## 0.1 AIC and BIC in Two-Step Cluster Analysis

(Removed from Last Week's Class due to Version Update)

Two-Step Cluster Analysis guides the decision of how many clusters to retain from the data by calculating measures-of-fit such as ***Akaike2s Information Criterion (AIC)*** or ***Bayes Information Criterion (BIC)***.

These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments.("Relative" means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value.)

***Important***: Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate an increased fit.

SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion. However, which criterion should we choose?

- AIC is well-known for overestimating the correct number of segments
- BIC has a slight tendency to underestimate this number.

Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results.

Once you make some choices or do nothing and go with the defaults, the clusters are formed. At this point, you can consider whether the number of clusters is "good". If automated cluster selection is used, SPSS prints a table of statistics for different numbers of clusters, an excerpt of which is shown in the figure below. You are interested in finding the number of clusters at which the Schwarz BIC becomes small , but also the change in BIC between adjacent number of clusters is small.

The decision of how much benefit accrued by another cluster is very subjective. In addition to the BIC, a high ratio of distance of measures is desirable. In the figure below, the number of clusters with this highest ratio is three.

*Autoclustering statistics*

| | | Schwarz's Bayesian Criterion (BIC) | BIC Change[1] | Ratio of BIC Changes[2] | Ratio of Distance Measures[3] |
|---|---|---|---|---|---|
| Number of Clusters | 1 | 6827.387 | | | |
| | 2 | 5646.855 | -1180.532 | 1.000 | 1.741 |
| | 3 | 5000.782 | -646.073 | .547 | 1.790 |
| | 4 | 4672.859 | -327.923 | .278 | 1.047 |
| | 5 | 4362.908 | -309.951 | .263 | 1.066 |
| | 6 | 4076.832 | -286.076 | .242 | 1.193 |
| | 7 | 3849.057 | -227.775 | .193 | 1.130 |
| | 8 | 3656.025 | -193.032 | .164 | 1.079 |
| | 9 | 3482.667 | -173.358 | .147 | 1.162 |
| | 10 | 3343.916 | -138.751 | .118 | 1.240 |
| | 11 | 3246.541 | -97.376 | .082 | 1.128 |
| | 12 | 3168.733 | -77.808 | .066 | 1.093 |
| | 13 | 3103.950 | -64.783 | .055 | 1.022 |
| | 14 | 3042.116 | -61.835 | .052 | 1.152 |
| | 15 | 2998.319 | -43.796 | .037 | 1.059 |

1. The changes are from the previous number of clusters in the table.
2. The ratios of changes are relative to the change for the two cluster solution.
3. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Figure 1: Schwarz Bayesian Information Criterion