Homework 8: Antibiotic Diffusion and Outlier Resistance

36-350, Fall 2013

Due at 11:59 pm on Thursday, 7 November 2013

We continue to work with the data on the diffusion of tetracycline, and the logistic curve model for product adoption, introduced in the lab. You can (and probably should) use code and results from the lab solutions.

Since real data is often contaminated with outliers, and rarely conforms to Gaussian distributions, it is common to replace the mean-squared-error criterion with other measures of error, which are less sensitive to the occasional large excursion from the model's predictions. One of these is the **Huber loss**, which we have seen in lecture:

$$\psi(h,c) = \begin{cases} h^2 & \text{if } |h| \leq c \\ 2|h|c - c^2 & \text{if } |h| > c \end{cases}$$

In this homework, we will fit the logistic-adoption model by minimizing the both the mean squared error and the mean Huber error.

For the purposes of this assignment, the vector y will stand for the proportion of doctors who have adopted tetracycline, with y_t the proportion who have adopted by month t. Similarly $p(t,\theta)$ will stand for the logistic function.

When a problem asks you to do something, give the command you used to do it. When it asks for numbers, report only to reasonable precision; an excessive number of decimal places will be penalized in the grading.

1. (30) *Fitting by MSE*

   (a) (10) Write a function, `logistic_mse`, which calculates the mean squared error (not the Huber error) of the logistic model on this data set. It should take a single vector of parameters, `theta`, and return a single number. This function cannot contain any loops, and must use your `logistic` function. It should *probably* also use the `prop_adopters` vector you calculated in lab.

   (b) (10) Use `optim` to minimize `logistic_mse`, starting from your rough guess in problem 2e in lab. 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 2b in lab. Does it seem like a reasonable match?

2. (5) Write a function, `huber`, to calculate the Huber loss. It should take two arguments, a vector `h` and a single number `c`, and return a vector containing $\psi(h_i, c)$ for each element of `h`. You may set the default value of c to be 0.02. How do you know that it works? For full credit, your function should contain no loops.

3. (10) Write a function, `mhe`, to calculate the mean Huber loss of the logistic-adoption model, applied to the data from lab. It should take one argument, a length-two vector called `theta`, giving the parameters of the logistic curve. How do you know that it works? For full credit, your function should contain no loops.

4. (10) Using one of the `surface` functions from lecture 9, make a contour plot of the mean Huber error as a function of b and t_0. Make a similar plot of the mean squared error. Comment.

5. (20) Use `optim` to find the parameter values which minimize the mean Huber error, starting from the parameter values that minimize the mean *squared* error, as found in problem 1b. Do they match what you expect from problem 4? Should they? How far off is the estimate from the minimizer of the MSE?

6. (5) Use `optim` to find the parameter values which minimize the mean Huber error, starting from the same rough guess as used in lab. Report the estimate and the value of the mean Huber error. How far off is it from the estimate you just got? How much does this make you worry about the sensitivity of `optim` to the starting guess?

7. (15) The survey was conducted on four different cities in Illinois. Fit the logistic-curve model to data for each city, using both mean-squared error and mean Huber error; report the parameter estimates. For full credit, use the split/apply/combine pattern, do not re-run things by hand. (Using `plyr` is optional.)

8. Extra Credit As explained in lecture 18, we can find the variance of the estimate by combining the Hessian of the objective function with the variance of the gradient:

$$\mathrm{Var}\left[\hat{\theta}\right] \approx \mathbf{H}^{-1}(\theta^*)\mathrm{Var}\left[\nabla f(\theta^*)\right]\mathbf{H}^{-1}(\theta^*)$$

where θ^* is the true optimum, f is the function we minimize to find the optimum, and $\mathbf{H}$ is its Hessian. Since we do not actually know θ^*, we cannot use this formula directly. We can use the Hessian at the estimate,

$$\mathbf{H}(\theta^*) \approx \mathbf{H}(\hat{\theta})$$

To get a value for $\mathrm{Var}\left[\nabla f(\theta^*)\right]$, we need to do something else. Define

$$f_t(\theta) = \psi(y_t - p(t,\theta))u_t(\theta) = \nabla f_t(\theta)$$

The vectors u_t are called the **scores**, and indicate how the parameters should be adjusted to best fit the various data points. Then it can be shown that

$$\mathrm{Var}\left[\nabla f(\theta^*)\right] \approx \frac{1}{n}\mathrm{Var}\left[u_t(\hat{\theta})\right]$$

with the variance on the right being found as the sample variance of the data points.

In this extra credit, you should work with the combined data from all cities, not the split-by-city data from problem 7.

(a) (5) Write a function, `huber.score`, to calculate the gradient of the Huber loss at a single data point. It should take as arguments a data frame or array `data`, consisting of a month t and the corresponding y_t, and a vector `theta`, giving the two parameters of the logistic function. It should return a vector of length two.

(b) (5) Create a 17×2 matrix, `scores`, giving the score vector for each month, at the estimate from problem 5. (Do this for the combined data for all cities, not separately for each city.) Do use your `huber.score` function, but do not use a loop. *Hint:* Make an array and use `apply`.

(c) (5) Find the 2×2 variance matrix of `scores`. *Hint:* `var`.

(d) (5) What is the **sandwich variance** of the estimate, $\mathbf{H}^{-1}\mathrm{Var}\left[\nabla f\right]\mathbf{H}^{-1}$?

(e) (5) What are the standard errors of $\hat{b}$ and $\hat{t}_0$? *Hint 1:* what is the relationship between the variance of an estimator and its standard error? *Hint 2:* the standard error of the mean is irrelevant here, and you will achieve nothing by trying to use it.