Exercises 9 · Logit models; test of proportions; cherry-picking

To turn in (Due Monday, April 6, 2015)

(1) Spam filtering

In this problem, you will use logistic regression to construct a basic e-mail spam filter, such as the kind that keeps all sorts of annoying stuff out of your inbox. On the course website you will find two data sets: "spam-fit.csv" (3000 observations) and "spam-test.csv" (601 observations). Each row contains information about a single e-mail message, and each file has the same ten columns:

Columns 1–6 record the relative frequencies of particular words or strings: "remove," "order," "free," "meeting," "re:," and "edu." These are listed as percentages. For example, an entry of 1 under the "free" column means that 1% of the words in the e-mail were the word "free."

Columns 7 and 8 record the relative frequencies of two punctuation marks: the semicolon (;) and the exclamation mark (!). These are also listed as percentages, as above; an entry of 1 under the semicolon column means that 1% of characters in the e-mail were semicolons.

Column 9 is the average length of consecutive strings of capital letters. For example, an e-mail that contained only the phrase "BUY OUR PRODUCT!" would have an average run length of 13 capitals. (Punctuation marks and spaces do not count as interruptions.)

Column 10 indicates whether the e-mail was spam/commercial e-mail (1) or not (0).

(A) Using only the data in "spam-fit.csv", fit a logistic regression model to predict whether an e-mail is spam.

(B) Use your fitted regression equation to compute the probability that each e-mail in "spam-fit.csv" is a spam message. These are your in-sample fitted values. Suppose that your e-mail program filters all messages that the model judges to be spam with greater than 50% probability. There are two kinds of errors one could make using such a threshold: a false positive, where one wrongly declares a non-spam message to be spam; and a false negative, where one

wrongly allows a spam message to pass through the filter. Compute the following three error rates that characterize your model's performance on the "spam-fit" data:

(1) The false positive rate (FPR), or the fraction of non-spam messages (true nulls) that were flagged as spam.
(2) The false negative rate (FNR), or the fraction of spam messages that were not flagged as spam.
(3) The false discovery rate (FDR), or the fraction of false positives among all messages flagged as spam. That is, if K is the raw number of false positives, and M is the number of total positives (i.e. messages flagged as spam), the FDR is the ratio K/M.

(C) Now use the same fitted regression equation to compute the probability that each e-mail in "spam-test.csv" is a spam message. (Do not fit a new regression model to spam-test; rather, use the model you estimated from spam-fit to generate your predictions for spam-test.) Again, suppose your e-mail program filters all messages that the model judges to be spam with greater than 50% probability. Compute the same three error rates from Part B. The `predict.glm` function is useful.

(2) Test for a difference of two proportions

(A) Consider the general problem of testing whether two population proportions are different on the basis of observed sample proportions. That is: $x_1 \sim \text{Binomial}(n_1, w_1)$ and $x_2 \sim \text{Binomial}(n_2, w_2)$, and we want to know whether the difference $\Delta = w_1 - w_2 = 0$. Calculate (by hand) the mean and variance of the sampling distribution of the estimator $\hat{\Delta} = \hat{w}_1 - \hat{w}_2$, where $\hat{w}_i = x_i/n_i$ for $i = 1, 2$, assuming the null hypothesis that $\Delta = 0$. Show your work.

(B) Suppose we genotype $n_1 = 57$ people who possess some particular binary phenotypic trait (group 1). We also genotype $n_2 = 63$ people without the trait (group 2). The trait is complex, so many genes affect it. But we have a particular favorite gene that we believe may play a role in the trait. Thus for each sample, we calculate the number of people who are homozygous dominant (AA) at this locus in the genome. We find that there are $x_1 = 23$ people in group 1 who are AA, while there are $x_2 = 10$ people in group 2 who are AA. Choose an α level and use your result above to test whether we can reject the null hypothesis that the proportion of AA

genotypes is the same in the two underlying subpopulations. You may use the large-sample theory here: i.e. assuming that the central limit theorem has kicked in, and that the ratio $z = \hat{\Delta}/\mathrm{SE}(\hat{\Delta})$ has an asymptotic normal distribution under the null hypothesis.

(3) Cherry picking

We've studied Neyman–Pearson hypothesis testing in the idealized setting where we test a single pre-specified hypothesis about the difference between two proportions. What happens, however, when we use the same data both to generate *and* test a hypothesis?

Consider a modification to the above testing problem. Suppose that for each group, we compute the proportion of homozygous dominant alleles at 5 different loci (versus a single locus in the problem above). That is, of the n_1 people in group 1, x_{11} are homozygous dominant at allele 1, x_{12} are homozygous dominant at allele 2, and so forth. Similar, of the n_2 people in group 2, x_{21} are homozygous dominant at allele 1, x_{22} are homozygous dominant at allele 2, and so forth.

For each allele, we compute the z-score corresponding to a test for a different in proportions. Then we pick the single largest z score across the five alleles and check whether it falls in our rejection region corresponding to $\alpha = 0.05$. If it does, we claim we've found a significant genomic predictor of the phenotype.[1]

(A) Set up a Monte Carlo simulation to assess the probability that this procedure generates a false positive under the "global" null hypothesis that at each of the five alleles, the population proportion of homozygous dominant genotypes is no different in population 1 (has trait) than in population 2 (doesn't have trait). That is, the null hypothesis for each individual test of proportions is assumed true.

(B) Above you looked for associations of genotype with phenotype at $K = 5$ loci. Find (approximately) the smallest value of K for which the actual probability of a false positive exceeds 80% under the global null hypotheis. Comment on the following statement in light of your findings: using the same data set both to generate and to test hypotheses is dangerous.

(C) Read the article "Selective reporting biases in cancer prognostic factor studies," by Kyzas et al,[2] Come to class prepared to discuss the connections between the phenomenon described in this paper and the simulation study you ran above. Try Googling "file drawer problem" or "publication bias" if you want some further reading.

[1] You'll have to assume something here about the population-wide proportion of homozygous-dominant alleles at each locus (e.g. 25% regardless of whether one has the binary trait.) Just be clear about whatever assumptions you're making.

[2] Selective Reporting Biases in Cancer Prognostic Factor Studies. Panayiotis A. Kyzas, Konstantinos T. Loizou and John P. A. Ioannidis. J Natl Cancer Inst (20 July 2005) 97 (14): 1043-1055. doi: 10.1093/jnci/dji184. Accessible through the UT Library portal or simply through a web search.