Lab 7

Statistical Computing, 36-350

Friday October 16, 2015

Today's agenda: Transforming data; combining information from multiple objects; practice with selective access; practice applying functions.

General instructions for labs. Upload an R Markdown file, named "[your Andrew ID].Rmd", to Blackboard. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. Include the name of your lab partner at the top of the file.

R Markdown setup. Open a new R Markdown file; set the output to HTML mode and click "Knit HTML". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission. Alternatively, you can start from the lab's R Markdown file posted on the course website, as a template.

Background. Now-common ideas like "early adopters" and "viral marketing" grew from sociological studies of the diffusion of innovations.One of the most famous of these studies tracked how a then-new antibiotic, tetracycline, spread among doctors in four small cities in Illinois in the 1950s. In this lab, we will go back to that data to look at one of the crucial ideas, that of the innovation (prescribing tetracycline) "spreading" from person to person.

On the class website, you will find two data files, http://www.stat.cmu.edu/~ryantibs/statcomp/labs/ckm_nodes.csv and http://www.stat.cmu.edu/~ryantibs/statcomp/labs/ckm_network.dat. The former has information about each individual doctor in the four towns. The latter records which doctors knew each other.

Part I: Patterns of Tetracycline

1. Load the data in ckm_nodes.csv into a data frame, called `ckm_nodes`. Check that it has 246 rows and 13 columns. Check that there are columns named `city` and `adoption_date`.

2. `adoption_date` records the month in which the doctor began prescribing tetracycline, counting from November 1953. If the doctor did not begin prescribing it by month 17, i.e., February 1955, when the study ended, this is recorded as `Inf`. If it's not known when or if a doctor adopted tetracycline, their value is `NA`. How many doctors began prescribing tetracycline in each month of the study? How many never prescribed? How many are `NA`s?

3. Create a vector which records the index numbers of doctors for whom `adoption_date` is not `NA`. Check that this vector has length 125. Re-assign `ckm_nodes` so it only contains those rows. (Do not drop rows if they have a value for `adoption_date` but are `NA` in some other column.) Use this cleaned version of `ckm_nodes` for the rest of the lab.

4. Create plots of the number of doctors who *began* prescribing tetracycline each month versus time. (It is OK for the numbers on the x-axis to be numbers rather than formatted dates. Produce another plot of the *total* number of doctors prescribing tetracycline in each month. The curve for total adoptions should first rise rapidly and then level out around month 6.

5. Create a Boolean vector which indicates, for each doctor, whether they had begun prescribing tetracycline by month 2. Convert it to a vector of index numbers. There should be 20 such doctors.

6. Create a Boolean vector which indicates, for each doctor, whether they began prescribing tetracycline after month 14, or never prescribed it. Convert it to a vector of index numbers. There should be 23 such doctors.

Part II: Doctor, My Friend

7. The file `ckm_network.dat` contains a binary matrix; the entry in row i, column j is 1 if doctor number i said that doctor j was a friend or close professional contact, and 0 otherwise. Load the file into R as `ckm_network`, so that you have a square matrix which contains only 0s and 1s, of dimension 246 x 246. Drop the rows and columns corresponding to doctors with missing `adoption_date` values. Check that the result has 125 rows and columns. Use this reduced matrix, and its row and column numbers, for the rest of the lab.

8. Create a vector which stores the number of contacts each doctor has. Do not use a loop. Check that doctor number 41 had 3 contacts.

9. Create a Boolean vector which indicates, for each doctor, whether they were contacts of doctor number 37, *and* had begun prescribing tetracycline by month 5. Count the number of such doctors without converting the Boolean vector to a vector of indices. There should be 3 such doctors. What proportion of doctor 37's friends do these 3 doctors represent?

Bonus: Visualizing the Network

The `igraph` package is contains useful tools that will allow you to visualize the doctor network. Install the `igraph` package with `install.package()`, and then load it with `library()`. Using `igraph`, convert `ckm_network`, called an "adjacency matrix" in graph theory, into a graph object, and plot it.

Behind the Scenes

The original study was published as

> James Coleman, Elihu Katz and Herbert Menzel, "The Diffusion of an Innovation Among Physicians" *Sociometry* **20** (1957): 253–270.

The files used here are taken from http://moreno.ss.uci.edu/data.html#ckm with some formatting changes. CKM actually measured three types of link among the doctors — friendship, general discussion, and asking for medical advice. To keep things simple, we are combining all three types of tie, and treating them as symmetric.