Assignment 2, Math 3346, 2009

Lecturer: John Maindonald

October 15, 2009

Due date: Sept 21 2009, at 5pm

General Comments

There were two issues that arise from the use of made of these data in the econometrics literature:

1. *Is there any reason to suspect the randomisation in the NSW experimental study?* The question was designed to elicit consideration about what might be expected if the randomisation was properly done, and to consider clues that might suggest that it was not properly done. If there is no reason to suspect the randomisation, then it ti reasonable to pool the control and treatment groups, for purpose of the comparison with the `cps1` and `psid1` control groups. (For these purposes, `re78` is of course ignored.)
2. *The `cps1` and `psid1` data, which was from observational studies individuals in neighbouring areas who had experienced similar employment difficulties, have been suggested for possible use as alternative controls. Might that have yielded reasonable results?* No, not using the data as they stand. The task set by the assignment had as a strong theme the assembling of the evidence, in a convincing manner. The questions were, essentially, hand-holding.

Questions

Function help page

Question 1: *Test the function `confusion()` by using the function `lda()` to do discriminant calculations, and then using this fucntion to determine the predictive accuracy.*

By typing

```
prompt(confusion)
```

you can get a skeleton for a help page for the function `confusion()`. Fill in the details for the help page. ***[3 marks]***

NB: help pages should be genuinely helpful!

Data transformation:

Question 2: *Why is it preferable to work with `logre75`, rather than `re75`?* ***[1 mark]***

In the untransformed data, a small number of individuals with large values of `re75` get large weight (or leverage).

Treated vs controls in nswdem, based on pre-treatment variables

Here is code that may be used to compare the control and treatment groups in the experimental nswdem *dataset:*

```
library(lattice)
form1 <- trt ~ age+educ+black+hisp+marr+nodeg+logre75
form2 <- trt ~ ns(age,3)+educ+black+hisp+marr+nodeg+logre75
  # form2 allows for the possibility that the effect of age mayh be nonlinear
form3 <- trt ~ (ns(age,3)+logre75)*(educ+black+hisp+marr)+nodeg
  # form3 allows for interaction effects involving continuous variables
## Try also the equivalent models with form2 and form3.
check.model(form1)
```

Question 3a: *With* form *taken to be whichever of* form1*,* form2 *and* form3 *gives the best accuracy, run the following code:*

```
nswdem1.lda <- lda(form, data=nswdem)
score <- predict(nswdem.lda)$x
plot(densityplot(~score, groups=nswdem$trt, auto.key=list(columns=2)))
```

Given that subjects had been randomly divided between treated and control (though with a greater number in control), is the graph much what might be expected? Explain.

[2 marks] Yes, it is. There are several ways that the answer could be justified:

- Bootstrap resampling;
- Simulation;
- Randomly permute labels between treatment and control groups.

No-one provided such a justification!

3b: *Compare the accuracies with the accuracy achieved by assigning all observations to the most frequent category. (In output from* rpart()*, this is the* root node accuracy*.) Does the best accuracy that was achieved give any reason to suspect the randomization? Explain.*

[2 marks]

Thoughts that this was intended to stimulate include:

- If the randomisation was effective, what should an effective dioscriminant analysis show?
- why the comparison the accuracy from asssignment to the most frequent group?

Pre-treatment comparison between nswdem, cps1 and psdi1

The following will use the smoothing spline methodology of the function gam() *in the mgcv package to check possible transformation of the variables* age*,* educ*, and* logre75*. The software automatically makes what should be a sensible choice of the amount of smoothing, corresponding to each of the terms* s(age)*,* s(educ)*, and* s(logre75)*.*

```
library(mgcv)
form <- gp1 ~ s(age)+s(educ)+s(logre75)+black+hisp+marr+nodeg
nswcps.gam <- gam(form, data=subset(nswplus, gp!="psid"))
nswpsid.gam <- gam(form, data=subset(nswplus, gp!="cps"))
```

Question 4: *Would the same transformations that are effective for the comparison between the experimental data and the `cps1` data be useful for the comparison between the experimental data and the `psid1` data. Justify your answer by reference to the graphs given by `plot(nswcps.gam)` and `plot(nswpsid.gam)`. You might want to do:*

```
opar <- par(mfrow=c(2,3), mar=c(3.6,3.6,0.6,0.6), mgp=c(2.25,.5,0))
plot(nswcps.gam)
plot(nswpsid.gam)
par(opar)
```

[2 marks]

The smoothing curves are broadly similar in their shape. So, at least as a starting point, use of the same transformation would probably be satisfactory.

Comparison between the three groups, using `lda()`

Question 5: *Use `lda()` to obtain a graphical comparison between the three groups. It may be desirable to account for nonlinear effects from any or all of `age`, `educ` and `logre75`. The graphs from Question 4 suggest that spline curves would be worth trying when we use `lda()`. (The function texttlda() does not have the functionality, implemented in `gam()`, for automatic choice of the degree of spline curve.) In what follows, we try normal splines of degree 4.*

```
library(splines)
fm1 <- gp ~ age+educ+black+hisp+marr+nodeg+logre75
fm2 <- gp ~ ns(age,4)+ns(educ,4)+ns(logre75,4)+black+hisp+marr+nodeg
```

Using whichever of `fm1` and `fm2` gives the best cross-validation predictive accuracy, fit the model, and graph the results. For example:

```
nswplus.lda <- lda(fm1, data=nswplus, prior=rep(1,3)/3)
scores <- predict(nswplus.lda)$x
library(lattice)
xyplot(scores[,1] ~ scores[,2], groups=nswplus$gp,
        auto.key=list(columns=3), par.settings=simpleTheme(cex=0.5, alpha=0.2))
```

The graph has a very striking feature. Can you shed any light on it? Hint: Consider looking at ranges of variables, within suitable splits of the data. ***[2 marks]***

The results (better accuracy, though not by a large margin) suggest use of `fm2`. Both graphs shows a separation into two groups, weaker in the `fm2` graph than in the `fm1` graph. The seaparation is between blacks and non-blacks. In part, this is however a difference in ages. Better modeling of the effects of age, `re75` and `educ` (especially) reduces the apparent effect of skin colour.

Comparison between the three groups, using `randomForest()`

Proximities from `randomForest` calculations can be used to get a plot that reflects the way that the data have been classified. Points are close together or widely separated according as they have, or have not, followed the same path down the trees to the terminal node. The proximity is, for each pair of points, the proportion of trees in which they have come together at the same terminal node. With more than several thousand points, the number gets so large that it creates problems for the computations.

```
## Extract subset of data for plotting
m <- cumsum(c(nrow(nswdem), nrow(cps1), nrow(psid1)))
n1 <- sample(1:m[1], 500)
n2 <- sample((m[1]+1):m[2], 500)
```

```
n3 <- sample((m[2]+1):m[3], 500)
take <- c(n1,n2,n3)
nswplus.rf <- randomForest(fm1, data=nswplus, sampsize=rep(722,3))
nswplus.pred <- predict(nswplus.rf, newdata=nswplus[take, ], proximity=TRUE)
sim <- 1-nswplus.pred$proximity
scores.rf <- cmdscale(sim)
xyplot(scores.rf[,1] ~ scores.rf[,2], groups=nswplus$gp[take],
        auto.key=list(columns=3), par.settings=simpleTheme(cex=0.5, alpha=0.5))
```

Question 6a: *Annotate the above code, explaining what each step does, and the choice of function arguments.* ***[3 marks]***

Question 6b: *Comment on the comparison between the `randomForest()` analysis and the `lda()` analysis.* ***[2 marks]***

In both plots, the three groups occupy different regions of the graph. The plot based on `lda()` is a low-dimensional projection, after rotation, of the space spanned by the columns of the model matrix. The plot from `randomForest()` reflects the way that the algorithm has classified the observations. The groupings do not reflect the split into black and non-black. That split is, in the `lda()` analysis, a proxy for other differences.

The overall accuracy for the random forest analysis was 79.9%, against an overall accuracy (leave one out CV) of 68.3% for the analysis that used `lda()`, with `fm2`. (If `fm1` is used, the accuracy drops to 62.1%.) Note that, for fair comparison with the `randomForest()` result, the prior must be specified as `c(1/3, 1/3, 1/3)`, both when calling `lda()` and when calling `confusion()`. Clearly `randomForest()` does a much more effective job of discrimination.

Overview

Question 7: *Write brief notes on what can be deduced from the several steps of the analysis. NB: Keep in mind the reasons for collecting and comparing these various sets of data.* ***[4 marks]***

See the "General Comments" at the beginning. There are huge differences between the three groups. As they stand, the `cps1` and `psid1` data are not suitable for use as controls.

[Some possibilities that have been investigated are:

- Identify subgroups that are similar.
  [Indications are that such subgroups would be very small; look at the plots]
- Propensity score methods try to identify a single (constructed) covariate that can account for differences in explanatory variables as a whole. Under quite strong assumptions, which are open to limited checking, the propensity score can be used to account for pre-treatment group differences.

]

TOTAL MARKS = 21