```
> head(h$merge)
     [,1] [,2]
[1,]  -22  -28
[2,]  -25  -57
[3,]  -21  -29
[4,]   -8  -12
[5,]   -9  -62
[6,]  -41  -60
```

We see that the logarithms of the 22nd and 28th observations have the smallest distance and therefore were the first cluster formed. These observations are:

```
> rownames(mammals)[c(22, 28)]
[1] "Horse"   "Giraffe"
```

and their logarithms are

```
> log(mammals)[c(22, 28), ]
            body    brain
Horse   6.255750 6.484635
Giraffe 6.270988 6.522093
```

Note that the cluster analysis will be different if the distances are computed on the original `mammals` data rather than the logarithms of the data, or if a different clustering algorithm is applied.

Exercises

2.1 (*chickwts* data). The `chickwts` data are collected from an experiment to compare the effectiveness of various feed supplements on the growth rate of chickens (see `?chickwts`). The variables are `weight` gained by the chicks, and type of `feed`, a factor. Display side-by-side boxplots of the weights for the six different types of feeds, and interpret.

2.2 (*iris* data). The `iris` data gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of three species of iris. There are four numeric variables corresponding to the sepal and petal measurements and one factor, `Species`. Display a table of means by `Species` (means should be computed separately for each of the three `Species`).

2.3 (*mtcars* data). Display the `mtcars` data included with R and read the documentation using `?mtcars`. Display parallel boxplots of the quantitative variables. Display a pairs plot of the quantitative variables. Does the pairs plot reveal any possible relations between the variables?

2.4 (*mammals* data). Refer to Example 2.7. Create a new variable r equal to the ratio of brain size over body size. Using the full `mammals` data set, order

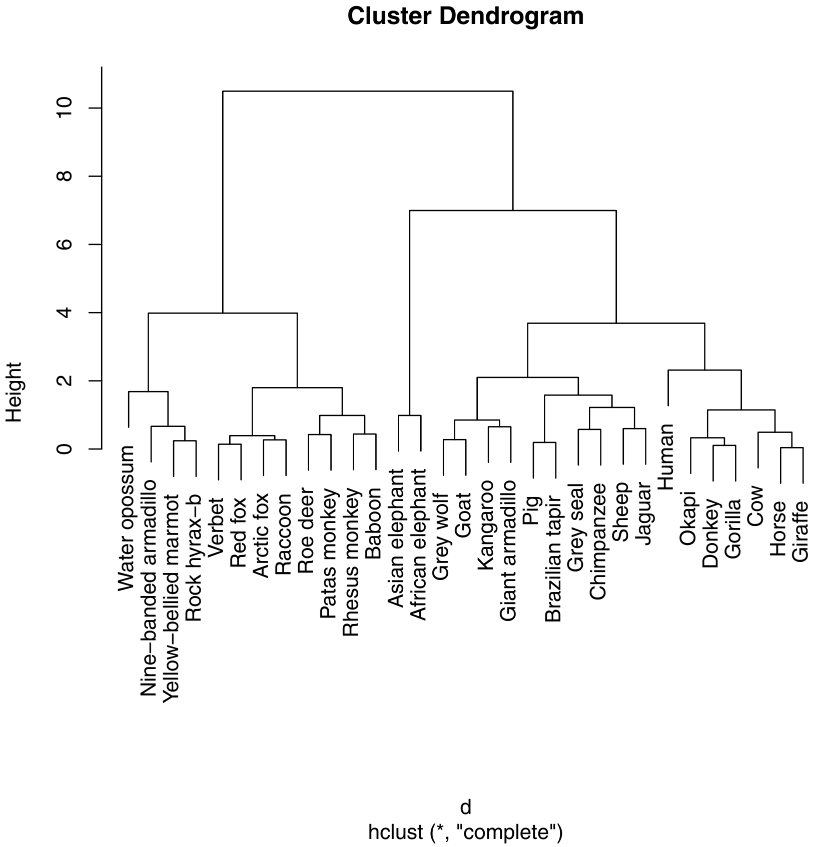

Fig. 2.20 Cluster dendrogram of log(mammals) data by nearest neighbor method in Example 2.1.

the `mammals` data by the ratio r. Which mammals have the largest ratios of brain size to body size? Which mammals have the smallest ratios? (Hint: use `head` and `tail` on the ordered data.)

2.5 (*mammals* data, continued). Refer to Exercise 2.5. Construct a scatterplot of the ratio $r = brain/body$ vs body size for the full `mammals` data set.

2.6 (*LakeHuron* data). The `LakeHuron` data contains annual measurements of the level, in feet, of Lake Huron from 1875 through 1972. Display a time plot of the data. Does the average level of the lake appear to be stable or changing with respect to time? Refer to Example 2.4 for a possible method of transforming this series so that the mean is stable, and plot the resulting series. Does the transformation help to stabilize the mean?

2.7 (Central Limit Theorem with simulated data). Refer to Example 2.6, where we computed sample means for each row of the `randu` data frame. Repeat the analysis, but instead of `randu`, create a matrix of random numbers using `runif`.

2.8 (Central Limit Theorem, continued). Refer to Example 2.6 and Exercise 2.7, where we computed sample means for each row of the data frame. Repeat the analysis in Exercise 2.7, but instead of sample size 3 generate a matrix that is 400 by 10 (sample size 10). Compare the histogram for sample size 3 and sample size 10. What does the Central Limit Theorem tell us about the distribution of the mean as sample size increases?

2.9 (1970 Vietnam draft lottery). What are some possible explanations for the apparent non-random patterns in the 1970 draft lottery numbers in Example 2.5? (See the references.)

2.10 ("Old Faithful" histogram). Use `hist` to display a *probability* histogram of the waiting times for the Old Faithful geyser in the *faithful* data set (see Example A.3). (Use the argument `prob=TRUE` or `freq=FALSE`.)

2.11 ("Old Faithful" density estimate). Use `hist` to display a *probability* histogram of the waiting times for the Old Faithful geyser in the *faithful* data set (see Example A.3) and add a `density` estimate using `lines`.

2.12 (Ordering the *mammals* data by brain size). Refer to Example 2.1. Using the full `mammals` data set, order the data by brain size. Which mammals have the largest brain sizes? Which mammals have the smallest brain sizes?

2.13 (*mammals* data on original scale). Refer to the `mammals` data in Example 2.7. Construct a scatterplot like Figure 2.19 on the original scale (Figure 2.19 is on the log-log scale.) Label the points and distances for cat, cow, and human. In this example, which plot is easier to interpret?

2.14 (*mammals* cluster analysis). Refer to Example 2.10. Repeat the cluster analysis using Ward's minimum variance method instead of nearest neighbor (complete) linkage. Ward's method is implemented in `hclust` with `method="ward"` when the first argument is the *squared* distance matrix. Display a dendrogram and compare the result with the dendrogram for the nearest neighbor method.

2.15 (Identifying groups or clusters). After cluster analysis, one is often interested in identifying groups or clusters in the data. In a hierarchical cluster analysis such as in Example 2.10, this corresponds to *cutting* the dendrogram (e.g. Figure 2.20) at a given level. The `cutree` function is an easy way to find the corresponding groups. For example, in Example 2.10, we saved the result of our complete-linkage clustering in an object `h`. To cut the tree to form five groups we use `cutree` with `k=5`:

```
g = cutree(h, 5)
```

Display `g` to see the labels of each observation. Summarize the group sizes using `table(g)`. There are three clusters that have only one mammal. Use `mammals[g > 2]` to identify which three mammals are singleton clusters.