Lab 8: How Antibiotics Came to Peoria

36-350

1 November 2013

Agenda: Fitting models by optimization; transforming data from one representation to another; handling missing data

Many theories of the diffusion of innovations (new technologies, practices, beliefs, etc.) suggest that the fraction of members of a group who have adopted the innovation by time t, $p(t)$, should follow a **logistic curve** or **logistic function**,

$$p(t) = \frac{e^{b(t-t_0)}}{1+e^{b(t-t_0)}} \tag{1}$$

Today and in the homework, we will look at a classic data set on the diffusion of innovations, which is supposed to show such a curve. It concerns a survey of 246 doctors in four towns in Illinois in the early 1950s, and when they began prescribing (adopted) a then-new antibiotic, tetracycline, and how they became convinced that they should do so (from medical journals, from colleagues, etc.).

Each row of `http://www.stat.cmu.edu/~cshalizi/statcomp/13/labs/08/ckm.csv` is a doctor. The column `adoption_date` shows how many months, after it became available, each doctor began prescribing tetracycline. Doctors who had not done so by the end of the survey, i.e., after month 17, have a value of `Inf` in this column. This information is not available (`NA`) for some doctors. There are twelve other variables, others of which may also be `NA`.[1]

1. (30) *The Model*

   (a) (10) Write a function, `logistic`, which calculates the logistic function (Eq. 1). It should take two arguments, `t` and `theta`. The `theta` argument should be a vector of length two, the first component being the parameter b and the second component being t_0. Your function may not use any loops. Plot the curve of the logistic function with $b = 0.05$, $t_0 = 3$ from $t = -30$ to $t = 30$.

   (b) (10) Explain why $p(t_0) = 0.5$, no matter what b is. Use this to check your `logistic` function at multiple combinations of b and t_0.

[1] For some of the other 12 variables, and the context, see `http://moreno.ss.uci.edu/data.html#ckm`, or Coleman, Katz and Menzel, *Medical Innovation: A Diffusion Study* (1966).

(c) (10) Explain why the slope of $p(t)$ at $t = t_0$ is $b/4$. (*Hint:* calculus.) Use this to check your `logistic` function at multiple combinations of b and t_0.

2. (40) *The Data*

   (a) (10) How many doctors in the survey had adopted tetracycline by month 5? *Hint:* `na.omit`, carefully.

   (b) (5) What *proportion* of doctors, for whom adoption dates are available, had adopted tetracycline by month 5?

   (c) (10) Create a vector, `prop_adopters`, storing the proportion of doctors who have adopted by each month. (Be careful about `Inf` and `NA`.)

   (d) (5) Make a scatter-plot of the proportion of adopters over time .

   (e) (10) Make *rough* guesses about t_0 and b from the plot, and from your answers in problem 1.

3. (30) *The Fit*

   (a) (10) Write a function, `logistic_mse`, which calculates the mean squared error of the logistic model on this data set. It should take a single vector, `theta`, and return a single number. This function cannot contain any loops, and must use your `logistic` function.

   (b) (10) Use `optim` to minimize `logistic_mse`, starting from your rough guess in problem 2e. Report the location and value of the optimum to *reasonable* precision. (By default, R prints to very unreasonable precision.)

   (c) (10) Add a curve of the fitted logistic function to your scatterplot from Problem 2d. Does it seem like a reasonable match?