

Lab 7

We are going to look at some data that I have been analysing that try to determine which factors affect local recombination rates in the human genome. Each row of the data contains the following observations for 74 different genes:

- rho.avg = estimate of local recombination rate in the gene.
- decode = estimate of recombination rate in a large region surrounding the gene, obtained from data comparing parents and offspring. Note the units of this are different than for rho.avg.
- div = diversity (think of this as the number of mutations) in the gene in a random sample of individuals.
- gc = GC content of DNA sequence in the gene. (DNA is made up of As, Gs, Cs and Ts).
- rep = proportion of the gene that is made up of repetitive bits of DNA.
- chi = proportion of the gene that is made up of a particular sequence known as the “chi” sequence.
- csar4 = proportion of the gene that is made up of a particular sequence known as the “csar4” sequence.
- deltag = a estimate of the energy holding the DNA sequence together.

1. Read the data file into R, and attach it, using

```
r=read.table
('http://www.stat.washington.edu/stephens/S423/recomdata.aa.txt',
header=T,sep =',')
attach(h)
```

Note that the `sep = ‘,’` command is used because the data items in the file are comma-delimited.

Use simple and multiple linear regression to answer the following questions. You should provide support for your answers, and demonstrate that the models you use conform approximately to underlying assumptions of the regression models (or if you are unable to get them to conform, you should say what the problem is).

- (a) Is there evidence that genes with higher diversity, tend also to have higher recombination rates? (In answering this question use rho.avg, or some suitable transformation of it, as the response variable in a simple linear regression.)
- (b) Show that both GC content of the gene and the deltag value for the gene seem individually to be associated with recombination rate. What, if anything, changes when both are used in a multiple regression to predict rho.avg (or some suitable transformation)? If you see a change, explain it.
- (c) It has been suggested that genes with a high proportion of repetitive bits of DNA will tend to have a higher recombination rate. Do the data offer evidence to support this idea?
- (d) Suppose you are given the following facts (A and B): A) high GC content tends to cause higher mutation rates, and thus higher diversity; B) high diversity tends to cause higher recombination rates. Assuming both A and B are true then clearly high GC content would be expected to be associated with high recombination rate. Question: can all the effect of GC content on recombination rate be accounted for by its effect on diversity, or does GC content appear to have an effect *over and above* its effect on diversity?
- (e) The values rho.avg may be hard work to obtain for genes not listed here, while the other variables are easy to obtain. Therefore it would be interesting to be able to predict rho.avg from the other variables. Develop a multiple regression model to do this. How often does your model get within a factor of 2 of the actual value of rho.avg for the genes given here? Compare this with what you get if you use only the recombination rate in a large region surrounding the gene (the decode variable) to predict rho.avg.