Biostats VHM 802 Course
Winter 2024, 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.

For this question we denote by y_i the i^{th} ratio of the SO_2 measurement by the Visiplume and standard methods, $i = 1, \ldots, 31$, and we let $\texttt{group}(i)$ denote the experimental condition (1−4) under which the measurements were taken.

A)

The statistical model is a one-way ANOVA model,

$$y_i = \mu_{\texttt{group}(i)} + \varepsilon_i,$$

where the ε_i's are i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$. The model assumptions are: normality, homoscedasticity and independence of the errors, in addition to equal means of the observations within each group. The listing and residual plots shows the distribution of the standardised residuals to be far from $\mathrm{N}(0, 1)$, and that the standard deviation is much larger in group 2 than in the other groups.

Both of these problems are most likely attributable to a very large residual for observation no. 20 — the last observation in the Smoke group. The observation itself is clearly much larger than any other value in the dataset. The visual discrepancy between this observation and the rest may be so large that we are willing to discard the observation as an error right away. Obviously, one would need to check the records for any evidence of an error; as these are published data, such evidence is unlikely to exist. A statistical assessment of whether this observation is outlying can be obtained from an outlier test based on the deletion residual. The computer listings do not give the value of the deletion residual; however, the plots show that the standardised residual is clearly above 4 and probably close to 5, and the deletion residual would be even more extreme than that. For the outlier test, one would need to use a $t(26)$ reference distribution. A t-distribution table (e.g., Table C of the Baldi & Moore textbook of VHM 801) shows that $t(.9995, 26) = 3.707$, and the corresponding P-value for the outlier test would be $P = 2 \cdot 31 \cdot .0005 = 0.031$. Given that the actual deletion residual is (much) larger than 3.707, we can safely assume there is strong evidence ($P \ll 0.03$) against observation 20 being in agreement with the model for the other observations.

B)

The Stata listing shows two Box-Cox analyses to determine an optimal transformation of the data:

- for the full dataset, an optimal power transformation is obtained with $\hat{\lambda} = -1.57 \approx -1.5$, and there is strong evidence in favour of the transformation,
- for the reduced dataset without observation 20, an optimal power transformation is obtained with $\hat{\lambda} = 1.19$; however, there is no evidence in favour of the transformation ($P = 0.83$ for $\mathrm{H}_0 : \lambda = 1$).

Analyses of the one-way ANOVA model for the power transformed data and for the untransformed data with observation 20 removed are given in the listings. Both of these analyses have better looking residual plots, and only minor differences between the within-group standard deviations. Still, the normal plot for the power transformed data does not look very good, and one may speculate that it could imply formal evidence against the normal distribution (in fact, different normality tests give P-values below and above 0.05). The model for untransformed data has one somewhat large negative residual, for observation 19, but even if this value seems somewhat low for its group it does not look nearly as extreme as observation 20, and it is probably not of interest to remove it. The choice between the two models is not obvious (if the residual plot for the power transformed is still considered acceptable). In favour of the model for transformed data speaks that no data point is lost, or omitted, but on the other hand the analysis needs to be carried out at an unnatural scale. In favour of the model for the reduced data speaks that if the observation was really an error, one might introduce a bias in the analysis by undertaking a "strange" transformation, and the appropriate approach is to simply ignore it. In terms of the statistical significance, the two models both show strong significance and apparently similar results. This could be taken as an argument to keep the outlying observation in the data. In any case, it should be reported that the two analyses give similar results.

C)

Results are shown here for both the acceptable models from B). The listed questions of interest can be answered by referring to the overall test for significance between groups, a pairwise comparison between the two coal-fired plants, and a contrast comparing the smoke group with an average of the two coal-fired plants, i.e., $\psi = \mu_2 - \frac{1}{2}(\mu_3 + \mu_4)$.

Question	Power transformation model	Outlier elimination model
groups overall	$F(3, 27) = 6.87,\ P = 0.0014$	$F(3, 26) = 7.94,\ P = 0.0006$
coal-fired plants	$t(27) = \frac{1.076-1.256}{\sqrt{.0536(\frac{1}{5}+\frac{1}{6})}} = -1.28,\ P > 0.20$	$t(26) = \frac{0.967-0.876}{\sqrt{.0169(\frac{1}{5}+\frac{1}{6})}} = 1.16,\ P > 0.20$
smoke vs. plants	$t(27) = -0.422/.0989 = 4.27,\ P < 0.001$	$t(26) = 0.265/.0569 = 4.66,\ P < 0.001$

We conclude there is evidence of an overall difference between groups, but not between the two coal-fired plants. The calculations for the contrast are shown below (for the power-transformed data),

$$\hat{\psi} = 0.744 - (1.076 + 1.256)/2 = -0.422,\ \mathrm{SE}(\hat{\psi}) = \sqrt{.0536(1^2/11 + (\tfrac{1}{2})^2/5 + (\tfrac{1}{2})^2/6)} = 0.0989.$$

There is strong evidence that the mean (median) ratio between the methods differ between the Smoke and Plants groups; the ratio is higher for air produced by the smoke generator. Finally, the comparison with the standard reading is done by looking for evidence against the hypotheses that the mean in each of the four groups equals 1, at both analysis scales. We can construct t-tests as above or confidence intervals (CIs) for the μ_j and inspect whether 1 is included in the CIs. With $t^* = 2.056$ and 2.052 for DFE = 26 and 27, respectively, each CI has a margin of error of $t^* \cdot \sqrt{\mathrm{MSE}/n_j}$, where n_j is the size of group j. The estimated means suggest that the CIs for groups 2 and 4 will not include 1, and the calculations (not included here) confirm this to be the case.

Question 2.

The data comprise 69 records corresponding to different sites. All measured variables are quantitative, and there is no indication of missing values.

A)

The shown analysis is a principal components analysis (PCA) based on the correlation matrix for the eight PCB variables, including the total PCB. The plots show the loadings for the first two principal components (the values are listed in the table of eigenvectors) and the scores for the 69 sites on the first two principal components. The first two components account for 85% of the variation in the data, and all the remaining eigenvalues are below 0.5. It is therefore reasonable to focus attention only on the first two components.

The first component loads positively and moderately strongly on all PCB variables, including total PCB, except that the coefficients for the `pcb28` and `pcb52` are lower, but still positive. Therefore the first component can be interpreted as an indicator of overall presence of PCB at a site. The second component is a contrast between on one hand `pcb28`, `pcb52` and `pcb118` and on the other hand the remaining PCB componets, except the total. Among the first three PCB variables, the loading of `pcb118` is much lower than for the first two which are the main contributors to this component. Therefore the second component is essentially an indicator for `pcb28` and `pcb52`, in particular when these are large without all the other components being large as well.

The score plot shows a curious pattern with a concentration of points between -2 and 2 for component one and a relative narrow range, maybe (-0.5,0.5), of component two. The remaining approximately 10 points are scattered around at large values of either compoment one or two, or both. At least some of these points could perhaps be seen as outliers, by their far distance from the main cluster of points. The fact that there are so many of them on the other hand makes it less attractive to view them as outliers. Their large values, in particular for component one, would have to be caused by very large values for some of the PCB variables. This fits with the descriptive statistics showing that all PCB variables are substantially right-skewed.

If reducing the dimension of the PCA variables was the main objective of the PCA, one would extract the scores for the first two components and use them to represent the PCB variables in further analysis. If gaining understanding about the relationships between the PCB variables was the main objective of the PCA, one would focus on the loadings and their interpretation. The correlation matrix between the variables should also be of interest. If exploration of patterns among the locations based on their PCB values was the main objective, the score plot is the main result. It would naturally be of interest to determine the actual sites involved and perhaps link the extreme values to site characteristics or spatial locations. We do not have information in our data to do that.

With the large differences in the range for the different PCB variables, there is probably little interest in working with the covariance matrix rather than the correlations (as was done), despite that the variables are on the same scale. The right-skewness of many variables could suggest extreme values, possibly outliers, in the right tail, and these could (as already mentioned) affect the results strongly; correlations are strongly affected by extreme values. It would be of interest to explore the distribution of individual PCB variables further, and transformation to a scale where the right tail is less influential might be useful. A log-transformation would probably help, with some modification for `pcb126`.

B)

A number of questions can be addressed with these data, and this solution will discuss some examples. The request was for "multivariable or multivariate analyses", and these are different things (in our terminology). For a multivariable analysis, we need to decide on a single outcome and a purpose of such an analysis. The text mentions that the PCB total is only available upon completion of a comprehensive and expensive panel of analyses, so it might be of interest to try to predict total PCB from the seven specific PCB variables available. We do not know whether these are the most important ones (they certainly do not sum up to anything close the total PCB) or whether they

can be obtained cheaply, but from the information we have the prediction objective seems relevant. This will simply be a multiple regression analysis with `pcb` as the outcome and `pcb28`,...,`pcb180` as predictors. The outcome is right-skewed on its own, so a Box-Cox analysis to assess the most appropriate analysis scale is suggested. Some of the predictors are also very skewed, perhaps due to single large values, and one would need to carefully look for influential observations and perhaps consider transforming some of the predictors as well. Other multivariable analyses for a predictive purpose could be for `teqpcb` or `teq` as the outcome and the PCB variables as predictors again. In this instance, all PCB variables are included among the predictors, and it could be considered to instead use the two major principal components identified in A).

The most obvious multivariate analyses that are entirely different from the PCA already shown would be to determine patterns among the fish samples from different lakes based on their similarity. The score plot from A) does this to some extent, but the TEQ variables could be added, or it could be done based on the TEQ variables only. Multidimensional scaling (MDS) with modern methods could extend the classical MDS corresponding to the PCA. A different type of representation of the sites could be achieved with a hierarchical cluster analysis, or it could be attempted to split the data into a fixed number of clusters with the k-means clustering method. Any information available about the locations could potentially be overlaid or explored from the resulting representations of clusters (or the spatial representation from MDS). For all these analyses, the choice of distance measures should be considered carefully. We already mentioned the skewness of many of the variables and the potential for single sites be strongly influential. If all three specific TEQ variables are included, the overal `teq` should not be included as well, due to it being the sum of the three components.

Other types of analysis become available if we have a known classification of sites. The text does not mention any such classification, but it might be of interest to classify sites based on where their toxicity (TEQ) mainly originate from. In the listing of the first five sites, it is seen that for two of them the PCB contributes the main toxicity but for the three other sites the toxicity mainly comes from dioxin. It might be of interest to explore whether such a classification can be related to patterns among the PCB variables (here we cannot include the TEQ variables). For any classification purpose, we would be able to choose between linear discrimination analysis, (multinomial) logistic and nearest neighbour classification. For the latter the distance measure is important, for the former two the proper choice of scale for the predictor variables is important, as already discussed.

Question 3.

Denote by p_i the probability of subject i experiencing a recurrence of the tumour during the (unspecified) follow-up period, $i = 1, \ldots, 286$. In terms of the binary event, y_i, we have $p_i = \mathrm{P}(y_i = 1)$.

A)

Both models fitted are logistic regression models for the binary outcome and with all 9 variables included as predictors. They differ in the way the predictors are modelled: the first model has all predictors as categorical variables, whereas the second model has them all as quantitative predictors with a linear (slope) term. We can represent the second model by the equation,

$$\begin{aligned}\mathrm{logit}(p_i) &= \beta_0 + \beta_1\,\texttt{malig}_i + \beta_2\,\texttt{age}_i + \beta_3\,\texttt{menop}_i + \beta_4\,\texttt{tsize}_i + \beta_5\,\texttt{invnodes}_i \\ &\quad +\beta_6\,\texttt{nodecaps}_i + \beta_7\,\texttt{breast}_i + \beta_8\,\texttt{tloc}_i + \beta_9\,\texttt{irrad}_i.\end{aligned}$$

For the first model, each of the predictors need to be represented by indicator (dummy) variables for effects relative to a baseline category. For example, for `malig` we would have $\alpha_{\texttt{malig}(i)}$ with

two resulting parameters: for the comparisons of malignancy category 2 versus 1 and for category 3 versus 1.

Interpretations of the effects for `malig` in the two models, say (1) and (2):

(1) overall significance not known, one would need a multiple Wald test (or likelihood ratio test) to assess that but it might be weakly significant; odds-ratios for comparisons with the reference category (1) for categories 2 and 3, respectively, are: $e^{-0.3765} = 0.69$ and $e^{0.9709} = 2.64$; it seems surprising that malignancy category 2 has an estimated lower risk than category 1, but it is not significantly different ($P = 0.407$), whereas category 3 has an estimated higher risk than category 1 which is also weakly significant ($P = 0.034$);

(2) clearly significant effect ($P = 0.001$); odds-ratio for going one category up = $e^{0.7655} = 2.15$; higher malignancy category is a risk for recurrence.

B)

Both models assume independent observations (we have no information to question that) and additive effects of the different predictors (it would require subject-matter knowledge to suggest necessary interactions terms). Additionally, the second model assumes linear relationships between predictors and the probability on logit scale (or log-odds). This assumption is probably invalid for the predictors with multiple categories, in particular when these are not ordered. Only the predictors `nodecaps`, `breast` and `irrad` are unproblematic, and among the other predictors only `malig` and `age` could have a sensible relation with the log-odds, but the assumption of linearity should be checked. All other predictors have non-ordered categories, in part due to missing/unknown observations, and the assumed relation is non-sensical. These predictors should be modelled as categorical.

The categorical model shows problems with non-estimated parameters and omitted observations. This is due to some categories having no events or no non-events — this happens for four predictors. The typical solution to such a problem is to combine the categories affected with adjacent categories to obtain non-zero totals of non-events and events in the combined category. Such an approach seems feasible for `age`, `tsize` and `invnodes`, but the missing/unknown category for `tloc` does not logically combine with another category. The same situation is present for `nodecaps` but because the missing/unknown category has both non-events and events it does not lead to estimation problems. Nevertheless, it is probably not attractive to retain the missing/unknown category. These subjects may be omitted in initial analysis and reintroduced into the dataset if the corresponding predictors do not show any important effects. Alternative methods exist in abundance, but handling of missing data is beyond the course curriculum.

The model with categorical predictors can be further criticized for estimating an excessive number of parameters because some of the categorical predictors (in particular `age`, `tsize` and `invnodes`) include categories with very low number of observations (non-events and events combined). Also here it is suggested to combined categories suitably. Another potential problem with the predictors is that the variables `age` and `menop` are functionally related and therefore potentially very strongly collinear. It is not obvious that these two predictors can exist meaningfully in the model together. It would require further cross-tabulations to assess their relationship.

C)

The Stata listing shows a backwards elimination from the full categorical model. The initial model has the same problems as the categorical model already discussed, and it would be necessary to rerun the model selection approach after the issues with the initial model have been rectified. For this

reason, the results shown cannot really be trusted too much. In particular, we already determined that the predictor `invnodes` needed to have its categories redefined, so the results for this predictor will change. It might be that the result for `malig` is close to final; we see that the main impact is for category 3 to have higher risk of recurrence than the first two categories. For this reason, modelling `malig` by a linear term does not look as a good idea.

The suggested approach, from a purely statistical perspective and not taking into account epidemiological model-building considerations, therefore consists in revising the categorical predictors as discussed under B) and then applying a backwards elimination approach similar to the one shown. It could be suggested to expand it to a stepwise approach, whereby inclusion steps are considered as well, but still starting from a full model. The issue of potential collinearity between `age` and `menop` should be dealt with separately, perhaps prior to the model selection.

One potential change in the outlined approach could be explored for the ordinal categorical predictors whose assessment by categorical effects does not take into account the ordered categories. We already discussed `malig` where the categorical modelling seemed appropriate, but it might give too noisy (and hence too weak) assessment of effects for predictors with many ordered categories, such as `age` and `tsize`. It is suggested to explore the relation with the outcome and decide whether categorical or perhaps linear modelling is preferable.