Biostats VHM 802 Course
Winter 2021, 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. It also includes more than the two required options for Question 2.

Question 1.

We use the following notation,

y_{ijk} = FEV1 measurement for patient i on test day j after treatment k,

where $i = 1, \ldots, 16$; $j = 1, \ldots, 4$; k =A,B,C,D, and not all combinations of (i, j, k) occur in the data.

A)

The design is a cross-over design where each patient is subjected to all 4 treatments at different times; thus, the experimental unit is a treatment/observation period (for a patient). The advantage of a cross-over design is that the treatments are compared within subjects (instead of between subjects) and therefore are unaffected by a potentially large between-subject variation. The orders in which the subjects go through the treatments is determined by the design. Inspection of the data layout reveals that the first four patients form a Latin square design, with rows ∼ patients, columns ∼ test days, and symbols ∼ treatments. In the same way, the other three sets of 4 patients form Latin squares as well. The advantage of a Latin square design is that it is balanced in rows, columns and symbols which allows a complete separation of the corresponding effects. It is also observed that the same Latin square was used for all four sets of patients. This particular Latin square has the property that every treatment is preceded by any other treatment exactly once. For example, treatment A is preceded by D for patient 1, by B for patient 2, is the first treatment for patient 3, and is preced by treatment C for patient 4. This balancedness could be useful if there was any carry-over effect from one treatment to the next.

Randomisation seems to have been done by randomly assigning patients to the rows $1 - 4$ within its Latin square, maybe after first randomly assigning patients to Latin squares (this is not necessary for a valid randomisation). No information is provided about which patients are women and men; gender could be built into the design (it probably was not, with the gender imbalance).

The data structure is longitudinal, or repeated measures on the same patient over time. The series is fairly short with only 4 measures on each patient, so one might view the data structure as hierarchical, with FEV1 measurements within patients. A hierarchical diagram would look as follows:

The graph is a profile plot showing the values of individual patients over time. There are fairly large differences between patients, both in their overall level and in their profile over time. Some patients have roughly the same FEV1 value at all four test days whereas other patients exhibit strong variations. There does not seem to be any general trend over time. Treatment effects are difficult to see from the graph because the patients have different sequences of treatments.

B)

The statistical model analysed in the listings is the following:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk},$$

with the ε_{ijk} independent and $\sim \mathrm{N}(0, \sigma^2)$. The α_i are patient effects, the β_j are test day effects, the γ_k are treatments, and all these effects are fixed effects. The hierarchical diagram suggests to take patient effects as random, but as all (current) predictors are at the lowest level and the data are completely balanced this would not change the inference for these predictors.

The residual plot (standardised residuals against fitted values) shows some indication of more variation to the left than to the right (an (inverse) "fan" or "cone" shape). The normal plot shows some deviations from the straight line, primarily at both ends of the distribution. The list of unusual observations (and the plots) show two standardised residuals beyond $(-3, 3)$. It seems that the tails in the residual distribution are too strong compared to a normal distribution. Two supplementary analyses are suggested for diagnostic purposes. The first one is to try the Box-Cox procedure for determining a power transformation of the outcome; the residual plot suggests that a power > 1 is to be expected. The second one is to inspect the two observations with extreme residuals, and perhaps evaluate their statistical significance as outliers using the deletion residuals. Compared with the overall range of FEV1 measurements, the observed values of 1.0 and 2.4 do not seem particularly extreme. Note that one should await any decision about transformation before assessing the significance of outliers.

One additional analysis of clear interest is for a model with a tx by test day interaction, because that interaction may include a carry-over effect from one treatment to the next. In order to fit such a model one would need to take patients as random effects. This is also true for inclusion of gender effects. Carry-over effects could also be incorporated by direct modelling, similar to the milk yield example (13.12) of the Oehlert text.

C)

The ANOVA table shows a strong patient effect ($P < 0.0005$) as to be expected from the graph, and also biologically because no attempt seems to have been made to select a homogeneous group of patients. There is no indication of a difference between test days, with a P-value much larger than 0.05. This was also expected from the graph. The overall comparison between the 4 treatments is non-significant at $P = 0.18$. Despite this non-significance, we would want to also assess the specific drug dose and formulation effects, which were most certainly pre-planned. Pairwise comparisons between the four treatments are not going to help (in particular if adjusted for multiple comparisons), due to the non-significant overall test. We can instead do contrasts, thereby effectively splitting the treatment variation according to the underlying 2×2 factorial (formulation × dose). That is, we form the contrasts for the main effects of dose and formulation, and their interaction. This could be done by model specification: define the factors dose and formulation appropriately, and substitute `tx` by `dose formul dose*formul`. From the information provided, we can work out the contrasts directly:

$$\begin{aligned}
\text{dose} \quad &: \quad \hat\theta = (\hat\gamma_B + \hat\gamma_D) - (\hat\gamma_A + \hat\gamma_C) = 0.044, \\
&\quad \mathrm{SE}(\hat\theta) = \mathrm{MSE}\sqrt{(1^2 + 1^2 + (-1)^2 + (-1)^2)/16} = 0.159, \\
&\quad t = \hat\theta / \mathrm{SE}(\hat\theta) = 0.28, \quad P \gg 0.05, \\
\text{formulation} \quad &: \quad \hat\theta = (\hat\gamma_C + \hat\gamma_D) - (\hat\gamma_A + \hat\gamma_B) = 0.356; \quad \mathrm{SE}(\hat\theta) = 0.159, \\
&\quad t = \hat\theta / \mathrm{SE}(\hat\theta) = 2.24, \quad P = 0.030 \quad (\text{in } t(42)), \\
\text{interaction} \quad &: \quad \hat\theta = (\hat\gamma_A + \hat\gamma_D) - (\hat\gamma_B + \hat\gamma_C) = 0.019; \quad \mathrm{SE}(\hat\theta) = 0.159, \\
&\quad t = \hat\theta / \mathrm{SE}(\hat\theta) = 0.12, \quad P \gg 0.05.
\end{aligned}$$

The analysis shows there is virtually no interaction between dose and formulation, that the dose effect is clearly non-significant, and that the effect of formulation is weakly significant. The solution formulation seems to give about 0.18 units (0.356/2) higher FEV1 measures than the suspension formulation. Note that the standard errors for the contrasts are all equal, and could also be obtained from the standard errors for group means by multiplication by $\sqrt{4}=2$ (because 4 means are involved).

Question 2.

This solution will only give answers for a few options, roughly corresponding to those chosen by the class for the exam, specifically the selections **a1**, **b1** and **c1**. The **d1,d2** objectives were not selected by any student. The idea behind those options was that a dimension-reduction technique could be used for one set of variables to come up with a few variables, which could then be regressed on the other set of variables (possibly also after a dimension-reduction for the second set), with multiple linear regression (with independent components, MANOVA does not offer real advantages). The challenge with such approaches would be to arrive at reduced sets of variables with good interpretations.

As a descriptive analysis of the new (protein source) variables would be included as background for many of the analyses, we begin with this part. The obvious starting point is a matrix plot, possibly with country groups indicated (Figure 1; in order to facilitate the presentation of the text, all major figures, such as this one, are deferred to an appendix). Several variables show a visual separation between Eastern and Western countries, so descriptive statistics are of interest both split by groups and combined.

Group	Statistic	rm	wm	egg	milk	fish	cer	stch	pno	fveg
Eastern	mean	7.38	7.00	2.13	12.1	1.38	44.9	3.63	4.00	3.75
	std.dev.	2.26	3.70	0.84	3.94	1.19	8.68	2.00	1.85	1.49
Western	mean	11.1	8.13	3.50	20.3	5.69	26.5	4.56	2.75	4.44
	std.dev.	3.30	3.79	0.97	6.87	3.46	5.92	1.32	2.02	2.16
Combined	mean	9.88	7.75	3.04	17.5	4.25	32.6	4.25	3.17	4.21
	std.dev.	3.46	3.72	1.12	7.13	3.54	11.1	1.60	2.01	1.96

The variables `rm`, `milk` and `cer` show substantial differences between Eastern and Western countries. The spread of the different variables varies quite a bit, by a factor of more than 5 within groups. Some variables appear very discrete, due to their integer values in a narrow range, but all variables are quantitative. The distributions appear reasonably (not perfectly) symmetric without any glaring outliers. Correlation coefficients, for simplicity without splitting by country groups, are shown below. Correlations are not very strong, or of consistent sign, ranging roughly between -0.7 and 0.7.

```
      |       rm       wm      egg     milk     fish      cer     stch      pno     fveg
------+-------------------------------------------------------------------------------
   rm |   1.0000
   wm |   0.2206   1.0000
  egg |   0.6069   0.5857   1.0000
 milk |   0.5361   0.3546   0.6659   1.0000
 fish |   0.0702  -0.2128   0.0410   0.1736   1.0000
  cer |  -0.5333  -0.4230  -0.6944  -0.6349  -0.5166   1.0000
 stch |   0.2032   0.2821   0.3827   0.2974   0.4503  -0.5697   1.0000
  pno |  -0.4466  -0.6554  -0.5803  -0.6911  -0.1158   0.6268  -0.4601   1.0000
 fveg |  -0.0668  -0.0702  -0.1626  -0.4109   0.2307   0.0397   0.0802   0.3550   1.0000
```

a1: *relationships/patterns among protein source variables*

The matrix plot and table of correlations showed pairwise relations between all variables, but we aim for simpler descriptions. One option is to explore dimension-reduction techniques, for the purpose of determining which variables have similar loadings on the main components/factors; note that the countries and country groupings are not of direct interest for this objective. Another option is to transpose the data so that the 9 protein source variables become the observations and use distance-based methods, such as hierarchical cluster analysis, for those. Because the Euclidean distance (squared) for standardized variables becomes proportional to 1 minus the correlation coefficient, an analysis based on this distance can be done directly without transposing, and this is what happens in Minitab's "Cluster Variables" menu. Two resulting dendrograms, for single and average linkage, are shown in Figure 2. As one major approach for objective **b1** will be a cluster analysis, we devote most of our effort here to the dimension-reduction techniques.

Analyses should be based on the correlation matrix, due to the different spread of the variables. The total variable was not included with the data, thereby avoiding singularity of the correlation matrix. The listing below shows the PCA eigenvalues. The PCA does not in itself make any assumptions, apart from the correlations being meaningful.

Eigenvalue	4.1683	1.6516	1.0158	0.8659	0.4387	0.3560	0.2587	0.1352	0.1097
Proportion	0.463	0.184	0.113	0.096	0.049	0.040	0.029	0.015	0.012
Cumulative	0.463	0.647	0.760	0.856	0.904	0.944	0.973	0.988	1.000

It takes four eigenvalues to exceed 80% of the variance, but the fourth eigenvalue is below 1. This points towards considering three or four components for further analysis. The table below shows the PCA loadings (eigenvectors) as well as the loadings obtained after varimax rotation (with Kaiser normalization) in a factor analysis based on 4 factors. Factor analysis is used here as a descriptive tool to enhance interpretability of the components/factors. Loading plots for the first two components/factors are shown in Figure 3, but at least three components should be considered. With four factors, all communalities are fairly high (with three factors, `stch` and in particular `rm` are not well explained).

Variable	PCA loadings				FA loadings				Commun-
/Components	1	2	3	4	1	2	3	4	ality
`rm`	0.323	-0.039	-0.092	-0.700	0.930	-0.059	-0.031	-0.005	0.870
`wm`	0.307	-0.255	0.608	0.241	0.184	-0.943	0.042	-0.020	0.926
`egg`	0.415	-0.106	0.151	-0.245	0.714	-0.518	-0.142	-0.113	0.811
`milk`	0.402	-0.133	-0.338	-0.056	0.634	-0.247	-0.265	-0.537	0.821
`fish`	0.130	0.657	-0.320	0.115	0.088	0.276	-0.898	0.088	0.898
`cer`	-0.424	-0.256	0.037	0.046	-0.586	0.337	0.629	0.089	0.860
`stch`	0.288	0.391	0.137	0.396	0.066	-0.414	-0.759	0.027	0.753
`pno`	-0.417	0.143	-0.020	-0.293	-0.367	0.639	0.308	0.441	0.832
`fveg`	-0.122	0.484	0.602	-0.363	-0.033	0.035	-0.153	0.952	0.931

The first component and factor give similar (negative) loadings to `cer` and `pno`, contrasting the other variables. The first factor also has similar (positive) loadings for `rm`, `egg` and `milk`, which for the first PCA component are joined by `wm` and `stch`. These findings indicate similarities between the correponsponding variables, which is also confirmed by the dendrograms of the cluster analysis. The second PCA component has high loadings on `fish`, `stch` and `fveg`, but the former two instead show up with similar loadings on the third factor, and `fveg` loads strongly on the fourth factor. In turn, the second factor has its strongest loading for `wm`, which was one of the main contributors to the third PCA component. These patterns are not so easy to translate into relations among the variables, but

we can say that a similarity between `fish` and `stch` is indicated, and that the strong loading on the fourth factor for `fveg` could indicate this variable is separated from the rest. The latter interpretation agrees with its low correlations with the other variables, as well as its position in the dendrograms.

b1: *relationships/patterns among countries based on protein source variables, with interpretation in the context of relevant groups*

Analyses here will use the same data, but the focus will be on the countries instead of the variables. The natural choices for analysis are different types of ordination, where we in this solution will cover classical and modern MDS, and cluster analysis, where we will include hierarchical cluster analysis but, in order to keep this solution reasonably short, skip over K-means clustering. All these methods are based on distances between countries, again preferably computed from standardized variables. The relevant grouping for the protein source variables is the division into Eastern and Western European countries, and group membership can be indicated by symbols on relevant plots.

We start by revisiting the PCA from the first part, now interpreted as a classical MDS. Score plots for the first two components are shown in Figure 4. A pretty good separation between Eastern and Western countries is obtained, with Eastern countries mostly low on the first component ($\sim$ high on `cer` and `pno`) and not high on the second component ($\sim$ not high on `fish` and `fveg`), with Western Mediterranean countries low on the first but higher on the second component, and most of the other Western European countries higher on the first component. As before, one should also look at higher components (at least include the third one), and one could also try the factor scores after varimax rotation, but neither of these additions seem to add substantially to the separation already obtained (not shown).

The default modern MDS (with stress strain and no transformation) has a fairly large loss value of 0.196, suggesting that improvements can be achieved with other versions or inclusion of higher dimensions. Non-metric MDS with monotonic transformation does indeed produce a lower loss of 0.122, but the patterns from its first two components are similar. Adding a third dimension does not substantially lower the loss. Figure 5 shows the country configurations for the standard MDS and the non-metric MDS; in terms of separating the Eastern and Western European countries these solutions do not offer much improvement over the pretty good results of classical MDS (or PCA).

We finally show results of hierarchical cluster analysis based on Euclidean distance for standardized variables. Among the three linkages discussed in the course, average and complete linkage give similar results that in turn differ somewhat from those of single linkage. Figure 6 displays the dendrograms for the latter two of these linkages. The most satisfactory clustering of the countries is perhaps offered by complete linkage, which separates out the four Mediterranean countries (with neighbouring Spain and Portugal in close distance) as well as two groups of Eastern European countries, that can even be interpreted as Balkan countries versus more continental countries. Additionally, the Scandinavian countries appear in a separate cluster. With single linkage, there is no separation of Czech/Slovakia and Poland from a group of Western Central European countries, and Albania is split out from the Balkan Easter European countries (with its place taken by Hungary); these modifications of the dendrogram also make good sense from the geographical location of the countries.

In summary, both ordination and hierarchical clustering methods managed to show differences between Eastern and Western European countries, and also showed closer relationships between Mediterranean countries (irrespective of the East-West dichotomy). One advantage of the PCA ordination is the direct interpretation of its components in terms of the loadings of the variables, but its loss of information (by focusing on the first few components) may have offered a less detailed picture of the clusters in the data.

c1: *ability to determine relevant group membership based on protein source variables*

For this objective we should try to predict group membership from the protein source variables, rather than to explore the patterns these variables generate among the countries (and compare them to country groups). That is, the group membership should be part of the analysis, making it a (supervised) classification problem. With only two groups, logistic classification is an obvious option, but we can also try linear discrimination or the kth nearest neighbour method (not in the course syllabus). MANOVA with `group2` as a predictor (equivalent to Hotelling's T2), and the corresponding univariate two-sample t-tests, will also utilize the division into groups, but these methods do not directly lead towards predicting group membership, so they perhaps fit better under **a1** as a way to describe the variables.

With only 24 observations (countries) and 9 variables, it is not surprising that perfect classification can be achieved, when evaluated on the same data; both linear and logistic discrimination manage this, with uniform priors. For such a small sample it is however more relevant to use leave-one-out cross-validation to assess the performance of classifiers. Still with uniform priors, LDA misclassifies Albania and Poland as Western European countries. Logistic classification misclassifies Poland as a Western European country, and runs into estimation problems for the sample without Spain. By fitting this model manually, it can however be determined that the predicted classification for Spain is correct. The misclassifications and the difficulty in obtaining predictions show that the prediction problem is bordering on being too extreme ("too easy") for the classifiers to work normally: 9 predictors for 24 observations is simply too much. It can be suggested to use a dimension-reduction technique to extract fewer, and independent, predictors. This would clearly go beyond what is expected for an exam.

Question 3.

Denote by y_i the content of vitamin B2 turnips harvested at plot i, $i = 1, \ldots, 27$, and by $\texttt{RA}_i$, $\texttt{MT}_i$ and $\texttt{AT}_i$ the three predictors measured at each plot as well (explained in the question).

A)

The multiple linear regression model using these 3 predictors is

$$y_i = \beta_0 + \beta_1 \texttt{MT}_i + \beta_2 \texttt{RA}_i + \beta_3 \texttt{AT}_i + \varepsilon_i, \tag{1}$$

where β_0 is the intercept, the other β's are regression coefficients corresponding to the respective predictors, and the ε_i's are the errors which we assume to be i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$. Brief parameter interpretations: $\hat{\beta}_0 = 82.1$ is the intercept, corresponding to the value when all 3 predictors equal 0; this is clearly not a relevant scenario, and corresponds to a strong extrapolation from the data. Each of the regression coefficients $\beta_1 - \beta_3$ correspond to the change in the predictive equation resulting from a one unit change of the corresponding predictor, when the two other predictors are held constant. The meaning of a one-unit change depends on the scale of the predictor, and as these are different the three regression coefficients are not directly comparable by their magnitude. The estimate $\hat{\beta}_2 = -0.78$ is negative, corresponding to a decrease in vitamin B2 content for an increase in `MT`, whereas the two other predictors are associated with increases in vitamin B2 content for increasing values. Finally, the error standard deviation $\hat{\sigma} = 9.9$ quantifies the unexplained variation about the predictive equation.

The assumptions of the model are

i) normality, i.e. normal distribution of the errors ε_i,

ii) homoscedasticity, i.e. same variance (σ^2) of all the errors ε_i,

iii) linear equation, i.e. zero mean (no bias) of the errors ε_i,

iv) independence of the errors (and of the observations).

B)

Assessments of the validity of the model's assumptions i)–iv) are made from the residuals plots.

i) The normal plot looks reasonably straight, the P-value is >0.10, and there are no extreme outliers (2 out of 27 standardized residuals exceed ± 2, and the most extreme is -2.34, which even without knowing the value of the corresponding deletion residual can safely be predicted to not test significant using the Bonferroni procedure).

ii) The plot of residuals against fitted values looks strange but this is largely due to two groups of predicted values (indicating that some groups exist in the data). The variation seems somewhat larger at higher predicted values but this could also be because there are more points. We conclude that there is no evidence to indicate heteroscedasticity.

iii) The plots against the predictors may indicate problems with the assumption of linearity. The plots against `RA` and `AT` look fine (no particular pattern seen). The plot against `MT` shows that there are in fact only 3 different `MT`-values present in the dataset (`MT` $= 2, 7, 47.4$; as seen in the data listing). Furthermore, the fit at `MT` $= 7$ does not look good, the residuals are not centered at zero but positively biased. Some negative bias is also seen for `MT` $= 2$. This shows the linear relation assumed for `MT` to be unsatisfactory. With only 3 and very non-equidistant values it would anyway be more natural to model `MT` as a categorical variable. Even without access to the full data, it is clear that the means for the 3 groups are very different: about 100 for `MT` $= 7$, about 90 for `MT` $= 2$ and closer to 60 for `MT` $= 47.4$.

iv) The last plot shows a negative trend, within each `MT`-group, in the plotted residuals against observation number, indicating observation number to be a potential confounder. It is, however, not interpretable for us as we don't know in which order the observations are listed. It should be suggested to the experimentator to check the data and explore this potential effect.

Given the problems we identied with model (1), it is difficult to draw definitive conclusions from the results and we will have to make a reservation for any changes that might occur after the model has been modified. The variables `AT` and `RA`, that both seem to be modelled appropriately in the model, show clearly non-significant parameter estimates. This does *not* by itself imply that they are both redundant (because of possible collinearity), but the sequential sum of squares shows how to test the effect of `RA` after `AT` has been eliminated: $F(1, 24) = [116.9/1]/[(2243.2 + 30.5)/24] = 1.23$, which is clearly non-significant ($F(1, 24, .90) = 2.93$). Therefore, both `AT` and `RA` are non-significant in model (1). The regression coefficient for `MT` is not of much use because we argued that it should be replaced by categorical modelling. However, its estimate is clearly significant (and the sequential sum of squares shows that this would also be true in a reduced model), which suggests that `MT` would also be strongly significant as a categorical predictor. If the other predictors remain non-significant, and the observation number is not included in the model, we could base our inference only on the `MT` group means (given above). The error variance can be expected to decrease when we improve the model, but using the value $s = 9.9$ as a rough (over)estimate gives us a standard error of a group mean of $s/\sqrt{9} = 3.3$. With differences in group means of up to about 40, there is no doubt that `MT` is clearly significant as a categorical predictor in a univariable analysis.

C)

The most important extra analysis is for a multiple regression model with `MT` as a categorical variable. Further analyses would depend on the results obtained in that model, both respect to the validity of model assumptions and significance of the effects. We can test by an F-test whether model (1) is a non-significant reduction, but it does not seem likely. If the new model is valid, we should continue with backwards elimination. When a final model has been reached, we should add interaction terms between the significant effects, as well as quadratic terms for effects modelled as continuous (`AT` and `RA`). It might also be worthwhile to compute a matrix of simple correlations between all predictors, to check for collinearity and the univariable associations with the outcome. Last but not least, the potential confounding effect of observation number should be explored. The simplest way is to add observation number as a continuous predictor, and check its significance. Whether it should be actually included in the final model (or in the model reduction process), depends on the interpretation of this variable.

Appendix: Figures for Question 2

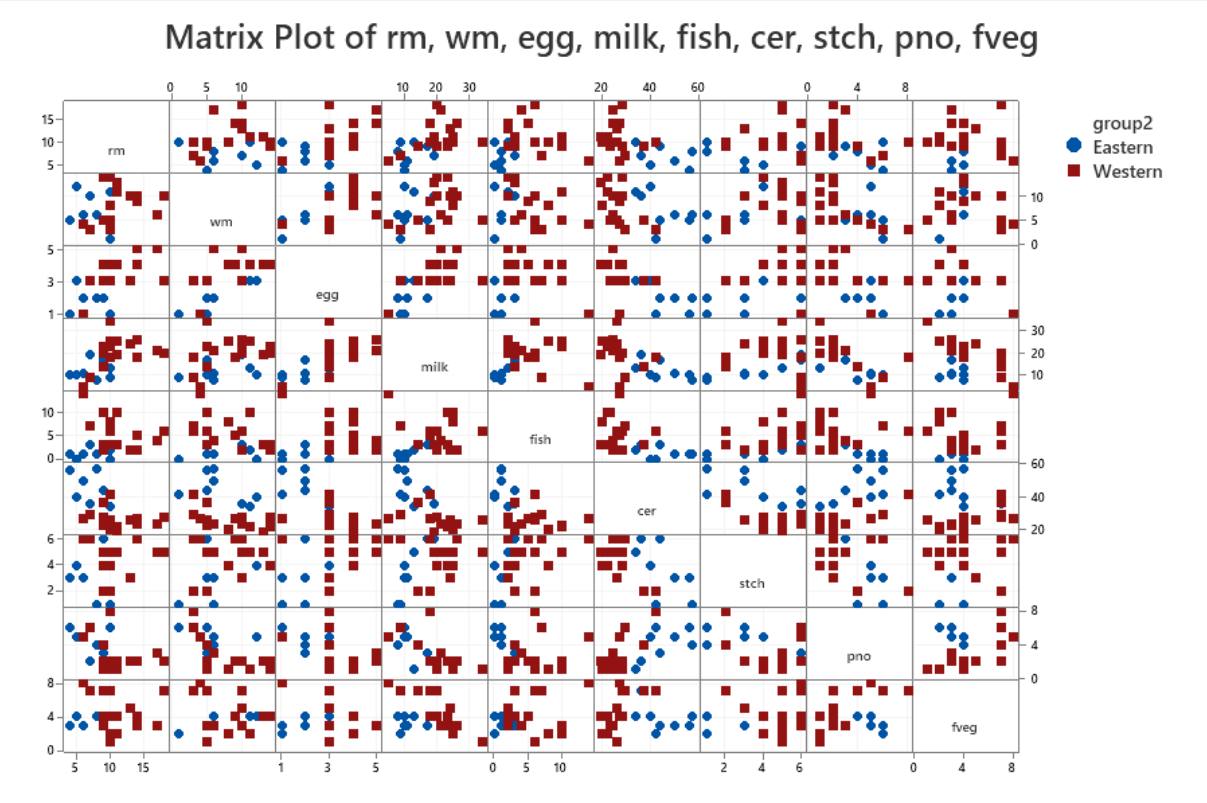

Figure 1: Matrix plot for protein source variables, with different plotting symbols for Eastern and Western countries.

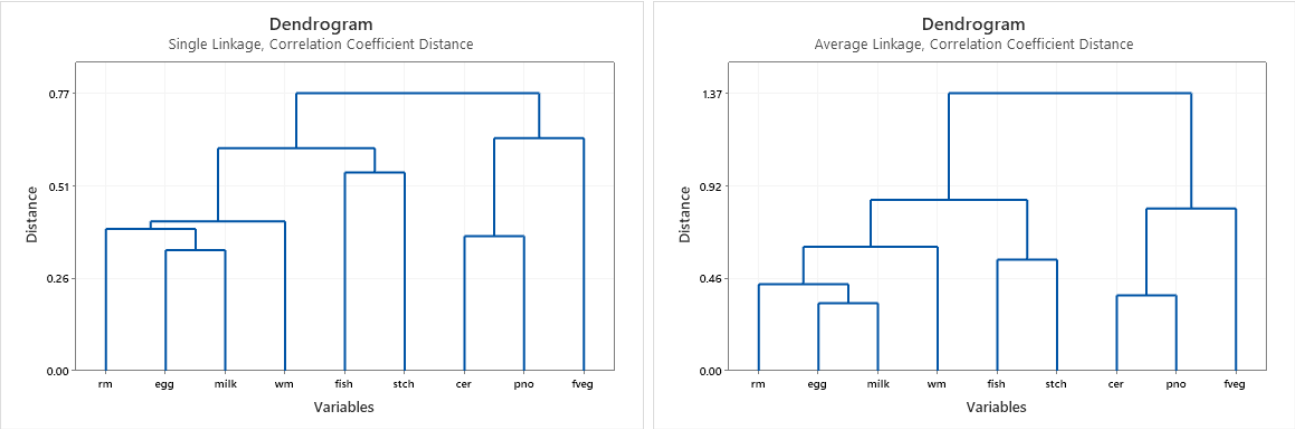

Figure 2: Dendrograms for hierarchical cluster analysis for protein source variables, based on Euclidean distance between standardized variables.

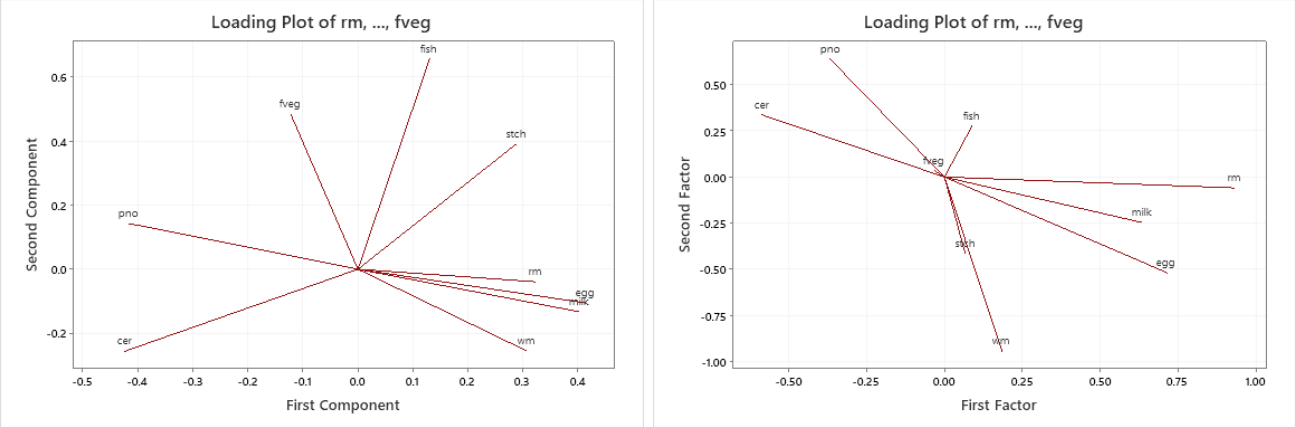

Figure 3: Loading plots for first two components (left)/loadings (right) of PCA/FA for (standardized) protein source variables.

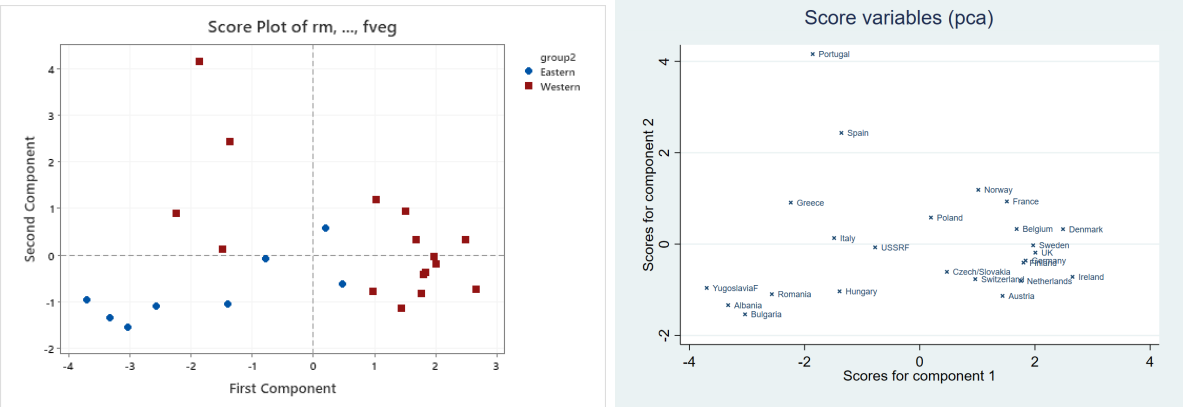

Figure 4: Score plots for first two PCA components, with country groups (left) and country names (right) indicated.

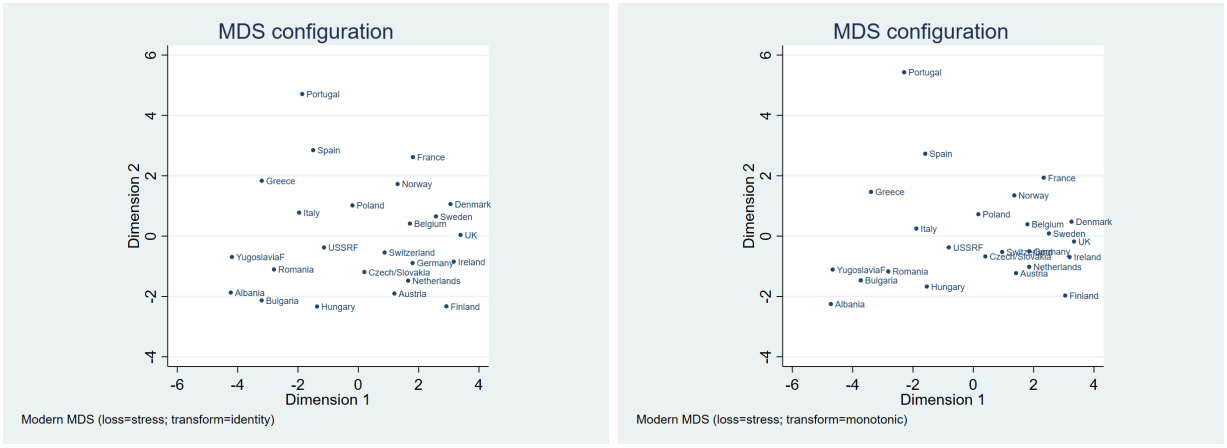

Figure 5: Ordination plots for the first two components of two versions of modern MDS.

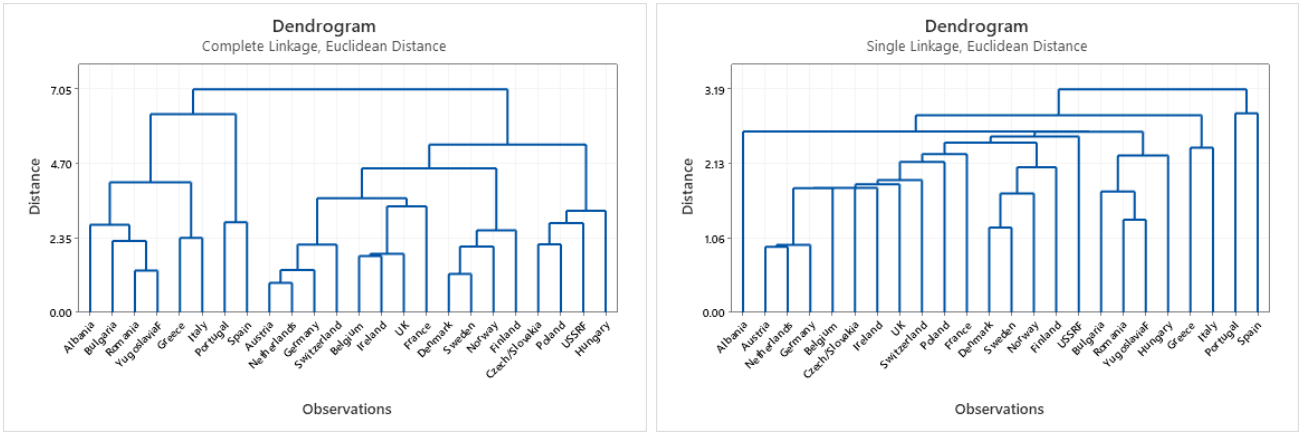

Figure 6: Dendrograms for hierarchical cluster analysis for countries, based on Euclidean distance from standardized protein source variables.