Biostats VHM 802 Course
Winter 2015, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to final exam

The solution is more detailed (and verbose) than required for a 100% mark. It includes all questions (1–3), where Question 3 was answered only by students taking the "full" (3 credit) VHM 802 course.

A)

The design is a block design with judges corresponding to blocks. We could also say that the assessments by each judge are repeated measures, but it seems more natural to consider the judges as a (necessary) grouping of observations; hence, judges would be a blocking factor because there is no real interest in judge effects. The experimental (and measurement) unit is a solution. There are four treatments factors: the different concentrations of the four compounds A−D, for a total of 16 different treatments. As each judge can only assess four solutions, the design is an incomplete block design. We will further discuss the block design in E). The natural statistical model would include factorial effects of the compounds and their interactions, possibly of all orders, which would correspond to the 15 degrees of freedoms for treatments, plus additive block effects. The model used in the listings did not include all interactions (for reasons discussed later in this solution), and for simplicity we describe this model. We use the following notation,

> y_{ijklm} = threshold concentration determined by judge m for the solution determined by compound levels i, j, k and l for compounds A−D, respectively,

where $i, j, k, l = 1, 2 \sim$ low/high; $m = 1, \ldots, 8 \sim$ judges.

$$y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + \theta_m + \varepsilon_{ijklm},$$

where

- ε_{ijklm}'s are the errors (between solutions), assumed i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$,
- α_i's, β_j's, γ_k's, δ_l's are the main effects for compounds A−D, respectively,
- $(\alpha\beta)_{ij}$'s, $(\alpha\gamma)_{ik}$'s etc. are two-factor interaction terms between the compounds,
- θ_m's are judge effects, which can be taken either as fixed effects or as random effects, in which case they are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma_J^2)$.

B)

The model analysed use the logarithmic threshold concentration as the outcome, and included all two-factor interactions between compounds except A∗D and no higher-order interactions. The sequential and partial sum of squares of the ANOVA table are identical, meaning that we can draw conclusions about all effects without removing non-significant terms. The R^2 is high, so the model explains a very large proportion of the variance. The F-tests show that only one 2-factor interaction is significant, namely A∗B at $P = 0.044$. In addition, the main effect for C is clearly non-significant ($P = 0.47$), and the main effect for D is clearly significant ($P = 0.001$). Finally, the judges have a clearly significant effect ($P < 0.0005$) and this effect explains approximately 20% of the sum of squares, so the use of a block design seems to have been a good idea (by reducing the unexplained variation). As the means for judges are not included in the listing, we will assume these to be of minor interest. For the other factors, we have the following interpretations:

– compound C has no significant impact, with estimates ± standard errors (SEs) at the two levels of 6.90 ± 0.21 and 6.68 ± 0.21;

– compound D has a strongly significant impact, with estimates of 7.43 and 6.16 for low and high levels, respectively, and thus lower threshold concentration at high levels of D. We can compute the margin of error for 95% CIs by multiplying the SE by $t^*(15) = 2.131$, giving $0.21 \cdot 2.131 = 0.45$;

– compounds A and B are involved in a significant interaction and will therefore need to be considered together, and it might be useful to sketch an interaction plot based on the four means of A∗B. The four means show lower threshold concentrations for high than low levels of both compounds, regardless of the level of the other compound. It is seen that the combined effect of having both compounds at high level is stronger than the added effects at low levels of the other compound (i.e., $3.08 < 9.85 + (7.34 - 9.85) + (6.89 - 9.85)$); thus, we can say the interaction effect is synergistic. In order to do (unadjusted) pairwise comparisons, we compute $\text{LSD}_{.95} = t^* \cdot \sqrt{2} \cdot 0.2968 = 0.89$. It is seen that the difference between low and high levels of either compound is significant, regardless of the level of the other compound.

In summary, the lowest threshold concentration would be obtained with high levels of compounds A, B and D, whereas compound C has no noteworthy impact.

C)

The wide range of values of the outcome, from values fairly close to zero (lowest value: 4) to values in the tens of thousands (highest value: 90 293) suggests that it would be very difficult to meet the assumption of constant variance on original scale. This is because values close to zero (the concentrations cannot be negative) will inevitable have less variation than values at the other end of the scale. Generally speaking, when the values of the outcome span several orders of magnitude it is (almost) always necessary to carry out a transformation. Whether this assumption is correct for the present data, could be explored by carrying out a Box-Cox analysis.

D)

On purpose, this incomplete block design was constructed to not allow estimation of A∗D but instead allow balanced estimation of all other two-factor interactions (such designs are beyond the scope of the course but discussed in Chapter 15 of the textbook). The way this can be seen in the experimental design is that the interaction A∗D varies at the block (judge) level, not within blocks. For example, the first judge only assesses two combinations of A×D, not all four combinations; this applies to all judges. Therefore, there is no information about the interaction A∗D within any of the blocks, and because the analysis has fixed block effects we can not estimate A∗D between blocks. This is the same situation as when we try to estimate a herd-level factor in a two-level model with fixed herd effects. It follows that we could get estimates and tests for the interaction by taking random effects for judges. These statistics would however be rather weak because of the low number of judges, and would rely on the judges representing a population. The idea of the experimental design is to in advance give up on estimating certain effects in return for getting better estimates of the other effects.

E)

With 3 compounds there are $2^3 = 8$ solutions, so the question is about constructing an incomplete block design with 8 (g) treatments and block sizes of 4 (k). The only incomplete block design covered in the course that could apply to this situation is a balanced incomplete block design (BIBD), where

balancedness means that any pair of treatment combinations "meet" the same time within a block. A selection of such designs is given in Appendix C.2 of the textbook, and the listing includes a design with 14 (b) blocks for a total of 56 observations (BIBD 18). This meets the requirement for a not too high number of judges. The main advantage of the design is that it would allow estimation of the full factorial formed by the 3 compounds, with equal precision for all pairwise comparisons.

Question 2.

We use the following notation,

$$\begin{aligned} y_{ij} &= \text{RBC survival (\%) for } j\text{th analysis of sample from mouse } i, \\ \text{group}(i) &= \text{group for } i\text{th mouse}, \\ x_j &= \text{ionic concentration for } j\text{th measurement within each mouse}, \end{aligned}$$

where $i = 1, \ldots, m = 30$ and $j = 1, \ldots, n = 6$ with $(x_1, x_2, \ldots, x_6) = (0.30, 0.35, 0.40, 0.45, 0.50, 0.55)$.

A)

The data structure is longitudinal, or repeated measures on the same mouse across the ionic concentrations. Note that even if the repeated measures are in this case not over time, we would still consider the ionic concentrations as an ordering that is similar to an ordering across time. For example, RBC values at close ionic concentrations would be expected to be more similar than values at concentrations further apart. The data structure could also be viewed as hierarchical, with RBC values within mice, although this is a simplified view because it ignores the ordering within mice. The experimental design for the mice is a one-way design with 3 groups and replicates within each group. The study description does not allow us to determine whether the mice were randomized onto the groups or the groups represent inherently different mice (e.g., different strains).

The provided graphs are profile and mean plots. All plots show a dose-response relationship with ionic concentration whereby the survival is initially (at low concentrations) high and drops down in mid-range concentrations of $45 - 50$ to values near zero at a concentration of 55. The relation appears to be clearly non-linear. The curves have a similar shape for the 3 mice groups, but the mean plot seems to indicate that the drop in survival happens at somewhat later concentrations for group 1 than the other groups. The 3 groups seem to have similar responses at both the lowest and highest ionic concentrations. The profile plots show that RBC values are quite consistent across mice within a group, with some exceptions (in particular, two values for conc. 45 in group 3). It is difficult to tell whether mice generally tend to be low or high, presumably this means that the within-mice correlations are not very strong. The profile plots (and also the partial data listing) show that the values at conc. 55 are all very similar and low; in fact almost all of the values equal 0, corresponding to no surviving red blood cells.

B)

The analysis carried out corresponds to a hierarchical or split-plot analysis for repeated measures with random subject (mouse) effects. The statistical model can be written,

$$y_{ij} = \mu + \alpha_{\text{group}(i)} + \beta_j + \alpha\beta_{\text{group}(i)j} + A_i + \varepsilon_{ij}\,, \quad A_i \sim N(0, \sigma_A^2),\ \varepsilon_{ij} \sim N(0, \sigma^2)\,.$$

The ANOVA table shows a strongly significant group by concentration effect, in addition to (less important) significant effects of groups and concentrations. From the plots it is no surprise that the strongest effect is of the 6 concentrations (modelled as categorical). Note that the Stata listing

contains Wald tests of parameters for groups, concentrations and the interaction; due to the clearly significant interaction, only the Wald test for the interaction is valid as a test for overall effects. The estimated between-mouse variation is fairly small, and the within-mouse ICC is estimated at $3.856/(3.856 + 38.943) = 0.09$. The Stata listing gives a likelihood-ratio test for the mouse random effects with the corresponding $P = 0.05$, but even if this is borderline significant we have no intention of removing the mouse random effects (because it would assume measurements on the same mice to be independent which seems very unrealistic). The corresponding F-test in the ANOVA table has $P = 0.044$. The ANOVA table also gives $R^2 = 98.2\%$ — a very high value, reflecting that there is little variation about the mean response curves.

Interpretations of the estimated effects should be based on the mean plot (= the interaction plot for group×conc.). These means are listed as least squares means (Minitab). The main interest is most likely in comparing the groups at the different concentrations, and we would therefore need standard errors for the differences between group means. Because of the unequal numbers of mice in the 3 group, there will be different standard errors depending on which groups are compared. The comparison between groups involves different mice, and the variance term therefore involves both variance components, e.g. for groups 1 and 3,

$$\mathrm{SE}(\bar{y}_{1j} - \bar{y}_{3j}) = \sqrt{(\hat{\sigma}_A^2 + \hat{\sigma}^2) \cdot (2/11)} = \sqrt{(3.856 + 38.943) \cdot (2/11)} = 2.790.$$

This calculation was based on the two groups with 11 mice. It is seen that this standard error is also listed in the comparison against baseline for group 3 in the Stata listing. The standard error for comparisons of group 2 against any of the other groups is 3.034. It is seen that the group means are within one standard error at concentrations 30, 35, 40 and 55, where therefore no significant differences exist. At concentrations 45 and 50, the differences between group 1 and the two other groups are much larger than two standard errors, suggesting statistical signficance. In order to carry such pairwise comparison tests out exactly, one should do pairwise comparisons within the group by conc. interaction, e.g. using the `pwcompare` command in Stata. Adjustment for multiple comparisons could also be considered.

C)

The hierarchical model/analysis can be criticised on several points. One is the "usual" one that with a long series, the underlying assumption of within-mouse correlations that are independent of "time" is unrealistic and most likely violated by the data. Given that the within-mouse correlation was estimated to be fairly low, this may not have any major impact on the results and conclusions. A more serious critique is that the assumption of constant variances is violated, by a very small variance at conc. 55 and possibly by larger variances at the mid-range concentrations 45 and 50 than at the lower concentrations. Related to this, the residual plot has "edges" at both ends, corresponding to very many data values of 0% and 100%. One would definitely expect this to violate the equal variances and possibly also the normality assumption. It might be possible to transform the outcome by the standard (arc-sine square-root) transformation for proportion data, but even that may not be sufficient. Another idea is to drop the values at conc. 55 which do not give much information anyway (because they are almost all zero). If none of this leads to roughly homoscedastic data, a non-parametric analysis may be attempted; this was not covered in the course but has been described e.g. in Davis (2000), *Statistical Methods for the Analysis of Repeated Measurements*, Springer. Finally, there is one very extreme negative standardized residual of -9.17. Such an extreme value can never occur in a normal distribution, and therefore this observation (the very low value for mouse 5 in group 3 at conc. 45) should be inspected and probably be discarded.

D)

One possible way of analysing repeated measures data is by separate analyses at different "time points" (ionic concentrations). As discussed above, one would probably not include conc. 55 among those analysed, and the strongest interest seems to be in concentrations 45 and 50. The statistical model would be a one-way ANOVA with 3 groups, and if the model fails to meet the linear model assumptions it could be carried out as a non-parametric analysis using Kruskal-Wallis test. Another analytical approach is to construct response features from each curve, and some possibilities are the area under the curve (because of the equidistant concentrations, this would be equivalent to analysing means across concentrations) and the so-called LD50 value: the dose (concentration) at which 50% of the red blood cells die. Such LD50 values could be computed by linear interpolation (corrsponding to reading off the concentration where the lines in the profile plots cross 50%) or using a mathematical equation for how the response depends on the concentrations; a large number of such dose-response curves exist in the literature. Once a response feature has been calculated for each, the analysis would again consist of a one-way ANOVA (possibly non-parametric, as above).

Question 3.

A)

The shown analysis corresponds to a multiple linear regression model with 5 predictors which can be written as,

$$\texttt{ES}_i = \beta_0 + \beta_1 \texttt{area}_i + \beta_2 \texttt{anear}_i + \beta_3 \texttt{dist}_i + \beta_4 \texttt{distSC}_i + \beta_5 \texttt{elev}_i + \varepsilon_i,$$

where β_0 is the intercept, the other β's are regression coefficients corresponding to the respective predictors, and the ε_i's are errors assumed to be i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$. The interpretation of the parameters are as follows,

- β_0: number of endemic species for an island of area zero, elevation zero, and distances/areas to nearest islands zero as well; in short, a value with no meaningful interpretation itself, but part of the regression equation,
- β_1: the difference in number of endemic species between two island whose area differ by 1 km^2 but are otherwise equal,
- $\beta_2, \ldots, \beta_5$: similar to β_1, except that the difference between the two islands considered is not in area but in one of the other characteristics (`anear`,...,`elev`),
- σ: the standard deviation about the regression curve or of the unexplained variation in the model.

B)

The normal probability plot looks quite straight but there is one somewhat extreme residual, for island 13 (Isabela). Its deletion residual is -3.47, after Bonferroni correction this corresponds to $P = 2 \cdot 27 \cdot P(t_{25} < -3.47) \approx 54 \cdot 0.001 = 0.05$. There is some evidence for this island to be an outlier in the model (assuming the model assumptions are met). The residual plot however does not look good, with some cone/fan-shape and a "lower edge" for small fitted values (corresponding to a lower bound of zero on `ES`). The large negative residual just considered is also the observation with the largest predicted value, and the extreme residual is therefore likely a result of heteroscedasticity (and the observation should definitely not be excluded based on the outlier test). It seems likely that

a transformation (such as square-root or log transformation) would improve the residuals. It could be noted that a square-root transformation is often suggested for counts.

The model diagnostics show that the islands 9 (Fernandina) and 13 (Isabela) have very large leverages, and that Isabela in addition has a ridiculously large value for Cook's distance. This is caused by the large areas and elevations for the two islands, plus the fact that Isabela is the nearest island to Fernandina. The values for area span several orders of magnitude in the data (0.01–4669), and it is clearly inappropriate to have a linear effect for area in the model. To remove the two large islands from the dataset would only alleviate, not solve the problem, and it would make the inference much less interesting.

C)

We first address how to deal with the predictor `area` (and `anear`). In order to reduce the range of values it would be natural to transform `area` by e.g. the log transformation. This transformation would be particularly attractive if the outcome was log-transformed as well. Other transformations are possible, and one may try to fit different transformed variables together to see which of them has the stronger association with the outcome. The distance and elevation variables might also need a transformation, but this is less clear from the information presented. The model diagnostics should be rerun when suitable transformations for `ES` and `area` have been determined.

None of the correlations between the predictors are alarmingly large, but a correlation of 0.75 between `area` and `elev` is notable. However, transformation of `area` will change the correlations with this predictor, so the correlations should be recomputed when transformations have been determined. At this point there seems to be no need to exclude any predictors from the modelling on grounds of high correlations. On the other hand, inclusion of `NS` in the model would be meaningless because it is an outcome similar to `ES` (in fact, `NS`$\geq$`ES`). If one wanted to incorporate the total number of species into the analysis, it would be better to combine the two variables into a single one, e.g. the ratio `ES`/`NS`. Finally, the suggested approach to modelling is a backwards selection based on a model with all predictors (suitably transformed) at a significance level of 0.05. Interactions between the continuous predictors would be difficult to interpret and are therefore probably of limited interest.