Biostats VHM 802 Course
Winter 2014, 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.

A)

The study is an experiment, and the design is completely randomized with pigs as experimental units. The pigs were randomized to a total of 12 treatments comprised by combinations of two experimental factors: protein source (fish, milk, soy) and protein concentration (9, 12, 15, 18%). There were 3 replications per treatment, except for one treatment group with only 2 pigs (perhaps because one pig dropped out of the study). Thus the design is slightly unbalanced but complete. The obvious statistical model is a two-way ANOVA model with main effects of the two factors as well as an interaction between them, and this is also the analysis presented in the listings. The ANOVA table shows strongly significant main effects ($P < 0.001$ for both factors) and also a strong interaction ($P = 0.003$). This means that both factors affect the leucine concentration, but that their effects depend on the level of the other factor.

B)

As already noted above, there are significant differences between protein sources, but due to significant interaction these differences will depend on the protein concentrations. The interaction plot illustrates these findings. Visually it appears that fish protein is associated with lower leucine levels irrespective of the protein concentration, and that the difference between fish and soy protein is similar across all concentrations. On the other hand, milk protein seems to yield higher leucine levels than the other protein sources as the concentrations increase. For exploration of statistical significance between factor combinations, we compute

$$\mathrm{LSD}_{.95} = t_{.975}(23) \times \sqrt{\mathrm{MSE} \times 2/3} = 2.069\sqrt{20.93 \times 2/3} = 7.73\,.$$

This value assumes 3 pigs per treatment for all treatments, and therefore does not take into account the slight imbalance in the design. It also does not account for multiple comparisons. In order to adjust for comparisons among sources within all concentrations, the Bonferroni method would require us to adjust the significance level for each comparison from 0.05 to 0.05/12, because there are $4 \times 3(3-1)/2 = 12$ such comparisons. (Alternatively, we could adjust for all pairwise comparisons involving both source and concentration comparisons, see C) for details.) For the solution, we will use unadjusted comparisons. It is seen that at protein concentrations of 15% and 18%, all 3 sources are significantly different. For the 9% and 12% protein concentrations, milk and soy sources are not significantly different, but both of these yield higher leucine levels than fish (for 9%, this can also be read off the Stata listing). In summary, fish protein gave lower leucine levels than the other sources irrespective of concentrations, and milk gave higher leucine levels than soy but only for concentrations 15% and 18%.

C)

As already noted above, there are significant differences between protein concentrations, and the source by concentration interaction also implies that comparisons between concentrations will differ by protein sources. The $\text{LSD}_{.95}$ value computed above applies to (unadjusted) within-source comparisons as well. The Bonferroni method could adjust for $3 \times 4(4-1)/2 = 18$ comparisons between concentrations, or for the combined number of comparisons for sources and concentrations $(12 + 18 = 30)$. If only comparisons within sources and concentrations but not across both factors is desired, it is unnecessary (and too conservative) to adjust for all possible comparisons between the 12 treatments (i.e., $12(12-1)/2 = 66$ comparisons). For fish protein, it is seen that no significant differences exist between concentrations (this can also be gleaned from the Stata listing). For soy protein, the only significant comparison is an increase from 9% to 18%, with intermediate values for 12% and 15% concentrations. For milk protein, leucine concentrations appear to increase with protein concentrations, and the only non-significant comparison is between 12% and 15%.

A linear relationship between protein and leucine concentrations may exist for all sources, but the interaction only suggests a non-zero slope for milk protein. Because of the interaction between source and concentration, a linear modelling of concentration cannot easily be explored by manual calculations; one would have to run the model with non-parallel slopes on concentrations for the sources, and compare the two models by an F-test (it turns out to be completely non-significant). It is possible to explore a linear relation for milk protein by computing polynomial contrasts based on the 4 leucine means for milk protein. The table below gives contrast weights, estimates and standard errors for the linear, quadratic and cubic contrasts on protein concentrations.

Contrast	Weights	Estimate	SE	t
linear	$(-3, -1, 1, 3)$	99.76	11.81	8.45
quadratic	$(1, -1, -1, 1)$	8.50	5.28	1.61
cubic	$(-1, 3, -3, 1)$	-11.72	11.81	-0.99

For example, the linear contrast and its SE were computed as follows,

$$\widehat{w} = -3 \cdot 35.40 - 43.47 + 49.93 + 3 \cdot 66.50 = 99.76, \quad \text{SE}(\widehat{w}) = \sqrt{\text{MSE}(3^2 + 1^2 + 1^2 + 3^2)/3} = 11.81\,.$$

Among these contrasts, only the linear contrast is significant, providing support of the adequacy of linear modelling for concentration. The different magnitudes of t-values also show that the linear contrast explains the vast majority of variation among concentrations for milk protein. In summary, a linear modelling of concentrations seems well supported by the data.

D)

If 4 pigs were selected from each of 9 litters (for a total of 36 pigs), the statistical analysis would need to account for correlations between pigs from the same litter. This would require random effects for litters, and we could utilize this additional hierarchical level in the data to set up a split-plot experimental design, as follows:

- whole plots = litters, whole-plot factor = protein sources,
- split plots = pigs, split-plot factor = protein concentrations,
- 3 replications for each protein source, hence no blocks.

The hierarchical structure would consist in pigs nested within litters. If there was substantial between-litter variation in leucine levels, the pig-level variance would be lower than for the original design,

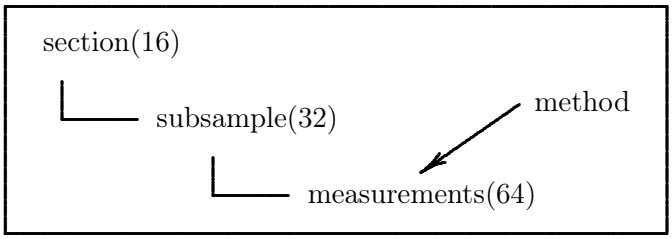
hence giving more precise comparisons for protein concentrations. The comparisons between protein sources, however, might have a similar precision as in the original design.

Note that the split-plot design arises not only from adding the litter level to the hierarchy; it also involves a decision to allocate concentrations within litters and sources between litters. If the two factors were allocated randomly across and within litters, both factors would have variation at both the pig and litter levels. In addition, comparisons between factors levels would no longer have the same precision because such comparisons depend on how many times a particular pair of factor levels occur within the same litter. Experimental designs with such "randomness" in the statistical properties of estimates are generally avoided, if possible.

Question 2.

A)

The study is an experiment, designed to explore variability in aflatoxin concentrations in peanuts; it is described in Tiemstra (1969), A study of the variability associated with sampling peanuts for aflatoxin, *J. Amer. Oil Chemists' Soc.* **46**, 667-672. The design includes multiple units: sections/samples (16), subsamples (32) and the measurements (or measurement units for the two analytical methods). The units are hierarchically nested. There is only one factor: the analytical method (`BF` or `CB`), and the design is balanced in both methods and the higher hierarchical units. The diagram below shows the hierarchical structure.



Although the units are hierarchically nested, the design lacks one important feature of a split-plot (in this case, split-split-plot) design, namely the presence of factors randomised to the units at the different levels. There is only one factor, and it is not clear from the description whether any explicit randomisation took place. Therefore, to call this a split(-split)-plot design is not quite correct.

B)

We use the following notation,

$$\begin{aligned} y_{ijk} &= \text{aflatoxin conc. obtained by method } k \text{ for subsample } j \text{ of section } i, \\ & \quad i = 1, \ldots, 16;\ j = 1, 2 \sim (\text{A,B});\ k = 1, 2 \sim (\texttt{BF,CB})\,. \end{aligned}$$

As our modelling aims at quantifying the variation between sections, subsamples and measurements, we will use random effects for sections and subsamples, in addition to a simple fixed effect for methods. The model equation becomes

$$\begin{aligned} y_{ijk} &= \mu + A_i + B_{ij} + \gamma_k + \varepsilon_{ijk}, \quad \text{with} \\ & \quad A_i \sim \mathrm{N}(0, \sigma_A^2);\ B_{ij} \sim \mathrm{N}(0, \sigma_B^2);\ \varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)\,, \end{aligned}$$

where A_i is the random effect of section i, B_{ij} is the random effect for the jth subsample of section i, γ_k is the effect of method k, and ε_{ijk} is the error term. The variations between and within sections are σ_A^2 and σ_B^2, respectively, whereas the error variance represents the variability associated with measurements after adjusting for overall method differences (in γ_k).

C)

The value 0 contains valid information, therefore it is inappropriate to remove it. To replace it by 0.01 is most likely problematic because that value is very much smaller than all non-zero values, and it will most likely become a strong outlier after log-transformation. Replacing 0 by the lower detection limit would seem acceptable, whereas using the lowest observed value non-zero value probably results in a too large value (unless that value is below the detection limit). Finally, the average concentration for Section 8 is an inadequate replacement for a 0 value because it does not reflect the true meaning of the measurement. If the lower detection limit is not known, it could be suggested to replace 0 by a value more or less midway between zero and the lowest non-zero value, or another value in that range that results in a not too extreme residual in the log-scale analysis.

D)

The statistical model in the Minitab and Stata listings is indeed the model described in part B). We will compare the two scales with regard to how well the model assumptions appear to be met. The normal plot is more straight on log-scale. The plot of standardised residuals against fitted values from original scale looks strange, with all fitted values in the range $0 - 80$, except for a single pair of fitted values around 300 with the most extreme residuals. The plot from the analysis on log-scale looks more conventional and shows no major problems. Two concerns arise from the original scale plot: heteroscedasticity is suggested because the most extreme residuals were associated with the largest fitted values, and the two points with extreme fitted values may be very influential. In the data listing, these points can be identified as subsample B of section 2, with much larger values than all other measurements. The fact that both methods produced a large value indicate that these values are probably not errors. The estimated variance for subsamples is much larger than the other variance components, and this must be caused by these two values. Presumably this means that the distribution of estimated subsample random effects has a large outlier for subsample B of section 2. In summary, several problems and concerns with model assumptions have been identified for the analysis on original scale, leading us to the log-scale as the preferred scale.

E)

We will base the discussion on the log-scale results. The ANOVA-based method in Minitab produced a negative variance component for sections, and as variances cannot be negative one would need to reset this value to zero. In this situation, the likelihood-based estimates from Stata will be more appropriate; however, conclusions are similar. There is a strongly significant ($P < 0.001$) difference between analytic methods, with method `CB` giving values 0.377 units larger on log-scale, corresponding to the `CB`-values being larger by an estimated factor of $e^{0.377} = 1.46$, i.e. almost 50% larger. The 95% CI for the `CB`/`BF` ratio is $(e^{0.188}, e^{0.566}) = (1.21, 1.76)$.

The total unexplained variance on log-scale equals $0 + 0.509 + 0.148 = 0.657$, of which $0.509/0.657 = 0.77 = 77\%$ resides at the subsample level and the rest within subsamples. No between-section variance was found. The between-subsamples variance component is strongly significant ($P < 0.001$). As the theory about the variability in aflatoxin concentrations in peanuts predicted little between-section variance and large between-subsample variance, it appears that the results of the study confirmed this theory.

F)

The analytical precision/variability is represented in the model by the measurement variance σ^2. As the model assumes this variance to be constant across the entire dataset, and hence also for the

two analytic methods, this model and the results from it cannot be used to explore whether the analytical precision differs between methods. One would need to either analyse data for the two methods separately or consider a model that allows for different error variances (such models can be fit in Stata).

Question 3.

A)

The study is purely observational and could be called a partial time series (because of the gaps between different years). It is used as in illustration in Ruppert and Carroll (1980), *J. Amer. Statist. Assoc.* **75**, 828-38. The statistical analysis shown for A) corresponds to a multiple linear regression model, which can be written as:

$$\texttt{salinity}_i = \beta_0 + \beta_1 \texttt{lag}_i + \beta_2 \texttt{water}_i + \beta_3 \texttt{year}_i + \beta_4 \texttt{period}_i + \varepsilon_i,$$

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 $\sim \mathrm{N}(0, \sigma^2)$. All predictors are modelled by linear effects, and their estimates can be summarized as follows:

- `lag`: positive slope, strongly significant,
- `water`: negative slope, moderately significant,
- `year`: positive slope, close to significant,
- `period`: totally non-significant.

The `lag` variable controls for the correlation with the previously measured value (autocorrelation), and probably has no direct interpretation. The negative slope for `water` means that higher river discharge lowers the salinity (the river contains freshwater, so that makes sense).

B)

The assumptions involved in the multiple linear regression model are independence, normality and homoscedasticity of the errors, and a linear relation between outcome and each predictor. The distribution of the standardised residuals does not agree well with a normal distribution (indicated both by the not very straight quantile plot and the significant normality test), perhaps mostly due to two quite large positive residuals, for obs. 9 (2.568) and 16 (2.900). An outlier test based on deletion residuals would for the most extreme value (3.562) give $P = 2 \times 28 \times \mathrm{P}(t(22) > 3.562)$. A t-distribution table gives $\mathrm{P}(t(22) > 3.505) = 0.001$, from which we get $P > 0.056$. This shows that the most extreme observation (16) is close to being a significant outlier.

The plot of standardised residuals against fitted shows no obvious pattern indicative of heteroscedasticity. One might want to carry out a statistical test to support this assertion, such as the Cook-Weisberg test (e.g., using `estat hettest` in Stata).

Observation 16 has the by far largest leverage (0.562), which exceeds the most liberal cut-off for extreme leverage ($3 \times 5/28 = 0.54$). The value for `water` for this observation is seen to be much larger than all other values in the data. The Cook's distance value (2.155) is also by far the largest for this observation, and indeed much larger than the usually recommended thresholds (1 or 4/28). The strong influence of this observation is undoubtably due to its extreme value for `water`. This could be confirmed by computing df-beta values for `water`.

C)

We first review the 3 additional analyses:

- The correlation matrix for outcome and predictors shows strong linear associations between the outcome and all predictors except `period`, a strong correlation between `lag` and `year` (not too surprisingly, the year differences also express themselves in the lagged salinity values) and fairly modest correlations between the other predictors.
- The regression model with categorical effects of `year` and `period` show a much improved model fit (e.g., the MSE is nearly halved), due to non-linear effects of both predictors: among the years, 1972 and 1973 were much lower than the later years which themselves do not show a clear linear trend, and nor do the periods show any clear linear trend: the estimates show some high and low values (relative to baseline) without any clear pattern.
- The regression model with a quadratic term for `water` also shows a much improved fit, and the quadratic term is strongly significant.

The additional analyses show that linear effects of `period` and `year` are inadequate, so all subsequent models should treat these predictors as categorical. The `period` effect may in the end be non-significant and in that case it should be removed (it was also non-significant in the unconditional associations). To start with, we should compute overall significance tests for the categorical predictors, either by representing the results in an ANOVA table or by specific tests (e.g., using `testparm i.year` in Stata).

In order to explore the impact of the influential observation no. 16 in expanded models, diagnostics should be recomputed and reevaluated. There may be a concern that this observation becomes even more influential in a model with a quadratic effect of water. It is also suggested to analyze with and without this observation. Therefore, the most obvious continuation of the analysis is to fit a model with categorical effects of `year` and `period`, a linear effect of `lag` and a quadratic effect of `water` — in two versions: with and without observation no. 16.