

Cluster analysis (variables)

Clustering variables and some (fabricated) binary data

So far we have only looked at clustering cases. We have also only looked at dealing with interval data. So, while we have a brief look at clustering variables, we also use a different kind of data, namely binary data. Once again we have fabricated it to bring out an important point. There is a range of symptoms associated with TBI. Various combinations of symptoms may be present, and it may be that an exploration of what combinations (or clusters) of symptoms tend to occur together would be revealing. From a long list of possible symptoms, we have selected just eight for the purpose of this example analysis.

sypmt1	fatigue
sypmt2	feelings of helplessness
sypmt3	disorientation
sypmt4	confrontational attitude
sypmt5	blurred vision
sypmt6	explosive temper
sypmt7	depression
sypmt8	irritability

Table 10.5 gives the first few rows of presence (coded 1) and absence (coded 0) data on eight symptoms (in reality, there would be many more) for 50 patients who have recently suffered mild to moderate TBI (the full dataset is `med.cluster.variables.sav` on the book website). Three cluster analyses are carried out on these data: (1) the average (within group) linkage method is used with the simple matching coefficient as the distance measure, (2) the same method is used with the Jaccard coefficient as the

distance measure and (3) the average (between groups) linkage method is used with the simple matching coefficient as the distance measure, as in the first analysis.

Table 10.5

The first few rows of presence/absence of 8 symptoms in 50 patients (full set is med.cluster.variables.sav on the book website)

sympt1	sympt2	sympt3	sympt4	sympt5	sympt6	sympt7	sympt8
0	0	0	1	0	0	0	1
0	0	0	1	0	1	1	0
0	0	0	0	0	1	0	1
0	0	0	1	0	1	0	0
0	0	0	1	0	0	1	1

Binary variables: requesting the cluster analysis

Looking at the table of presences and absences certainly does not convey much impression of how the variables may be related, if they are. To see whether cluster analysis throws any light on this, we once again used **Analyze, Classify** and **Hierarchical Cluster** to get SPSS Dialog Box 10.1. We entered all of the variables, sympt1 – sympt8 into the **Variables** box and, this time, we clicked the radio button for **Cluster Variables**. We then clicked the **Plot** button and again asked for a **dendrogram** and omitted the icicle diagram. For the Cluster **Method**, we chose **Within-groups linkage** (SPSS Dialog Box 10.3), and clicked the **Binary** radio button in the **Measure** group. We used the drop down list to choose a similarity measure (which will be converted to a distance measure). There is a long list, as mentioned in the discussion of distance measures in the introduction.

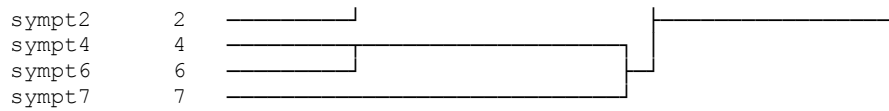
Distance measures: simple matching coefficient

We began by choosing the **Simple matching coefficient**. SPSS Output 10.8(a) shows the resulting dendrogram. This suggests that symptoms 1 and 8 may be linked, also 3, 5 and (less closely) 2. Symptoms 4 and 6 also make a pair, but not grouped as closely

as 1 with 8, or 3, 5 and 2. Symptom 7 joins the 3, 5, 2 cluster but only just before it is joined by the loose 4 and 6 pair. Looking at the list of symptoms, these clusters do seem quite plausible. Symptoms 1 and 8 are linked, which makes intuitive sense; a physical state of fatigue may lead to the emotional state of irritability. Also symptoms 3, 5 and 2 are linked and it does seem plausible that the physical state of blurred vision may be associated with cognitive disorientation and, eventually, feelings of helplessness. Symptom 7, depression, joins the 3, 5, 2 cluster just before it is joined by the 4 and 6 pair of symptoms (confrontational attitude and explosive temper).

Distance measures: the Jaccard coefficient

But what happens if we use a different similarity/distance measure? SPSS Output 10.8(b) shows the dendrogram obtained using the Jaccard coefficient. Apart from the cluster of symptoms 1 and 8, this looks very different from SPSS Output 10.8(a). The choice of distance measure for binary data is particularly difficult and, as here, can have a marked effect on the results. In this case, many people would suggest that the simple matching coefficient is more appropriate than the Jaccard, because none of the symptoms is rare. Symptom 1 occurs in 12 out of 50 cases, and all others are more common than that. Because this technique is exploratory, if your results help you to understand your data you do not need to be too concerned if you do not have clear justification for your choice of distance measure or linkage method. But of course if results that help your understanding are not robust to changes in the choices you make, you must be careful not to claim too much for them, but only use them as a guide to further work.



SPSS Output 10.8. Dendrograms for variable clusters using different distance measures and clustering methods

It is of course possible to cluster the cases here as well as the variables, and consideration of each cluster of patients may suggest helpful interventions or better ways to understand the symptoms. If so, then the technique has been useful, whether or not you can fully justify your choice of distance measure or linkage method.