AUSTRALIAN NATIONAL UNIVERSITY

Second Semester Take Home Examinations 2006

Math3346

Data Mining

The notes given below relate to questions 2 and 3. Questions are given in italics, followed by answers to the more searching questions and (in some cases) additional notes on points that arise.

Question 2 *(Each question carries up to 5 marks. Answers should occupy no more than a page in total.)*

(a): Gene-disease association studies: *A gene-disease association study collects data on patients, hospitalised for reasons unconnected with the disease, who had one of two gene variants A and B in their genetic makeup. The following data were obtained:*

	Variant A	*Variant B*
No disease	*11,880*	*1140*
With disease	*120*	*60*

What are the respective proportions with the disease, for variant A and for variant B? Before concluding that the disease is more strongly associated with one of the variants and that the variant is perhaps "causing" the disease, what checks should you make? Create an artificial data set that illustrates a major reason why such an inference may not be warranted.

(b): Errors in explanatory variables: *Discuss consequences for regression or classication studies when one or more explanatory variables is measured inaccurately, while others are measured with acceptable accuracy. Illustrate by reference to diet-disease association studies. Your answer should address the following:*
(i) Does failure to find an association necessarily imply no association?
(ii) Comment on implications, in a large dataset, for identifying explanatory variables that may be "important".

(a) The proportions are 0.01 and 0.05. Checks include:

- Individuals are not a random sample from the population. Biases that are different for different combinations of without/with disease and gene variant may affect the numbers who are hospitalized. This is difficult to check.
- We are looking at a subset of the total population – those who are hospitalized. Results may not generalize to the wider population. Note however that a genuine association in this subset of the population would still be of interest, and suggestive that an association in the population more generally is likely.
- Any dataset that has the following characteristics will meet the case: (i) for each level of a third factor that classifies the data, the proportions with and without the disease are different, but in each case the same for both variants; (ii) the relative numbers for the two levels of the third factor differ between variants.
  For example:

	Level 1 of factor C		Level 2 of factor C		Total	
	Variant A	Variant B	Variant A	Variant B	Variant A	Variant B
No disease	900	90	95	950	995	1040
With disease	100	10	5	50	105	60

  The overall proportions are 10.6% and 5.8%. Factor C might for example be the sex of the patient, or it might categorize variants of another gene.

(b) Assuming a single explanatory variable x and that measurement errors in x are independent of errors in the dependent variable, the magnitude of the regression estimate is reduced, and a greatly increased sample size may be required in order to find an effect that stands out relative to statistical variation.

Failure to find an association, as in several large diet/disease association studies, does not therefore imply lack of association. The food frequency questionaire, which has been the primary measurement instrument, is too inaccurate, with person-specific errors that have sometimes been of the same order of magnitude as between person variation in the dietary attribute under investigation.

In examining the effect of variables individually, an important variables that is measured with large error may show no (or a very small) effect, while variables that are measured accurately may, because of almost inevitable small biases in large data sets, show effects that stand out relative to statistical variation.

With multiple regression, there are additional complications that were not discussed in the lectures.

Question 3: *Data for this question can be loaded into an R session by typing:*

```
load(url("http://www.maths.anu.edu.au/˜johnm/r/misc-data/nsw.RData"))
```

Data are from two sources:[1]

- *Data from a work training experiment that was conducted over the period 1975–1977, under the aegis of the the United States National Supported Work (NSW) Demonstration program. This study randomly assigned individuals who had a history of employment and related difficulties to one of two "treatments": either to a 6–18 months training program (`trt`), or to a control group (`ExpCtl`) that did not participate in the training program.*
- *A group of non-experimental controls (`nonExpCtl`) that were taken from another study. These were chosen to be as similar as possible, on individual pre-intervention indicators, to the experimental subjects in the NSW study.*[2]

Variables are

```
trt (levels are ExptCtl, nonExptCtl, and trt)
age (years)
educ (years of education)
black (0=white 1=black)
hisp (0=non-hispanic 1=hispanic)
marr (0 = not married 2=married)
nodeg (0=completed high-school 1=dropout)
re75 (real earnings in 1975)
re78 (real earnings in 1978; this was the outcome variable of interest)
```

The original study was designed to test the effect of the work training on 1978 income (`re78`). Here, however, our interest is in exploring possible differences between the three sets of data. Especially, the interest is in systematic differences in the pre-intervention variables.

(Each of the following carries up to 3 marks per question.)

(a): Describe code steps: *Itemize the successive steps performed by the following code:*

```
sapply(split(nsw[,-1], nsw$trt), function(x)sapply(x,mean))
```

From this output, do you see any substantial differences between two or more of the three groups?

[1] The web site `http://www.nber.org/~rdehejia/nswdata.html` has additional data, and references. You may find it interesting to briefly peruse this web site. It is however unlikely that you will find anything, additional to the information given here, that will help with this exercise.

[2] This study had the name "Panel Study of Income Dynamics".

(b): Comparison between the three groups: *For variables that take a number of different values, suggest, and provide, a graphical alternative to Item that would allow a more refined comparison.*[3] *For which of the variables is the mean a satisfactory summary?*

(c): Annotation of code: *Run the following code. (Note that* `nsw[, -8]` *has been corrected to* `nsw[, -9]`*) Then add annotation that describes what each line does:*

```
library(randomForest)
nsw.rf <- randomForest(trt~., data=nsw[, -8, proximity=T)
cmdpoints <- cmdscale(1-nsw.rf$proximity)
library(lattice)
xyplot(cmdpoints[,2] ~ cmdpoints[,1], groups=nsw$trt,
       par.settings=list(pch=1:3), auto.key=list(columns=3))
```

(d): Interpretation of plot: *From the evidence from Items and , and from the plot in Item , comment on any differences between the three groups, as indicated by pre-intervention measures? Why may it be important to investigate such differences?*

(e): Is a low-dimensional representation adequate? *Repeat the above calculation of* `cmdpoints`*, now with* `k = 4`*. Calculate the sums of squares of the elements in the successive columns of* `cmdpoints`*. What does this tell you about the adequacy of the two-dimensional representation of the ''distances''?*

(a): Describe code steps:

(b): Comparison between the three groups: The mean is a satisfactory summary for binary 0/1 variables. There were evident large differences in the proportion who were black, who were married, and who did not graduate from high school. For other variables a boxplot or density plot is better, allowing a comparison of the distributions. For `age` and `educ` the distributions are very different between `nonExpCtl` and the other two groups, while for `re75` there are very clear differences. Differences in `re78` are to be expected. However the clear difference in the shape of the distribution, between `nonExpCtl` and the other two groups, is a concern, if the aim is to replace `ExpCtl` by `nonExpCtl` when values of `re78` are compared.

(c): Annotation of code: Here, `nsw[, -8]` should have been `nsw[, -9]`. Surprisingly, this did not (usually; the result is a bit different from one run of the `randomForest`

[3] Hint: Consider code of the form `boxplot(split(nsw$age, nsw$trt))`

calculation to the next) much affect the appearance of the plot. Differences between nonExpCtl and the other two groups in the pre-intervention variables were of more consequence for the two-dimensional separation of the groups than differences in re78.

Note: One student modified the code for xyplot(), setting col and pch directly rather than using par.settings If this is done the change must, additionally, be included in the call to auto.key. Otherwise the information in the key will be wrong. The par.settings mechanism ensures that the change is made in both places.

(d): Interpretation of plot: Differences in several of the pre-intervention variables were noted above. In the plot, points from nonExpCtl were clustered in the lower left corner, while points for the other two groups were (usually) pretty much interspersed. This indicates likely problems in any attempt, in any subsequent analysis, to replace ExpCtl by nonExpCtl

(e): Is a low-dimensional representation adequate? The sum of squares of elements in column i of cmdpoints is a measure of the separation of points in dimension i. Based on the use of nsw[, -8], I obtained:

```
## NB: Values in each column of cmdpoints are centered at 0
> apply(cmdpoints, 2, function(x)sum(x^2))
[1] 59.11386 28.21434 24.29253 19.68627
```

The third and fourth dimensions do contribute substantially to the separation between points. Based on the use of nsw[, -9], I obtained:

```
> apply(cmdpoints, 2, function(x)sum(x^2))
[1] 60.00933 29.90927 20.61945 20.18455
```

The first two dimensions do capture a slightly larger proportion of the total variation than when nsw[, -8] was used in the randomForest calculation.

Note: Papers that are noted on the web site canvass the issue of whether, using a suitable methodology, an analysis that replaces ExpCtl by nonExpCtl can yield the same result as when the comparison is with ExpCtl. The exam questions have the character of exploratory analysis of the data, as preliminaries to pursuing that question.