Biostats VHM 802 Course
Winter 2016, 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 can be described from different perspectives. The multiple measurements over time on the same dog point to a repeated measures design. The fact that the dogs received different treatments at different time points leads to a cross-over design. We could also simply consider the dogs as blocks, hence a block design. There are four treatments in total, namely the four combinations of the two factors preparation and dose (SL, SH, TL, TH). Because each dog only receives three of these treatments, we have an incomplete block design, and it turns out to be a balanced incomplete block design (BIBD) with the parameters (in the GO textbook notation),

$$g = 4,\ b = 20,\ r = 15,\ k = 3,\ \lambda = 10.$$

It is seen that the required relations ($rg{=}bk$ and $\lambda(g-1){=}r(k-1)$) are both met. This view of the design does not take into account the periods (represented by `day`); as the design is balanced in the periods we should consider whether it contains any structure related to Youden squares. Indeed, it is seen that each of the five sets of four consecutive rows (i.e., rows 1−4, 5−8, ..., 17−20) form Youden squares. The first two squares are identical, but the remaining four squares are all different. In further description of the design, we could say that it has a hierarchical structure, with measurements within dogs, and that measurements are the experimental units for the treatments.

A natural (initial) model for the design would include effects of the treatments, here the main effects of preparation and dose and their interaction, the blocks (dogs) and the additional blocking factor (day). If we denote the calcium value for measurement i by y_i, for $i = 1, \ldots, 60$, we can write this model in a single-index notation as follows,

$$y_i = \mu + \alpha_{\texttt{prep}(i)} + \beta_{\texttt{dose}(i)} + (\alpha\beta)_{\texttt{prep*dose}(i)} + \gamma_{\texttt{day}(i)} + \delta_{\texttt{dog}(i)} + \varepsilon_i, \tag{1}$$

where the error terms, ε_i's, are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$. The dog effects ($\delta_j$) can be taken as either fixed or random; in the former case, the analysis relies only on intra- (or within-)block information. If the dogs were not selected in a way so that they could reasonably represent a suitable population, it is often most natural to use fixed dog effects, in order to avoid any confounding effects from dog characteristics on the treatment estimates. A hierarchical model for repeated measures would have random dog effects, but because there are no predictors at the dog level it would also be legitimate to have fixed dog effects. From the perspective of a carry-over design, model (1) is the usual starting point for analysis, including the treatments, periods and subjects.

B)

The model analysed is an extension of model (1), by including also interactions between treatment terms (main effects and interaction) with the periods. The ANOVA table can be used to assess the significance of the different terms in a balanced design, and more generally when partial (or adjusted) and sequential sum of squares are equal. It is seen that all effects involving `day` meet this requirement,

a results of the balancedness in days. Some minor differences exist between the sum of squares for the other terms, and one should interpret the tests based on partial sum of squares with caution. Among the `day` effects, only the interaction `prep*day` is significant ($P = 0.003$). The interaction `prep*dose` is clearly non-significant, but there is a strong significance ($P < 0.0005$) for `dose` (even when considering the uncertainty involving the sum of squares). Also the dog effects are strongly significant, but this is probably expected and of minor interest. In summary, all three factors show significant impact on the outcome, and our focus for interpretation should be on the main effect for `dose` and the interaction `prep*day`.

- The higher dose is associated with a higher serum calcium level, as is seen from the two fitted means in the Minitab listing (15.16 vs. 14.29), and 95% confidence intervals for these means would have a margin of error of $t^* \cdot \text{SE} = 2.045 \cdot 0.125 = 0.26$, where $t^* = t_{.975}(29) = 2.045$. This is indeed the biologically expected impact of higher doses of a parathyroid extract.

- The `day*prep` interaction is shown in the interaction plot, perhaps most clearly in the top right corner of the display. The curves for the two preparations appear as non-parallel, because the test preparation takes the highest value on days 1 and 3, and the lowest on day 2. With days having no obvious interpretation except as a blocking factor, it seems natural to focus mainly on the comparison between preparations on each day. The resulting three pairwise comparisons can be done either without or with adjustment for multiple (3) comparisons; for simplicity we consider only the former. In order to continue the analysis, we would need the SE for such differences between the two preparations on each of the days. Due to incomplete blocks and the restriction to specific days in the comparison, no standard formula will work to compute these SEs. Our best guess based on the information at hand would be to treat the means for S and T as independent, leading e.g. for day 1 to the calculation: $\text{SE}^2 \approx \sqrt{0.234^2 + 0.234^2} = 0.33$ (and for days 2 and 3 to a value of 0.36). (Calculation in Minitab/Stata, by requesting pairwise comparisons, will give the exact values as 0.37 and 0.42, so the simple approximations underestimate the SE somewhat.) Working with the approximate values would lead us to consider differences beyond a magnitude of $t^* \cdot \text{SE} \approx 0.7$ as significant, and we see that the differences at days 2 and 3 exceed this value by a good margin whereas the differences at day 1 is clearly lower. We would therefore conclude that the test preparation was higher on day 3, lower on day 2 and not too different on day 1.

In summary, the dose effect was clear and as expected, whereas the effect of preparations appears limited to a variable/inconsistent pattern across the days (periods) where the study was conducted. If such an effect was of interest, it could be suggested to confirm it by a follow-up experiment.

C)

Even if the optimal power (λ) in the Box-Cox analysis was far away from 1, the 95% confidence interval is very wide and still includes 1, corresponding to no transformation. Therefore the benefits of transforming the outcome may be minor. One way to assess that is to compare the residual plots for the untransformed and inverse transformed outcomes. Both normal plots are reasonably straight with the same cluster of points deviating from the line; note that after inverse transformation the order of the observations are reversed, so that the residuals will often switch from the right to the left tail, and vice versa. None of the plots seem to deviate very strongly from a normal distribution, and no clear outliers are seen. Most importantly for our discussion, no obvious improvement is seen after transformation. Also the two plots of standardised residuals against fitted values seem similar, and none of them show any obvious reason for concern. Taking these findings together, there seems to be no substantial improvement in the residuals after transformation, and indeed the untransformed

residuals look pretty good. Therefore there does not seem to any real advantage in transforming the data, and we would then prefer the simpler interpretation when analysing on original scale.

D)

In the context of an incomplete block design, it would usually be of interest to explore the impact of switching to random block effects, hereby including any inter-block information about the treatments in the analysis. As discussed previously, such an analysis is most interesting when the blocks can be thought to represent a suitable population.

The repeated measures on dogs have not been accounted for in the analysis so far with any specific correlation structure; we would usually think of the fixed dog effects as corresponding to the same (compound symmetry) correlation structure as random dog effects. Therefore, it could be explored whether other correlation structures allowing for autocorrelation over periods within each dog will improve the model fit.

From the perspective of a carry-over design it would be of interest to explore whether any carry-over effects of treatments from one period to the next can be seen. One approach is to explore whether any effects of squares exist because the 4 different squares do not have the same carry-over combinations. Another approach is to construct indicator variables representing specific carry-over effects from the treatments and include those in the model, as was done in Example 13.12 of the GO textbook. The fact that one column is "missing" in the Latin square (and we hence have less replication to assess the carry-over effects) gives less power to such model terms, but they could still be explored.

Question 2.

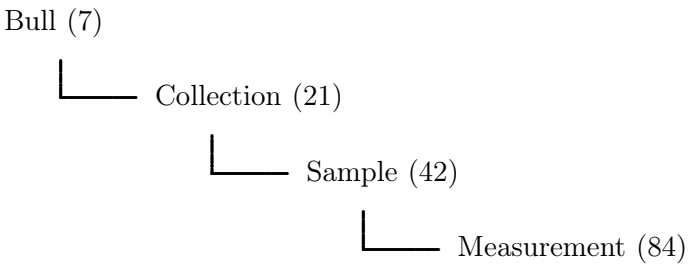
The dataset includes the following four variables and a total of 84 concentrations measured for each of the two methods:

`bull`	levels (or categories) $1-7 \sim$ index i
`day`	levels $1-3$ within each bull $\sim$ index j
`sample`	levels $1-2$ within each bull and day $\sim$ index k
`repl`	levels $1-2$ within each bull, sample and day $\sim$ index l.

Denote by y_{ijkl} and z_{ijkl} the live semen concentrations measured by methods 1 and 2, respectively.

A)

The data structure is hierarchical, as shown in the diagram:



The hierarchical structure requires us to include random effects of all hierarchical levels above the lowest one, possibly except the top level, which may also be modelled by fixed effects. It is not mentioned in the text whether the 7 bulls are thought to represent a population, or whether they

were selected by convenience. There are no factors (treatments) to examine at the bull level, so the bulls can be modelled by fixed effects (as shown in the Minitab and Stata listings) or by random effects. The collections are identified in the data by the factor `day(bull)`, i.e. days nested within bulls, or equivalently by all combinations of days and bulls. Similarly, the samples are identified as `sample(day bull)`, i.e. samples nested within days and bulls. The statistical model for y (or z) is:

$$y_{ijkl} = \mu + \alpha_i + B_{ij} + C_{ijk} + \varepsilon_{ijkl}, \tag{2}$$

where $B_{ij} \sim \mathrm{N}(0, \sigma_B^2)$, $C_{ijk} \sim \mathrm{N}(0, \sigma_C^2)$ and $\varepsilon_{ijkl} \sim \mathrm{N}(0, \sigma^2)$, and all random variables are independent. In this model, B_{ij} and C_{ijk} represent the random effects of collections and samples, respectively. The bull effects are in the fixed α_i-parameters. Furthermore, σ_B^2 and σ_C^2 are the variances of the collection and sample random effects, respectively, and σ^2 is the error variance.

B)

The requested statistics for the two methods, with formulae for computation, are given in the table below. The three variance components may be interpreted as follows. The collection variance, σ_B^2, represents the variation between different collections from the same bull. The sample variance, σ_C^2, represents the variation between different samples and operators from the same collection from a bull. Finally, the error variance, σ^2, represents the variation between duplicate measurements (by the same operator) on the same sample.

Statistic	Formula	Method 1	Method 2
collection variance	$\hat{\sigma}_B^2$	0.04259	0.02365
sample variance	$\hat{\sigma}_C^2$	0.00139	0.01793
measurement variance	$\hat{\sigma}^2$	0.00074	0.01324
total variance $\hat{\sigma}_t^2$	$\hat{\sigma}_B^2 + \hat{\sigma}_C^2 + \hat{\sigma}^2$	0.04472	0.05482
prop. collection var.	$\hat{\sigma}_B^2/\hat{\sigma}_t^2$	0.952	0.431
prop. sample var.	$\hat{\sigma}_C^2/\hat{\sigma}_t^2$	0.031	0.322
prop. measurement var.	$\hat{\sigma}^2/\hat{\sigma}_t^2$	0.017	0.242
repeatability r	$2.83\,\hat{\sigma}$	0.077	0.326
reproducibility R	$2.83\sqrt{\hat{\sigma}^2 + \hat{\sigma}_C^2}$	0.131	0.500
expected collection diff.	$2.83\sqrt{\hat{\sigma}^2 + \hat{\sigma}_C^2 + \hat{\sigma}_B^2}$	0.598	0.663

The last value in the table gives the value, which with probability 95% would not be exceeded by a difference between two measurements on different collections from the same bull when samples are processed by different operators. To compute the corresponding value when the samples are processed by the same operator, simply omit the contribution from $\hat{\sigma}_C^2$ in the formula.

Clearly, there is a huge difference between methods 1 and 2. Method 1 has very small variance components for sample and measurement, both in absolute values and as a proportion of the total variation. This leads to much smaller repeatability and reproducibility for method 1. Due to the more accurate measurements, also the expected difference between two collections is smaller, although the major component of this variation is more biological than measurement variation. From the point of view of the precision of the measurements, method 1 is therefore clearly preferable to method 2.

C)

When all samples are done by the same operator and all replications listed as `repl=1` are taken first, the factors `sample` and `repl` can be given an independent meaning (and are no longer purely nested factors in the design). Therefore, we can introduce fixed effects of sample and replication in model (3) below. Including also the interaction between these two factors, corresponding to a different replication effect for the two operators (in other words, the operators may not work at the same speed), we arrive at the following model:

$$y_{ijkl} = \mu + \alpha_i + \beta_k + \gamma_l + (\beta\gamma)_{kl} + B_{ij} + C_{ijk} + \varepsilon_{ijkl}, \tag{3}$$

where β_k is the effect of operator k, γ_l that of replication l and $(\beta\gamma)_{kl}$ their interaction effect. This model can be run in Stata by adding the (fixed) interaction term `sample##repl` to the `mixed` model statement. However, Minitab will not run this model due to restrictions put on the hierarchical structure. If the actual time between replicates was known, a new variable (say `timelag`) could be constructed to have this value when `repl=2` and the value 0 when `repl=1`. The new variable could then be substituted for `repl` in the fixed part of the model to yield the interaction term `sample##c.timelag`. If larger time lags were associated with a consistent decrease in live cell counts, this model might give a better fit than model (3) with the first interaction considered.

D)

It would be possible to combine the two datasets, and the hierarchy would be essentially unchanged with now four measurements (two by each method) per sample. The temporal sequence of these four measurements could be taken into account with repeated measures correlation structures. An alternative view would introduce an additional sample×method level, in order to allow a stronger correlation between the two duplicate measurements for the same method. In both cases, models accounting for the data structure could be run in statistical software. Such models would however most likely strongly violate any assumptions of homoscedasticity, because our analyses demonstrated substantial differences in variability between measurements by the two methods. Therefore it would be quite legitimate to abstain from any attempts to analyse all the data in a single model, solely on the basis of the different distributions of variability for the two methods.

Question 3.

A)

The study is observational, more precisely a retrospective case-control study. The cases and controls are not matched. Age is an obvious potential confounder for associations between exposures and cancer risk. If the main exposure of interest is alcohol consumption, the other exposure variable, tobacco consumption may be considered as a potential confounder for the association between alcohol and cancer. The temporal relation between tobacco and alcohol consumption is not clear, but there is no obvious reason to believe that smoking would be temporally after alcohol (and thus an intervening, or intermediate, variable).

B)

This is a logistic regression model with linear effects assumed for all predictors. Therefore, with Y_i indicating the disease status of subject i we would have

$$p_i \quad = \quad \mathrm{P}(Y_i = 1), \qquad \text{and}$$

$$\text{logit}(p_i) \quad = \quad \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{tob}_i + \beta_3 \texttt{alc}_i,$$

where β_0 is the intercept, and the other β's are logit-scale regression coefficients corresponding to the respective predictors. The interpretation of the parameters are as follows,

- β_0: in a case-control study, the intercept cannot be interpreted in terms of the prevalence in the population; it's the estimated logit-scale prevalence within the study sample for a hypothetical subject with all predictors equal to zero, but that is usually of no practical interest (regardless of whether the predictors are centred or not);
- β_1: the coefficient for `age` was estimated at $\hat{\beta}_1 = 0.072$, corresponding to an odds-ratio (OR) for a one-year increase in age of: $\text{OR}_1 = \exp(.072) = 1.075$, or a 7.5% increase in cancer risk;
- β_2: the coefficient for `tob` was estimated at $\hat{\beta}_2 = 0.039$, giving for an increase in tobacco intake of 1 g/day: $\text{OR}_2 = \exp(.039) = 1.040$, or a 4% increase in risk;
- β_3: the coefficient for `alc` was estimated at $\hat{\beta}_3 = 0.0265$, giving for an increase in alcohol intake of 1 g/day: $\text{OR}_3 = \exp(.0265) = 1.027$, or a 2.7% increase in risk.

All 3 predictors are highly significant ($P < 0.0005$). The model assumptions in the logistic model for a case-control study are: *i)* independent observations of the predictors for all selected study subjects, and *ii)* the linear and additive relations between the predictors and logit(p) as expressed in the equation above. The Stata listing includes two goodness-of-fit tests. The Pearson chi-square test is totally useless here because the very large number of covariate patterns leaves too little replication within each pattern. The Hosmer-Lemeshow test is valid and shows a weak significance ($P = 0.045$), indicating some problems with the model fit. Further exploration would be needed to determine the reasons for the lack of fit.

C)

In the second analysis, the predictors `age`, `tob` and `alc` have all been replaced by categorical predictors with 6, 4 and 4 categories, respectively. The parameter estimates shown in the listing represent the corresponding 5, 3 and 3 comparisons with the respective baseline categories. The two models fitted are not submodels of each other, so a formal test to compare them cannot be computed. Their fit can be compared by the values of $\log L$, and the model with categorical predictors has a better fit. Considering that it also includes 8 additional parameters, the difference in $\log L$ is not large enough to give it the lowest AIC (specifically, $2(359.14 - 352.14) = 14 < 2 \cdot 8 = 16$). The second model seems however to have a better fit because it's Hosmer-Lemeshow goodness-of-fit test is no longer critical ($P = 0.84$). This suggests that in some aspects the categorical modelling is a substantial improvement of the model fit; we will consider this for each predictor separately.

- `age`: the age groups are equidistant (apart from the last open-ended interval), so in a linear relation one would expect roughly equal changes in logit(p) from one group to the next; this pattern is not found in the estimates where the differences are: 1.97, 1.80, 0.56, 0.56 and −0.07. This pattern reflects a strong increase in risk up to age 45–54 and then an only slightly increasing risk in later ages. Based on these estimates one would guess that a linear effect of `age` has poor fit, and that either the categorical version or another non-linear (possibly quadratic) function of age would be the best choice;
- `tob`: also here the groups are equidistant, so we calculate the differences in the same way as above: 0.44, 0.07, and 1.13. This predictor also shows some non-linearity, but the pattern seems less clear. An exploration of the quantitative relation of risk with `tob` would be helpful, but

perhaps a linear relation is not such a bad approximation. Because the predictor had grouped values from the beginning, so it could also be suggested to try to estimate coefficients for all the original (nine) groups;

- `alc`: again we calculate the differences: 1.44, 0.52, and 1.63. These values do not indicate a strongly non-linear relation with `alc`. It is suggested to try to add a quadratic term to the linear relation from the first model, but if that is non-significant to retain the linear relation.

D)

Both models assume additive effects of the three predictors. It would be obvious, indeed recommended, to try to add interaction terms, preferably to begin with an interaction for each pair of predictors. The interactions will take different forms depending on whether the predictors involved are modelled as quantitative or categorical.

Our initial study description identified age as a potential confounder, and also tobacco consumption could be interpreted as a potential confounder. In order to explore these hypotheses further, it would be of interest to carry out further analysis. For example, for `age` one would need to run the final model with and without `age` and note the change in coefficient(s) for `alc`. It would also be relevant to determine whether there was any association between `age` and `alc` in the controls; this could be assessed with a correlation coefficient or a linear model for alcohol consumption as a function of age.

To aid in validation and interpretation of a final model it could also be suggested to compute and inspect the residuals (where one would need to ensure a sufficiently low number of covariate patterns), and to quantify the predictive ability of the model, e.g. as sensitivity/specificity or area under the ROC-curve.