Homework 7

Statistical Computing, 36-350

Due Wednesday Nov 4, 2015

Your homework must be submitted in R Markdown format. We will not (indeed, cannot) grade homeworks in other formats. Your responses must be supported by both textual explanations and the code you generate to produce your result. (Just examining your various objects in the "Environment" section of R Studio is insufficient—you must use scripted commands.)

We continue examining the diffusion of tetracycline among doctors in Illinois in the early 1950s, building on our work in Lab 7. You will need the data sets `ckm_nodes.csv` and `ckm_network.dat` from the lab.

1. Clean the data to eliminate doctors for whom we have no adoption-date information, as in the labs. Only use this cleaned data in the rest of the assignment.

2. Create a new data frame which records, for every doctor, for every month, whether that doctor began prescribing tetracycline that month, whether they had adopted tetracycline before that month, the number of their contacts who began prescribing strictly *before* that month. Explain why the dataframe should have 5 columns and 2125 rows. Try not to use any loops.

3. Let

   p_k = Pr(Doctor starts prescribing tetracycline this month |
   Doctor did not previously prescribe, and number of doctor's friends prescribing before this month is k)

   When computing p_k it is important to note that the number of doctors who *could* start prescribing this month does *not* include doctors have already started prescribing. In order to compute p_k correctly you must be sure that the denominator you use does not include doctors who have already started prescribing. If the denominator is zero you will find p_k=NaN. In these cases we can't estimate p_k, and you should exclude those values of k from your analysis

   We suppose the p_k are the same for all months.

   a. Explain why there should be no more than 20 values of k for which we can estimate p_k directly from the data.
   b. Explain how to estimate p_k from the data on hand.
   c. Create a vector of estimated p_k probabilities, using the data frame from (2). Plot the probabilities against the number of prior-adoptee contacts k; you should find an **increasing** trend, even by eye.

4. a. Suppose a model $p_k = a + bk$. This would mean that each friend who adopts the new drug increases the probability of adoption by an equal amount. Find the least squares regression solution of this model by minimizing the squared loss on the data using `grad.descent()` from Lecture 12, and the p_k values you constructed in (3b). Report the parameter estimates. (Hint: The squared loss on the data for a linear regression model in our case is $\sum_k (p_k - a - bk)^2$. Construct a function `squared.loss()` which calculates the squared loss on the training data for your model. Then use `grad.descent()` to minimize this loss, with initial value `c(0,0)`, step size of 0.001, stopping derivative of 0.001, and with a maximum of 10000 iterations. It is okay for this function to use variables from the global environment e.g. variables for p_k and k.) How many iterations did it take to converge? What were the final parameter estimates?

b. Suppose another model $p_k = e^{a+bk}/(1+e^{a+bk})$. Explain, in words, what this model would imply about the impact of adding one more adoptee friend on a given doctor's probability of adoption. (You can suppose that $b > 0$, if that makes it easier.) Estimate the model by minimizing the logistic loss, using the values you constructed in (3b). (Hint, the logistic loss in the data is $\sum_k \{\log(1+\exp(a+bk)) - p_k(a+bk)\}$. As before, create a function `logistic.loss()` which calculates the logistic loss in the data, and use `grad.descent()` from lecture notes, along with initial value `c(-2,0)`, step size and stopping derivative of 0.001, and with a maximum of 50000 iterations). How many iterations did it take to converge? What were the final parameter estimates?

c. Plot the values from (3b) along with the estimated curves from (4a) and (4b). (You should have one plot, with k on the horizontal axis, and probabilities on the vertical axis .) Which model do you prefer, and why?

d. **Bonus** (2pt) : Compare your answers in (a) and (b) to the results from `lm()` and `glm()` respectively (Hint: use `family='binomial'` in the latter to run a logistic regression). Do they match? In particular, glm() should have a visible discrepancy with your results in (b), which leads the next mission (should you choose to accept..)

Bonus bonus (2pt): The p_k values from problem 3 aren't all equally precise, because they come from different numbers of observations. Also, if each doctor with k adoptee contacts is independently deciding whether or not to adopt with probability p_k, then the variance in the number of adoptees will depend on p_k. Say that the actual proportion who decide to adopt is $\hat{p}_k$. A little probability (exercise!) shows that in this situation, $\mathbb{E}[\hat{p}_k] = p_k$, but that $\text{Var}[\hat{p}_k] = p_k(1-p_k)/n_k$, where n_k is the number of doctors in that situation. (We estimate probabilities more precisely when they're really extreme [close to 0 or 1], and/or we have lots of observations.) We can estimate that variance as $\hat{V}_k = \hat{p}_k(1-\hat{p}_k)/n_k$. Find the $\hat{V}_k$, and then re-do the estimation in (4a) where the squared error for p_k is divided by $\hat{V}_k$. Likewise in (4b), we are treating each p_k as if it came from the same number of observations, when in fact they each come from n_k observations, hence the discrepancy with `glm()`! To correct this, each term in the logistic loss should be multiplied by n_k. How much do the parameter estimates change? How much does the plotted curve in (4c) change?