Biostats VHM 802 Course
Winter 2019, 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.

Question 1.

We use the following notation,

$$\begin{aligned} y_{ij} &= \text{tumour volume of } i\text{th mouse at } j\text{th time point} \\ \text{group}(i) &= \text{treatment group for } i\text{th mouse} \\ t_j &= \text{day (post day of injection) of } j\text{th measurement}\,, \end{aligned}$$

where $i = 1, \ldots, m = 30$ and $j = 1, \ldots, n = 7$ with $(t_1, t_2, t_3, t_4, t_5, t_6, t_7) = (7, 11, 12, 13, 14, 15, 17)$.

A)

The data structure is longitudinal, or repeated measures on the same mouse (= experimental unit) over time. The data structure could also be viewed as hierarchical, with volume measurements within mice, although this is a simplified view. The experimental design for the mice is a completely randomized one-way design, and mice should have been allocated randomly to the three treatment groups.

Both the profile and mean plots show an increasing trend over time in the tumour volumes (except for two mice in group C). So overall the tumours grow (not too surprisingly). Also the differences between mice increase over time; at day 7 all mice have tumours of about the same size, and at day 17 there is wide range between mice. There seem to be strong mice effects (and within-mice correlation) as mice tend to follow their own curves. The mean plot shows the average tumour size in the three groups to develop similarly over time, with group C lowest throughout. However, the profile plots show a clear difference between the groups with respect to the variation between mice. In group A all mice develop similarly, there is more variation in group B, and group C shows huge differences between mice with two mice having clearly less tumour growth than the rest. These differences in variation are not reflected in the mean plot.

B)

The shown analyses are separate analyses at each time point by one-way ANOVA methods. For example, the analysis for day 5 uses the model

$$y_{i1} = \mu_{\text{group}(i)} + \varepsilon_{i1},$$

with $\varepsilon_{i1} \sim \text{N}(0, \sigma^2)$. In order to obtain a valid (but weak) analysis, the separate analyses for each time point should involve a Bonferroni correction for the number of different analyses carried out. As all 7 time points are after treatment administration, it seems that one could not apriori exclude some of them, and therefore the Bonferroni correction should be for 7 analyses. This means that all P-values for tests comparing treatments should be multiplied by 7. The listing shows that all P-values were

non-significant even without the correction (the lowest one is 0.115 for day 14), so there is absolutely no evidence of difference between treatment groups in these analyses.

While we cannot assess the validity of assuming a normal distribution in the separate time analyses, the listing shows that the assumption of equal variances in the three groups may be seriously violated in many analyses. The shown Bartlett test for equal variances is significant for days 11,13,15,17; these are also the days for which the (IPS) rule of thumb that the ratios between standard deviations should not exceed 2 is violated. We already noted (in A) that one major difference between groups is the difference in variation; therefore, it seems inconsistent to analyse the data under the assumption of equal variances. An obvious alternative analysis is a non-parametric one-way ANOVA (Kruskal-Wallis), which is available directly from the menus in Minitab/Stata.

C)

A response feature is a statistic summarising the profile (set of values) for each subject (mouse). As the profiles are roughly linear, the feature immediately suggested is the slope over time for each mouse. To make the slopes more representative of the profile one might restrict the time period covered to days $11-17$; in this interval, the profiles are approximated well by a straight line, whereas the tumour growth rate up till day 11 for many mice seems to be different. On the other hand, a slope for days $11-17$ does not represent the entire profile and therefore potentially only captures a partial treatment effect. As a second response feature, one could consider the gain (difference) from initial to last measurement, although some problems exist. The initial values are pretty close and also quite small relative to the final tumour size. A gain computed over the entire span of the treatment (from day 5 to day 17) would have been more meaningful than from day 7 onwards, although one could argue the tumours to have grown only little from day 5 to 7. One could also argue that all tumours started at volume zero, and that a better statistic would simply be the final tumour volume. Less obvious response features are the curvature (because the curves do not show any substantial curvature) and the mean tumour volume across all time points (because it does not reflect the development over time and has no obvious biological interpretation).

The statistical analysis of any response feature would follow the same lines as the analysis for separate time points, except that no Bonferroni correction should be applied. In particular, the concerns about differences in variability would be the same, and a non-parametric analysis and an analysis for differences in variance could be performed in addition to a standard one-way ANOVA.

D)

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 strong day effect, but non-significant group and group×day effects. There is substantial between-mouse variation, and the within-mouse ICC is estimated at $4344/(4344 + 4417) = 0.50$.

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. Correcting this problem would probably only lead to increased P-values for effects involving time, and might therefore not change the basic conclusion of no significant group effects. A more serious critique is that the assumption of constant variances is violated, by an increasing variance over time (as is obvious from the graphs and also seen in the

separate time descriptive statistics) and by different between-subject variances in the three groups. Possible solutions to the heteroscedasticity over time are a suitable transformation of the outcome (e.g., square-root or log transformation) and/or a restriction of the time period to days with similar variance. The heteroscedasticity across groups is more difficult to deal with, and can basically not be solved within the class of models covered in the course.

In summary, the hierarchical model is problematic for these data, and one should not put much emphasis on its results. The textbook Davis (2002), *Statistical Methods for the Analysis of Repeated Measurements*, suggests these data to be analysed by advanced non-parametric methods. Another suggestion is advanced mixed models involving correlation structure, random slopes and variance depending on time and/or groups, as described in Pinheiro & Bates (2000), *Mixed-Effects Models in S and S-PLUS*.

Question 2.

A)

The first version of the study has subjects (persons) as experimental units, and two classification variables of interest: sex and diet. Each combination of sex and diet will comprise 10 subjects, about whom no further information exists. The experimental design can therefore be described as a two-factor (2×3) design with 10 replications. We may consider sex as a blocking factor, in which case we would call it a completely randomized block design. Denote by y_i the decline for subject i in cholesterol level between pre- and post-diet measurements, $i = 1, \ldots, 60$. The natural statistical model is a two-way ANOVA model with interaction,

$$y_i = \mu + \alpha_{\text{Sex}(i)} + \beta_{\text{Diet}(i)} + (\alpha\beta)_{\text{Sex}\times\text{Diet}(i)} + \varepsilon_i, \tag{1}$$

with the errors (ε_i) assumed i.i.d. and $\sim \text{N}(0, \sigma^2)$. The statistical analysis is performed in Minitab using either the Balanced ANOVA or General Linear Model menus; in Stata, we would use the `regress` or `anova` commands. In any case, the model specification should include both the main effects and the interaction. The ANOVA table will take the following form,

Source	DF
Sex	1
Diet	2
Sex*Diet	2
Error	54
Total	59

Finally, randomization is performed by randomly selecting among the volunteers of each sex (i.e., within each block) the subjects to receive each of the 3 diets.

B)

When age is taken into account, the subjects can no longer be considered as replicates within each treatment group. One way to adjust for age is to record the subject's age and to include age in the model as a covariate. In the simplest version, with the same relation with age for all subjects (parallel regression lines), the statistical model extends to

$$y_i = \mu + \alpha_{\text{Sex}(i)} + \beta_{\text{Diet}(i)} + (\alpha\beta)_{\text{Sex}\times\text{Diet}(i)} + \gamma \cdot \text{age}_i + \varepsilon_i. \tag{2}$$

Possible further extensions include non-parallel regression lines for age across either sexes or treatments, or both. The ANOVA table for model (2) contains an additional line for Age with 1 degrees of freedom (and one degrees of freedom less for Error). The models with non-parallel regression lines contain additional interactions between Age and either Sex or Diet, or both.

Another option for taking age into account in the study, is to divide the subjects into groups of age ranges, e.g. < 35, $35 - 49$, $50 - 64$, and ≥ 65 years. Ideally, one would want approximately the same number of subjects of each sex in all age groups, but it may be difficult to achieve in practice. In the statistical model, Age-group would be a factor similar to Sex and Diet, and it may be included as an additive effect only or it may enter into interactions with the other factors. In the ANOVA table, the main effect would have the degrees of freedom equal to the number of group levels minus one (i.e., DF = 3 in the example), which would then be subtracted from DFE.

C)

To let subjects go through all 3 diets, changes the experimental design to a *cross-over design*. The experimental unit is no longer subject, but subject-period (one of the 3 dietary periods). With 60 subjects there would now be 180 such periods. Instead, the number of subjects could be reduced to 20. The general advantage of this type of design is that diets can be compared within each subject, thereby eliminating the (presumably) large subject variation for the comparison of diets. The main disadvantage is a somewhat more complex design, in particular the need for controlling after-effects of the diets.

Denote now by y_i the decline for subject-period i in cholesterol level between pre- and post-diet measurements, $i = 1, \ldots, 60$ (corresponding to 20 subjects). The repeated measures for each subject can be taken into account by a hierarchical or split-plot type model with subjects as the upper hierarchical level ("whole plots") and subject-periods as the lower hierarchical level ("split plots"). With only 3 measures per subject, this would usually be considered a satisfactory method of analysis. In particular, it would be part of the study layout to ensure sufficient wash-out periods between diets, so that correlations between successive periods should be small. The statistical model can be written (ignoring for this part any impact of age discussed in B),

$$y_i = \mu + \alpha_{\text{Sex}(i)} + \beta_{\text{Diet}(i)} + (\alpha\beta)_{\text{Sex}\times\text{Diet}(i)} + \delta_{\text{Period}(i)} + A_{\text{Subject}(i)} + \varepsilon_i. \tag{3}$$

In Minitab, analysis of this model would preferably use the Mixed Effects Models menu, with Subject as a random effect (possibly nested within Sex, if subjects are not uniquely numbered). The ANOVA table takes the form,

Source	DF
Sex	1
Subject	18
Diet	2
Sex*Diet	2
Period	2
Error	36
Total	59

and the mean square for Subject becomes the denominator of the F-test for Sex. In Stata, the `mixed` command would be used, with (unique) subjects as random effects. Additional refinements of the design and layout might try to control for possible effects of the order in which the diets are given to subjects; this would lead into a detailed discussion of carry-over effects in cross-over designs.

D)

For the design in A), the standard error of the difference of two diet means (means of declines) would be $\sigma\sqrt{2/20}$, and a 95% confidence interval would have a margin of error of $t(.975, 54)\sigma\sqrt{2/20}$. Inserting the expected value of $\sigma = 20$ (and $t^* \approx 2.009$) yields a value of 12.7. Therefore, a difference between diet means of 10 units would not be expected to be significant at the 5% level.

For the design in B), information is lacking about the reduction in the subject variation achieved by including age in the model, either as a covariate or as a factor. Therefore, to answer the question, one would have to guess a (reduced) value of σ. From then, the calculation would be entirely similar to the above calculation for the design in A).

For the design in C), the diets would be compared using the within-subject (or subject-period) variation, as opposed to the previously used between-subject variation. The total variance of one observation equals $\sigma_A^2 + \sigma_\varepsilon^2$, where σ_A^2 is the between subject-subject variation and σ_ε^2 is the within-subject variation. The given σ-value includes both of these, but with a known value of the within-subject correlation $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_\varepsilon^2)$, we can reconstruct the required value of σ_ε^2. Specifically, from $\rho = 0.3$ and $\sigma^2 = 20^2$, we get $\sigma_A^2 = 20^2 \times 0.3 = 120$ and $\sigma_\varepsilon^2 = 20^2 - 120 = 280 = 16.7^2$. Repeating the above calculation with this value yields a margin of error of 10.7. Therefore, a difference between diet means of 10 units would be expected to be close to significant in this design.

Question 3.

Denote by p_i the probability of subject i developing CHD during the follow-up period, $i = 1, \ldots, 3154$. 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, corresponding to the equation,

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{weight}_i + \beta_3 \texttt{smoke}_i + \alpha_{\texttt{behpat}(i)} + \beta_4 \texttt{arcus}_i,$$

where $\alpha_1 = 0$ and $\alpha_2, \alpha_3, \alpha_4$ are the coefficients for a comparison of the actual category and the reference category (1).

Interpretations of effects:

- `age`: highly significant effect ($P < 0.0005$); odds-ratio for a change/difference in age of 10 years $= \mathrm{e}^{10 \cdot 0.0690} = 1.99$; high age is a risk factor for CHD;
- `weight`: highly significant effect ($P < 0.0005$); odds-ratio for a change/difference in age of 50 pounds $= \mathrm{e}^{50 \cdot 0.01285} = 1.90$; high weight is a risk factor for CHD;
- `smoke`: highly significant effect ($P < 0.0005$); odds-ratio comparing smoker vs. non-smoker $= \mathrm{e}^{0.665} = 1.94$; smoking is a risk factor for CHD;
- `behpat`: overall significance not known, but probably significant; odds-ratios for comparisons with the reference category for categories 2, 3, 4, respectively, are: $\mathrm{e}^{0.0541} = 1.06$, $\mathrm{e}^{-0.728} = 0.48$, $\mathrm{e}^{-0.602} = 0.55$; behavioural groups 1 and 2 seem very similar ($P = 0.80$ for testing homogeneity), and groups 3 and 4 also seem similar (no test possible) and at lower risk of CHD than groups 1 and 2.
- `arcus`: weakly significant effect ($P = 0.041$); odds-ratio for presence of arcus sinus condition $= \mathrm{e}^{0.285} = 1.33$; the condition is a risk factor for CHD;

For the assessment of effects of `behpat` it would be useful to *i)* carry out a combined Wald test (`testparm`), and *ii)* to carry out the three remaining pairwise comparisons (`pwcompare`).

B)

In addition to the model equation above, the model assumes the outcomes y_i to be independent. The model equation involves an assumption of additivity of the different effects (i.e., no interaction) and an assumption of linearity of the continuous effects (i.e., age and weight) on logit scale. The model may furthermore fit poorly to the data if important predictors are omitted. It is therefore suggested to examine the additional predictors and include any of those with significant effect, and to examine interactions between all significant predictors. The linearity of the relation with continuous predictors can be assessed by adding non-linear terms (e.g., quadratic). Another method is to categorise the predictor into a suitable small number of categories and assess whether the estimates on logit scale follow a roughly linear pattern.

The goodness-of-fit test based on the Pearson chi-square statistic is useless because of the lack of replication; the listing gives a total of 2466 covariate patterns. The Hosmer-Lemeshow test does not suffer from this drawback and its P-value of 0.64 is perfectly valid. As the test is somewhat dependent on the number of groups, it is suggested to rerun the test with different numbers of groups, in particular the standard recommended 10 groups. The model fit statistics do not give any information about goodness-of-fit.

C)

For this question, the solution is confined to summarising the information that can be extracted from the listings provided and answering the specific questions.

The first model fit for Question 3.C is close to a full model except that `behpat` is not included. The model is fit on a different number of observations than the model Question 3.A so no direct comparison (by a likelihood-ratio test, AIC or BIC) is possible. The table of estimated correlations between parameter estimates shows extreme correlations (around ± 0.99) between the estimates for `weight`, `height` and `bmi`, and the pairwise correlations between the actual values of these variables also show some correlation (highest 0.81). Given the strong biological and numerical link between these variables it is suggested to retain only one of them. Each of them could be tried separately (together with the other predictors) to see which of them produces the best fit, as measured by the $\log L$ or AIC value. There is also a fairly strong correlation between the estimates for `sbp` and `dbp`, and between the raw values as well, and again it is suggested to retain only one of the two variables. The model fitted shows a strongly significant effect for `sbp` and a clearly non-significant effect of `dbp`, so it is suggested to drop the latter.

The last model fitted with only five predictors uses the same number of observations as the "full" model. It has an inferior (higher) AIC, and a likelihood-ratio test for the model reduction is computed as

$$G^2 = 2 \cdot (-799.124 + 804.300) = 10.35, \quad P = \mathrm{P}(\chi^2(4) > 10.35) < 0.05 \text{ (because } \chi^2_{.95}(4) = 9.49).$$

The calculation shows that there is statistical evidence against the simpler model. It seems likely that one of the variables `height`, `weight` and `bmi` needs to be included in a reduced model. On the other hand, the model for Question 3.A did not include an effect of blood pressures and `chol`. The effect of `arcus` switches between significant and non-significant between the models, so it should probably be included for now and further explored. The best model suggested from these listings therefore includes the variables: `age`, `weight`, `smoke`, `sbp`, `chol`, `arcus`, `behpat`, where the effect of the latter is as a categorical variable. It may be possible to reduce behavioural patterns from 4 categories to the 2 behavioural types mentioned in the introduction.