubic model is clearly significant, the best model needs to be of at least cubic order. The fourth order term is not significant ($P = 0.14$), so the cubic order model may be argued as the preferable one (one could also argue this non-significance to be irrelevant when the purpose is prediction). In order to test whether the cubic order model gives an acceptable fit compared to the categorical model, we compute the F-statistic:

$$F = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})}{\text{SSE}_{\text{full}}/\text{DFE}_{\text{full}}} = \frac{(57814 - 55673)/(89 - 85)}{55673/85} = 0.82,$$

which is clearly non-significant in a $F(4, 85)$-distribution. Therefore, there is no indication of lack of fit for the cubic regression model. Finally, we can predict the ASA for analysis temperature 28 °C at an "average" variety and growth temperature from the coefficients in the Minitab listing,

$$\hat{y} = 148.44 + 5.584 \cdot 28 + 0.5765 \cdot 28^2 - 0.015510 \cdot 28^3 = 416.3\,.$$

When using the Stata listing and omitting the terms for variety and growth temperature, the prediction is for variety B73 and growth temperature 13,

$$\hat{y} = 160.22 + 5.584 \cdot 28 + 0.5765 \cdot 28^2 - 0.015510 \cdot 28^3 = 428.1\,.$$