Biostats VHM 802 Course
Winter 2025, 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 should be answered only by students taking the "full" (3 credit) VHM 802 course. Some minor parts were waived in the marking.

A)

We use the following notation,

> y_{ijkl} = biomass for the section of tank l exposed to insects of parent type i, hatching condition j and growth condition k,

where $i, j, k = 0, 1 \sim$ non-native/native plants; $l = 1, \ldots, 8 \sim$ tanks.

The design is a block design with tanks corresponding to blocks. There are four sections within each tank, so the design may be considered as hierarchical with sections within tanks, or it may simply be considered as a block design. The experimental (and measurement) unit is a section within a tank. There are three treatments factors, for a total of 8 different treatments. As each block holds only four sections, the design is an incomplete block design. There is a system to how the blocks were constructed, but the system is beyond the material covered in the course. The natural statistical model (and indeed the one shown in the listing) has a full factorial effect of treatments and additive blocks:

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

where

- ε_{ijkl}'s are the errors (between sections), assumed i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$,
- α_i's, β_j's, γ_k's are the main effects of parent, hatching and growth types, respectively,
- $(\alpha\beta)_{ij}$'s, $(\alpha\gamma)_{ik}$'s etc. are interaction terms between the treatment factors,
- δ_l's are tank effects, which can be taken either as fixed effects or as random effects, in which case they are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma_D^2)$.

B)

The ANOVA table shows that the 3-factor interaction is clearly non-significant, that only `hatch*growth` is significant among the interactions (and only weakly so), whereas the main effect for `parent` is non-significant, and (less importantly) that `hatch` and `growth` are both strongly significant on their own (but also involved in the significant interaction). There is also a strongly significant tank effect.

None of the effects involving `parent` are close to significant, so where the parents were reared seems to be unimportant for the ability of the insect to control growth of the non-native plant. Without further information about the tanks, we cannot use tank effects to predict effectiveness of growth control. It remains to explore the effects of where the insects were hatched and grown to maturity. Low biomass is the desirable outcome, and it is seen that the factor combination (0,0) produced the lowest growth. This means that if the insects were hatched and grown on the non-native plants, they would also be most effective for control of its growth. The most clearly visible feature of the

interaction between `hatch` and `growth` is a higher biomass for (1,1) than expected from additivity. That is, if the insects were hatched and grown on the native plants, they show very little appetite for the non-native plants. This does not help us for devising growth control. The pairwise comparisons for the interaction show that if the insects were hatched on non-native plants, there is no significant difference between being grown on the two plants, although the difference between the estimates for biomass is substantial (9.60 and 13.48). The P-value is shown as $P = 0.045$, but with a Bonferroni correction it becomes $6 \cdot 0.045 = 0.27$ and with a Holm correction it becomes $2 \cdot 0.045 = 0.090$. The conclusion is that the key experimental condition for growth control is that the insects should be hatched on non-native plants, and that growth on non-native plants may also be beneficial.

C)

The comparison of residual plots obtained for the analyses on untransformed and square-root transformed scales show no clear improvement in the latter. The normal plot is somewhat more straight (and we might want to get P-values from normality tests), but neither of the plots seem too concerning. The most extreme residual on either scale is not alarmingly large, and overall the residual plots on both scales look quite acceptable. It is suggested to carry out a Box-Cox analysis to determine the evidence in favour of a square-root transformation.

The analysis on square-root transformed scale no longer shows a significant `hatch*growth` interaction although it is still close ($P = 0.096$). This seems as a major difference between the two models (but interaction is scale-dependent). The change in conclusion for growth control is that the additive model would suggest the best experimental conditions to consist in *both* hatching and growth on non-native plants.

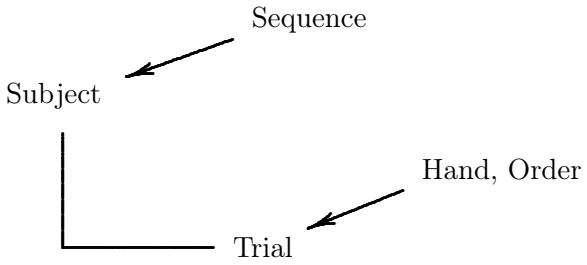
D)

In part A), we noted the design to be an incomplete block design. It is not a balanced incomplete block design (BIBD) because there are a total of $8 \cdot 7/2 = 28$ treatment pairs, and the design includes $8 \cdot 4 \cdot 3/2 = 48$ within-block pairs of treatments. Therefore, it is impossible for all treatment pairs to "meet" within a block the same number of times. In the pairwise comparisons for interactions, this can be seen by the standard errors for the differences equalling either 1.66 or 1.80. The standard errors for least squares means follow the usual rule for main effects ($0.832 = 3.32908/\sqrt{16}$) but not for interactions ($1.2251 > 3.32908/\sqrt{8}$); this is also a reflection of the design. In the ANOVA table, the sequential and adjusted sum of squares are identical for all effects except blocks, reflecting that blocks cannot be assessed independently of the treatments. The design has been constructed to partially confound the interactions with blocks (Chapter 15 of GO).

Question 2.

A)

A total of 24 measurements of fMRI pixel counts representing activated brain areas will result from this design, for two finger-tapping action trials for each of 12 subjects. If the hand used for finger-tapping is considered the "treatment" (which would make sense because a difference between left and right hands could be anticipated), the experimental unit (and measurement unit) is each trial where the finger-tapping occurs and the fMRI images are obtained. The variables of interest are: hand (left/right), order (first/second), subject ($1-12$) and possibly also the sequence in which the treatments were applied (left-right/right-left). The experiment is a cross-over design, and the data structure may be considered as hierarchical with two trials for each subject.



B)

Several models exist for cross-over designs, differing mainly in their assumptions about carry-over effects. In this case, carry-over effects may be assumed negligable, provided that a suitable pause ("wash-out period") is administered between the first and second trial for each subject. The treatment (hand-tapping) is not going to have any lasting impact on the subject. The model without carry-over effects is a hierarchical model corresponding to the 2-level structure above; note that without carry-over effects the actual sequence is unimportant once the treatment and order have been accounted for. The model can be formulated as follows for the response y_{ijk} obtained from subject i at order k with hand j ($i = 1, \ldots, 12$ and $j, k = 1, 2$):

$$y_{ijk} = \mu + \alpha_j + \beta_k + A_i + \varepsilon_{ijk},$$

where the subject random effects $A_1, \ldots, A_{12}$ are i.i.d. and $\sim \mathrm{N}(0, \sigma_A^2)$, and the errors ε_i's are i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$. The model parameters are,

- μ: overall mean,
- α_j's: hand effects,
- β_k's: order effects,
- σ_A^2: between-subjects variance,
- σ^2: between-trials (within-subjects) variance.

A hand by order interaction can be included in the model, but it requires likelihood-based estimation of the random effects model and therefore relies more strongly on the assumptions about the random effects. Sequence effects are captured by the effects of hand and order (and possibly their interaction) and should therefore not be included with separate terms. An alternative to the above model is analysis of suitable differences between the two measurements on each subject. Such analyses would only be advantageous if the assumptions of the above "full" model could not be met, e.g. due to heteroscedasticity.

C)

Generally speaking, sample size calculations can be based on either precision or power. In both cases, the statistic of interest to assess differences between hands is

$$\bar{y}_{\cdot 1 \cdot} - \bar{y}_{\cdot 2 \cdot} = \alpha_1 - \alpha_2 + \bar{\varepsilon}_{\cdot 1 \cdot} - \bar{\varepsilon}_{\cdot 2 \cdot}$$

Note that both the subject and order terms cancel, the latter due to the balancedness of the design. Therefore, we can do sample size and power calculations for the hand effect as if the design was two

independent samples, as long as we use the within-subject variance σ^2. For a power calculation, we need, in addition to this variance, the anticipated difference between groups and the actual sample size of each group (12). The assumptions are those of the model presented above, in particular the absence of both carry-over effects and an interaction between hand and order.

D)

The above model can be extended to include left-handed subjects if one also adds a fixed effect for whether a subject is left- or right-handed and its interaction with the actual hand activated. Subjects "handed-ness" will be a subject-level predictor, and hence the design has split-plot character with subjects as whole plots and trials as split plots, even if there is no randomization at the subject level. In order to compare differences between left- and right-handed responses between subjects of different "handed-ness", one will need to look at the interaction between the two factors. The usual tools will apply: an overall test for the interaction as well as pairwise comparisons among the four combinations for the two factors.

Question 3.

Denote by p_i the probability of household i switching to a new water source during the follow-up period, $i = 1, \ldots, 3020$. In terms of the binary event, y_i, we have $p_i = \mathrm{P}(y_i = 1)$.

A)

The model fitted is a logistic regression model for the binary outcome and with all four variables included as predictors. We can represent the model by the equation,

$$\mathrm{logit}(p_i) \quad = \quad \beta_0 + \beta_1 \, \texttt{arsenic}_i + \beta_2 \, \texttt{dist}_i + \beta_3 \, \texttt{assoc}_i + \beta_4 \, \texttt{educ}_i.$$

The predictor `assoc` is binary and coded as 0/1, and it can therefore be used as a dummy (indicator) variable for category 1; this means that β_3 represents the difference between the yes and no categories. The three other predictors are quantitative and modelled by their respective slopes. The intercept corresponds, on logit scale, to the probability for a household with an arsenic concentration of 0 (impossible because values were at least 0.5), a distance of 0 (also impossible), no membership of community organizations and 0 years of education. If it was of interest to interpret the intercept, the two first predictors would need to be centered at meaningful values.

Interpretations of the four predictor effects (in all cases assuming the other predictors held constant):

- arsenic: strongly significant ($P < 0.0005$), for a one-unit increase in arsenic levels (a meaningful increase), the odds of switching increases by a factor of 1.6: OR $= \exp(0.467) = 1.595$.
- dist: strongly significant ($P < 0.0005$), for a 50 m increase in distance to the nearest well (a 1 m increase is clearly too small to be meaningful), the odds of switching decreases by a factor of 0.64: OR $= \exp(-0.00896 \cdot 50) = 0.639$.
- assoc: not significant but somewhat close ($P = 0.106$), the odds of switching is lower by a factor of 0.88 if any household members were active in a community organization: OR $= \exp(-0.1243) = 0.88$; alternatively, the odds of switching is 13% higher if no members were active in a community organization.
- educ: strongly significant ($P < 0.0005$), for one extra year of education (a somewhat meaningful increase, 5 years could have been used as well), the odds of switching increases by a factor of 1.04: OR $= \exp(0.004245) = 1.045$.

B)

The model assumes independent observations (we have no information to question that, but it could be hypothesized that the 3020 households were located in multiple villages, and that decisions to switch were variable across villages — this would lead to a hierarchical data structure), additive effects of the different predictors and linear relationships between the quantitative predictors and the probability on logit scale (or log-odds). The two latter assumptions can be assessed to some extent by the additional information provided. For a start, the Hosmer-Lemeshow goodness-of-fit test is not significant and therefore does not demonstrate any problems with the model assumptions. It is, however, a fairly weak test, and its non-significance is by no means a proof that all model assumptions have been met. The Pearson goodness-of-fit test is inappropriate and useless by the lack of replication within covariate patterns.

Two approaches are shown to explore the linearity assumption for the three quantitative predictors: categorization by the `lintrend` command and addition of quadratic terms to the model equation. The categorization focuses on each predictor separately, but considering that the pairwise correlations between predictors are quite low, it may give a reasonable impression of the predictor effects in the multivariable model. The predictor with the strongest non-linear pattern seems to be `educ`, with a positive parabolic shape. The two other predictors show a negative parabolic shape, even if there seems to be one point outside the parabolic pattern for the second category of `dist`. The signs for the quadratic terms in the first extra logistic model match these impressions, but only the quadratic terms for `arsenic` and `educ` are significant (and quite strongly so). It therefore seems there is a potential to improve the model by adding non-linear terms, at least for these two predictors.

The second additional model fit includes multiple interaction terms between the predictors in the original model. One of these, between `dist` and `educ` is clearly significant, and another, between `arsenic` and `educ`, is not too far off, whereas the two remaining interaction coefficients are quite far from significance. These findings indicate that at least one and possibly several interactions would improve model fit. Some difficulties with this are that the quantitative effects are all assumed linear, contrasting the findings above, and that generally speaking interpretation of interactions between quantitative predictors is quite difficult. This would have to be addressed in further model building. In any case, the results do indicate that the assumed additivity between predictors is not unproblematic.

One major suggestion to further explore the model assumptions is to look at the residuals, even if it will have to be more or less individual observation residuals, which are generally less useful than residuals based on covariate patterns. Residuals could be plotted against predictor values with lowess curves overlaid to visualize the main patterns. Extreme residuals could also be explored, effectively corresponding to inspecting households that switched but were estimated to be little likely to do so, as well as households that did not switch but were estimated to be likely to do so.

C)

The discussion above showed that different model expansions should be considered. Models of different complexity can be compared by likelihood-ratio tests if they are submodels of each other or by a criterion such as the AIC. The main question for model expansion may however be how to combine the different model expansions.

The first step should be to decide about the functional relationships. It might be useful to additionally explore functional form with fractional polynomials; these have the advantage of being more flexible than simple quadratic terms but still apply to multifactorial models. In order to easier explore interactions, it would be beneficial to include as few terms as possible for each predictor, ideally only a single term. As a technical note, because `educ` takes the value 0, some value needs to be added to

make its values strictly positive.

Once functional form has been decided on, interaction terms can be included. It may be helpful to explore one interaction term at a time and perhaps restrict interactions to those with some biological justification and/or involving the strongest predictors. For predictors represented by a single term, the interaction is simply the cross-product. If such a cross-product is significant, it is strongly recommended to visualize it with predictive margins. It might also aid interpretation to categorize one of the two predictors involved and then explore the interaction between a quantitative and a categorical predictor.

For a relatively large dataset such as the present, model building by statistical significance and model fit (by AIC) may produce more complex models than desired. It can be suggested to also compute area under the receiver-operating curve and explore whether added terms substantially improve the model's predictive ability. Reliability is probably less of an issue with such a large dataset.

Generally speaking, the model building is likely to be manual. Automated model building procedures are not needed here to select between a large number of predictor effects, and the most important part is the construction of the effects to be considered (which automated procedures cannot do). It would also be relevant to involve causal considerations in the model building; some further information about the relationships between the variables would be helpful/necessary.