5

Worked Example for Cluster Analysis

This example uses the ASA software integrated into Excel or SPSS (www.asastat.com). ASA is, in part, a point-and-click interface to R but analyses can be conducted from within SPSS or Excel. All data are hypothetical. We assume you have read the primer on cluster analysis.

We will first conduct a cluster analysis on our data using hierarchical clustering to illustrate that approach. Then we will apply centroid clustering to illustrate it. Two measures of parenting style were obtained for 750 mothers as reported by their adolescent children. One was a measure of maternal expressions of warmth and affection and the other was how controlling the mother is. They are in the variables called *warmth* and *control*. Each was measured on the same 0 to 10 metric with higher scores indicating higher warmth and higher control. Of interest is identifying "clusters" of parenting styles with respect to these two variables that may exist in the broader population from which the sample was selected. We did not standardize the variables prior to analysis because they are on the same metric.

The ASA software routinely reports confidence intervals for key parameters in statistical models. There are different ways of presenting confidence intervals. One strategy is to report them directly. Another strategy is to report them as margins of error, much like the margins of error you see for political polls on television or in print media. In this case, one calculates the half width of the confidence interval and reports it in "plus or minus" format. For example, in a political poll, you might be told that the percent of people endorsing a candidate is 50% ±5%. In this case, the confidence interval is 45% to 55%. This is an efficient way of summarizing the interval. In some cases, confidence intervals are asymmetric. When this occurs, some researchers will report the lower and upper margin of error separately. Alternatively, the researcher might calculate the absolute difference between the lower limit and the parameter estimate as well as the absolute difference between upper limit of the interval minus the parameter estimate and then report whichever difference is larger using the ± format. Some analysts prefer the use of credible intervals in Bayesian analytic frameworks instead of confidence intervals for characterizing margins of error (see Curran, 2005).

PRELIMINARY ANALYSES

Our first step is to gain a sense of variable distributions to determine if issues with unusual shapes and outliers will arise. These analyses are presented in the Appendix. All was in order, so we proceed accordingly.

HIERARCHICAL CLUSTERING

We decide to use an average link algorithm and Manhattan distance scores. The latter are the most intuitive of the possibilities and there is no strong reason to use an index (such as Euclidean distance scores) that give larger weight to larger disparities. I need to think *a priori* about what a reasonable cut-point for defining clusters is for the distance score. Based on past research, I decide that if two parents differ by 2 units on the 0 to 10 metric of a given dimension, then I should treat them as having distinct parenting styles on that dimension. Since I have two dimensions/variables and the Manhattan distance scores are based on a sum of disparities across the dimensions/variables, a total disparity of 2 + 2 = 4 is the initial cut-point I decide to work with. I select the option "Specify cut-off" and enter the value of 4 for the cut-off value.

The first part of the output shows the merging process at each iteration and information surrounding the merging. Here is the first portion of the output:

MERGING ITERATIONS

	NUMBER OF CLUSTERS	OBJECT 1	OBJECT 2	MERGING CUTOFF
Step 1	749	-214	-392	0.01121
Step 2	748	-190	-192	0.01226
Step 3	747	-459	1	0.02039
Step 4	746	-17	-188	0.022

At step 0, every individual is treated as a separate cluster, so there are 750 clusters because my sample size is 750. At the first step, individual 214 and individual 392 are merged into a cluster, resulting in 749 clusters. These two individuals had the smallest distance score among all 750 "clusters" and that distance score was equal to 0.011. The negative signs in front of the objects indicate that the two "clusters" that were merged consisted of singletons (one individual each). Looking down the list, at Step 3, individual 459 was merged with a multi-individual cluster as reflected by the absence of a minus sign for Object 2. The individual was merged with the cluster formed at Step 1. The distance score for the merge between "cluster" 749 and the cluster formed at Step 1 was 0.020.

We are not interested in clustering at such low values because my *a priori* interest is with individuals with distance scores of 4 or greater. Moving down the output, here is the section where merging cur-off distance scores are around 4:

	NUMBER OF CLUSTERS	OBJECT 1	OBJECT 2	MERGING CUTOFF
Step 744	6	232	737	3.82542
Step 745	5	731	743	3.86198
Step 746	4	740	742	4.09305
Step 747	3	739	744	5.13972
Step 748	2	745	746	5.23524
Step 749	1	747	748	6.75941

We can see that the value of 4 occurs between a 4 and 5 cluster solution. Here are the cluster sample sizes from the output for the five cluster solution that ended up being generated based on my initial program input:

CLUSTER SIZES

Cluster	N	Percent total N
1	197	26.267
2	320	42.667
3	4	0.533
4	143	19.067
5	86	11.467

One of the clusters (Cluster 3) is very small and reflects the fine grained difference between my *a priori* cut-off value of 4 and the value of 4.09 (see above) that defines a four cluster solution. We decide, based on this, to impose a 4 cluster solution on the data, thereby loosening our cut-off criterion ever so slightly. Before doing so, however, let's examine the dendrogram for the analysis, which appears in Figure 5.1.

The dendrogram is a branching diagram that shows similarities between objects and captures the merging process. The vertical (Y) axis is the distance score at which a merge occurs; the horizontal (X) axis represents the clusters. Because there are 750 individuals in the current example, there are 750 demarcations on the horizontal axis to represent the singleton clusters at step 0. This leads to a crowded and unreadable axis. It will be easier to explain a dendrogram using a simpler example with fewer individuals, so we divert for the moment to the example in Figure 5.2. This example clustered 5 individuals in successive steps. The objects are demarcated on the X axis strategically, not in the order they were input, to accommodate graphical characterization of the merging process. The black, upward lines are connected by black horizontal lines that indicate where a merge

has taken place. The midpoints of these horizontal lines are called ***clades*** and each clade has a distance score associated with it represented by the clade's location on the Y axis. A clade occurs where a merge takes place.

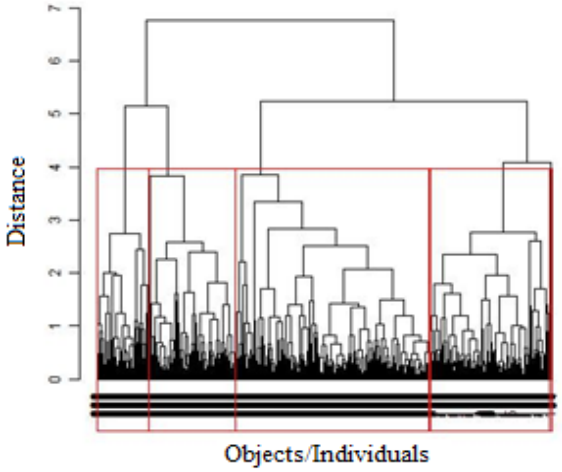


FIGURE 5.1. Dendogram for Example

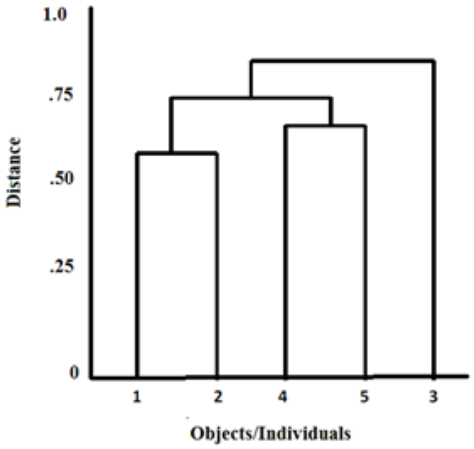


FIGURE 5.2. Simplified Dendogram

For example, per the diagram, individuals 1 and 2 were merged at the first step (reflected by the clade with lowest on the Y axis) at a distance score of about 0.55. At the next step, individuals 4 and 5 were merged at a distance score of about 0.65. At the next step, the first cluster of two individuals was merged with the second cluster of two individuals at a

distance score of about 0.75, creating a cluster of 4 individuals. At the final step, this 4 object cluster was merged with the remaining singleton cluster into one large cluster consisting of all objects.

Returning to Figure 5.1, the program adds red lines to the traditional dendrogram to demarcate the five clusters that resulted from the pre-specified cut-off value. One of the red lines covers an area so small you can hardly see it differentiated from the other red lines – it is the very small cluster. We personally prefer to examine the merging process data directly, as we did above, rather than rely on dendrograms.

We re-run the program but this time we check the box 'Specify number of clusters" and we enter the value 4 given my decision to use four clusters. Here are the cluster sample sizes that result:

```
CLUSTER SIZES

  Cluster     N          Percent total N

     1        201        26.800
     2        320        42.667
     3        143        19.067
     4        86         11.467
```

All of the clusters are reasonably sized. The program reports the cut-off distance score that was required to generate the four cluster solution:

```
CLUSTER CUTOFF DISTANCE VALUE TO ACHEIVE 4 CLUSTERS

  Cut-off distance value: 4.09305
```

which is consistent with what we saw in the merging iteration output.

The program plots the cases in each cluster with an ellipse around them. The plot uses the principal components strategy discussed in Pison, Struyf and Rousseeuw (1999), which fits the data using two components. The plot is reasonably trustworthy if the two principal components account for large portions of the variance in the target variables (***warmth*** and ***control***), and in this case, they do (the ASA program indicated the two components accounted for 100% of the variance). The plot is presented in Figure 5.3. The structure seems reasonably well articulated.

The program reports means and standard deviations for each cluster. Below is the output (the margins of error are based on 95% confidence intervals, but are only approximate – see the primer on cluster analysis for why):

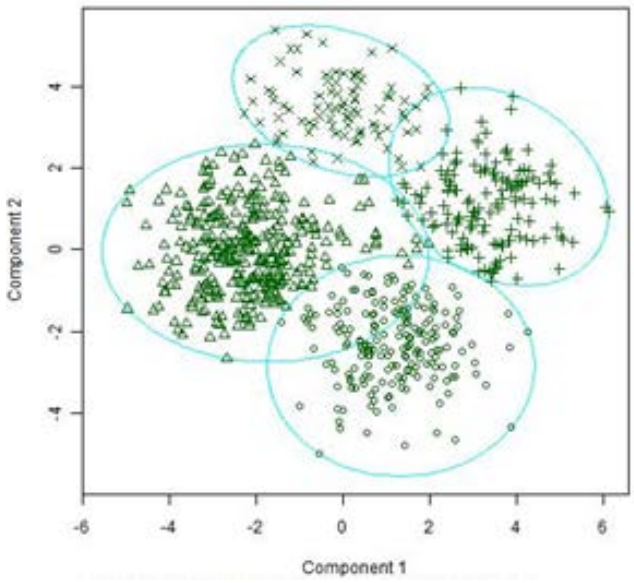


FIGURE 5.3. Cluster Plot

```
MEANS AND STANDARD DEVIATIONS FOR CLUSTERS

VARIABLE: WARMTH

  Cluster 1 Mean: 7.10265 +/- 0.13035
  Cluster 2 Mean: 6.91292 +/- 0.12166
  Cluster 3 Mean: 3.02865 +/- 0.15223
  Cluster 4 Mean: 2.81014 +/- 0.15989

  Cluster 1 Standard deviation: 0.93718
  Cluster 2 Standard deviation: 1.10613
  Cluster 3 Standard deviation: 0.92085
  Cluster 4 Standard deviation: 0.74574

VARIABLE: CONTROL

  Cluster 1 Mean: 7.00816 +/- 0.14059
  Cluster 2 Mean: 2.94286 +/- 0.11849
  Cluster 3 Mean: 7.06987 +/- 0.15899
  Cluster 4 Mean: 2.90733 +/- 0.22086

  Cluster 1 Standard deviation: 1.01082
  Cluster 2 Standard deviation: 1.07734
  Cluster 3 Standard deviation: 0.96179
  Cluster 4 Standard deviation: 1.03015
```

We find it helpful to create a table that places the means side-by-side, like this:

	Warmth	Control	Cluster Size
Cluster 1	7.10	7.01	201 (26.8%)
Cluster 2	6.91	2.94	320 (42.7%)
Cluster 3	3.03	7.07	202 (19.2%)
Cluster 4	2.81	2.91	301 (11.5%)

There appear to be four types of parenting styles, based on the above. First, there are parents who are both warm and controlling of their children, i.e., they are both affectionate and they seem to "look out for" and monitor their child (Cluster 1). We will refer to them as *authoritative* parents. Second, there are parents who are warm and affectionate with their children but who are low in control and supervision (Cluster 2). We will refer to these as *permissive* parents. Third, there are parents who seem to be fairly controlling of their adolescent child but who are not expressive of warmth (they are "cold" towards their child; see Cluster 3). We will call them *authoritarian* parents. Finally, there are parents who are relatively low in control and also low in warmth (that is, they seem to be disengaged – see Cluster 4). We will call them *neglectful* parents. Creating cluster labels is a common practice in cluster analytic applications. We often seek to give summary labels to each cluster that capture the gestalt of the mean patterns in a multivariate sense.

Writing It Up

Because of space limitations in journals, we do not have the liberty to describe the preliminary analyses and some of the checks we performed, but we would indicate in the Method section the general strategies we used for preliminary analyses and analytic checks and report that the analyses affirmed the use of the hierarchical procedure. Here is how we might write-up the results for the hierarchical analysis:

"A hierarchical cluster analysis was performed using unstandardized scores for warmth and control, an average link clustering algorithm, and Manhattan distance scores. Based on past research, a cut-off distance score of 4.0 was used to define clusters. An initial analysis using this criterion yielded five clusters, but one cluster was quite small (n = 4) and reflected a minor perturbation from the cut-off value of 4.09 that defined a four cluster solution. A four cluster solution was therefore imposed on the data. Table 1 presents the mean warmth and control values for each cluster as well as the within-cluster standard deviations. The four parenting styles implied by the clusters can be characterized as follows: First, there are parents who are both warm and controlling of

their children, i.e., they are both affectionate and they "look out for" and monitor their child (Cluster 1). They are termed *authoritative* parents and constitute 27% of the sample. Second, there are parents who are warm and affectionate but who are low in control and supervision (Cluster 2). They are termed *permissive* parents and constitute 43% of the sample. Third, there are parents who seem to be fairly controlling of their adolescent child but who are not expressive of warmth (i.e. they are likely somewhat "cold" to their child; see Cluster 3). They are hereafter referred to as *authoritarian* parents and constitute 19% of the sample. Finally, there are parents who are relatively low in control and also low in warmth (Cluster 4). They seem disengaged and are termed *neglectful* parents. They constitute 12% of the sample."

.

Table 1: Means for Clusters

	Warmth	Control	Cluster Size
Cluster 1	7.11 +/- 0.13 (0.94)	7.01 +/- 0.14 (1.01)	201 (26.8%)
Cluster 2	6.91 +/- 0.12 (1.11)	2.94 +/- 0.12 (1.08)	320 (42.7%)
Cluster 3	3.03 +/- 0.15 (0.92)	7.07 +/- 0.16 (0.96)	202 (19.2%)
Cluster 4	2.81 +/- 0.16 (0.75)	2.91 +/- 0.22 (1.03)	301 (11.5%)

(Table notes: Margins of error are half-widths of 95% confidence intervals. Standard deviations are in parentheses, except for cluster sizes, which are percents)

CENTROID CLUSTERING

We now analyze the same data but using a centroid clustering method. We illustrate a fuzzy cluster analysis based on partitioning, which is in the program "Fuzzy cluster analysis" in the folder "Cluster Analysis > Partitioning Methods." We assume you have read about it in the primer on cluster analysis. We use squared Euclidean distance metrics in this case, but other types of distance scores are reasonable as well.

Here are the results for the average silhouette widths using this algorithm (see the primer for a discussion of silhouette widths):

```
EVALUATION OF NUMBER OF CLUSTERS TO USE

  2 cluster model average silhouette width: 0.5884
  3 cluster model average silhouette width: 0.6582
  4 cluster model average silhouette width: 0.7131
  5 cluster model average silhouette width: 0.5943
  6 cluster model average silhouette width: 0.5201
  7 cluster model average silhouette width: 0.5277
```

```
8 cluster model average silhouette width: 0.5028
9 cluster model average silhouette width: 0.4978
10 cluster model average silhouette width: 0.5064
```

These values replicate the choice of a four cluster solution, which had the largest silhouette value. Other indices could be used as well, but we do not report them here in the interest of space.

The fuzzy cluster analysis program produces a plot of the four clusters, as did the hierarchical program (see Figure 5.3). The structure seems reasonably well articulated and is similar in form to the earlier plot from the hierarchical analysis:

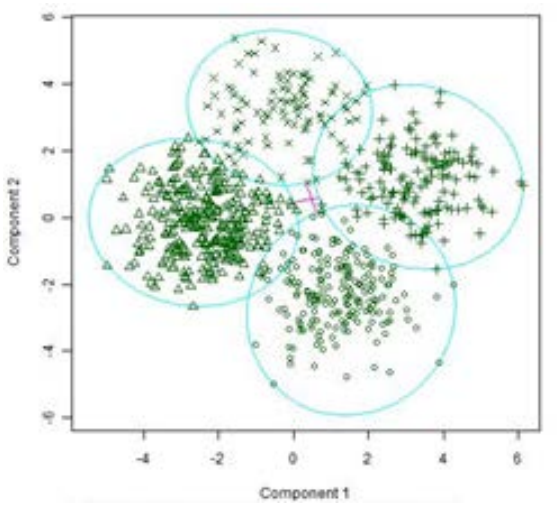


FIGURE 5.4. Cluster Plot for Fuzzy Cluster Analysis

Fuzzy clustering also produces a confusion matrix. Each individual is assigned a probability of being in each cluster, hence there are four such probabilities per person, one for each cluster. The individual is assigned to the cluster the individual has the highest probability of being in. For individuals classified into Cluster 1, we can calculate the mean probability they had of being in Cluster 1 as well as the mean probability they received for being in each of the other clusters. We expect the former probability to be large and the remaining probabilities to be small. We repeat this process for each cluster. The results for the analysis are as follows.

```
CONFUSION MATRIX

  FOR THOSE CLASSIFIED INTO CLUSTER 1

  Mean probability of being in cluster 1: 0.7923
```

```
Mean probability of being in cluster 2: 0.0774
Mean probability of being in cluster 3: 0.0862
Mean probability of being in cluster 4: 0.0404

FOR THOSE CLASSIFIED INTO CLUSTER 2

Mean probability of being in cluster 1: 0.0814
Mean probability of being in cluster 2: 0.7904
Mean probability of being in cluster 3: 0.0394
Mean probability of being in cluster 4: 0.0914

FOR THOSE CLASSIFIED INTO CLUSTER 3

Mean probability of being in cluster 1: 0.0847
Mean probability of being in cluster 2: 0.0388
Mean probability of being in cluster 3: 0.7905
Mean probability of being in cluster 4: 0.0853

FOR THOSE CLASSIFIED INTO CLUSTER 4

Mean probability of being in cluster 1: 0.0415
Mean probability of being in cluster 2: 0.0934
Mean probability of being in cluster 3: 0.0839
Mean probability of being in cluster 4: 0.7829
```

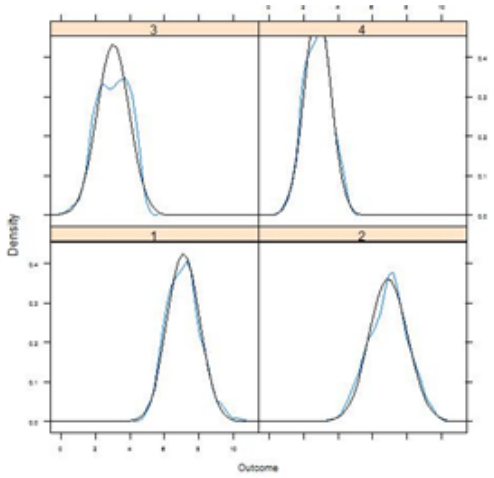
Overall, the confusion matrix is reasonably structured, which increases our confidence in the solution and our ability to unambiguously classify individuals.

Here is the table of means based on the fuzzy clustering solution (taken from the output, but re-ordered to map onto the prior table in the hierarchical analysis):

	Warmth	Control	Cluster Size
Cluster 1	7.05	7.00	202 (26.9%)
Cluster 2	7.09	2.83	298 (40.1%)
Cluster 3	3.10	7.03	150 (20.0%)
Cluster 4	3.09	2.89	100 (12.9%)

These means replicate well the prior solution, as do the sample sizes.

As a final check, we decide to examine the distribution of scores for the target variables within each cluster defined by the hierarchical analysis to determine if there are problematic distributions or blatant outliers. We saved cluster membership scores for each individual in our data set. We first examine kernel density plots for the individuals in each cluster for warmth using the program “Histograms and densities by groups” in the folder “Graphics (Excel and R) > R Histograms and Densities Graphs,” with normal distributions overlaid. Here are the side-by-side plots:



The distributions appear reasonable for our purposes. The same was true for the plots for the control variable.

Writing It Up

Here is how we might write-up the fuzzy cluster method:

"A centroid based cluster analysis was performed using the fuzzy cluster method described in Kaufman and Rousseeuw (1990). Inspection of the entropy values favored a four cluster solution. The values for a 2 cluster through 6 cluster model were 0.59, 0.66, 0.71, 0.59, 0.52, respectively. Table 1 presents the mean warmth and control values for each cluster as well as the sample sizes when individuals were assigned to a cluster based on membership probabilities. The four parenting styles implied by the clusters can be characterized as follows: First, there are parents who are both warm and controlling of their children, i.e., they are both affectionate and they "look out for" and monitor their child (Cluster 1). They are termed *authoritative* parents and constitute 27% of the sample. Second, there are parents who are warm and affectionate but who are low in control and supervision (Cluster 2). They are termed *permissive* parents and constitute 40% of the sample. Third, there are parents who seem to be fairly controlling of their adolescent child but who are not expressive of warmth (i.e. they are likely somewhat "cold" to their child; see Cluster 3). They are hereafter referred to as *authoritarian* parents and constitute 20% of the sample. Finally, there are parents who are relatively low in control and also low in warmth (Cluster 4). They seem disengaged and are termed *neglectful* parents. They constitute 13% of the sample. The confusion matrix for the analysis was reasonably well-articulated (see Table 2)."

Table 1: Means for Clusters

	Warmth	Control	Cluster Size
Cluster 1	7.05	7.00	202 (26.9%)
Cluster 2	7.09	2.83	298 (40.1%)
Cluster 3	3.10	7.03	150 (20.0%)
Cluster 4	3.09	2.89	100 (12.9%)

Table 2: Confusion Matrix

	Mean Probability			
Cluster Classified Into	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.79	0.08	0.09	0.04
Cluster 2	0.08	0.79	0.04	0.09
Cluster 3	0.08	0.04	0.79	0.09
Cluster 4	0.04	0.09	0.08	0.78

REFERENCES

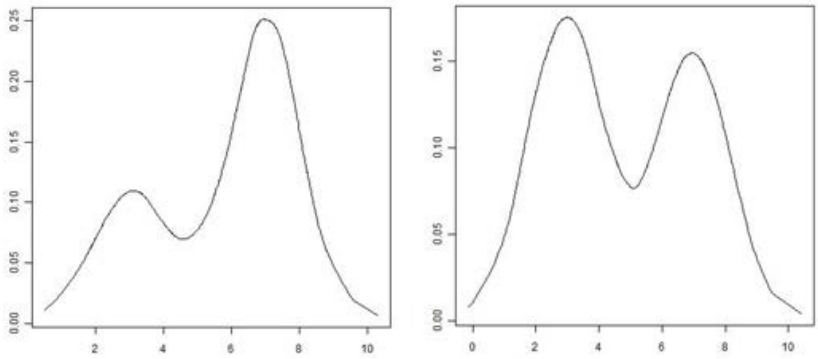
Curran, J. M. (2005). An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, Probability and Risk*, 4, 115-126.

Kaufman, L. & Rousseeuw, P.J. (1990) *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.

Pison, G., Struyf, A. and Rousseeuw, P. (1999). Displaying a clustering with CLUSPLOT. Computational Statistics and Data Analysis, 30, 381-392.

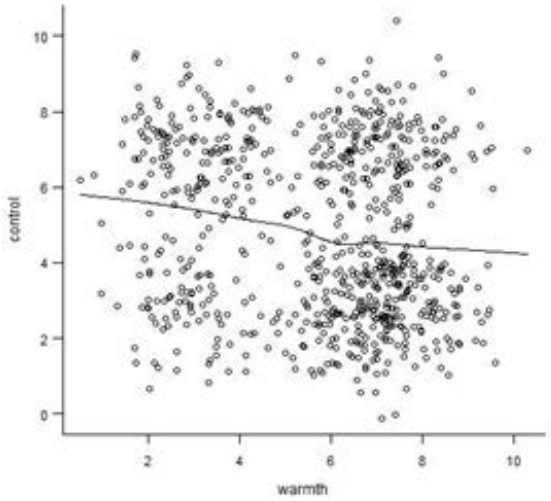
APPENDIX: PRELIMINARY ANALYSES

We examine kernel density plots for ***warmth*** and ***control*** using the program "Adaptive kernel density plot" in the folder "Describing Distributions (Frequencies, Means, SDs, Normality Tests > Frequency Distributions, Deciles, Distribution Shapes, Normality > Kernel Density Plots." Here are the plots (warmth on the left, control on the right):



Both distributions are decidedly bi-modal. This is not problematic. Indeed, such distributions can exist if there are distinct population clusters, each with roughly normal distributions on the target variables but with different means and/or variances. Such scenarios result in what are called *mixed normal distributions* and the above plots are consistent with this. In neither plot do outliers look to be particularly problematic.

Here is a smoother for *warmth* and *control* using the program "Traditional scatterplot with fit lines and smoothers" in the folder "Graphics (Excel and R) > "R Scatterplot and Regression Graphics":



The smoother line indicates little relationship between the variables and, again, there are no definitive signs of problematic outliers.