13

Regression on principal component or discriminant scores

Dimension reduction techniques reduce the number of candidate explanatory variables. Perhaps best known is the replacement of a large number of candidate explanatory variables by the first few principal components. The hope is that they will adequately summarize the information in the candidate explanatory variables. In favorable circumstances, simple modifications of the components will give new variables that are readily interpretable, but this is not always the case.

Propensity scores, often simply called propensities, may be helpful where a response is compared between two groups – a control and a treatment group – that have not been assigned randomly. The response may for example, in a medical context, be death rate in some interval of time. Variables that are not of direct interest, but which may in part explain any differences between the two groups, are commonly known as *explanatory variables*. Results from such analyses are likely to be suggestive rather than definitive, irrespective of the methodology used to account for explanatory variable effects.

Propensities aim to capture, in a single variable, the explanatory variable effects that are important in accounting for differences between two groups. The propensity score, commonly derived from a discriminant analysis, then becomes the only explanatory variable in the regression calculation.

Other types of ordination scores may be used in place of principal component scores. There are a variety of other possibilities.

13.1 Principal component scores in regression

The data set `socsupport` has the following columns:

```
1.  gender: male or female
2.  age: 18-20, 21-24, 25-30, 31-40, 40+
3.  country: Australia, other
4.  marital: married, single, other
5.  livewith: alone, friends, parents, partner, residences, other
6.  employment: full-time,  part-time, govt assistance, parental
    support, other
7.  firstyr: first year, other
8.  enrolment: full-time, part-time, blank
9   10. emotional, emotionalsat: availability of emotional support,
    and associated satisfaction (5 questions each)
```

```
11  12. tangible, tangiblesat: availability of tangible support
    and associated satisfaction (4 questions each)
13  14. affect, affectsat: availability of affectionate support
    sources and associated satisfaction (3 questions each)
15  16. psi: psisat: availability of positive social interaction
    and associated satisfaction (3 questions each)
17. esupport: extent of emotional support sources (4 questions)
18. psupport: extent of practical support sources (4 questions)
19. socsupport: extent of social support sources  (4 questions)
20. BDI: Score on the Beck depression index (total over
    21 questions)
```

The Beck depression index (`BDI`) is a standard psychological measure of depression (see for example Streiner and Norman, 2003). The data are from individuals who were generally normal and healthy. One interest was in studying how the support measures (columns 9–19 in the data frame) may affect `BDI`, and in what bearing the information in columns 1–8 may have. Pairwise correlations between the 11 measures range from 0.28 to 0.85. In the regression of `BDI` on all of the variables 9–19, nothing appears significant, though the F-statistic makes it clear that, overall, there is a statistically detectable effect. It is not possible to disentangle the effects of these various explanatory variables. Attempts to take account of variables 1–8 will only make matters worse. Variable selection has the difficulties that we noted in Chapter 6. In addition, any attempt to interpret individual regression coefficients focuses attention on specific variables, where a careful account will acknowledge that we observe their combined effect.

Hence the attraction of a methodology that, prior to any use of regression methods, has the potential to reduce the 11 variables to some smaller number of variables that together account for the major part of the variation. Here, principal components methodology will be used. A complication is that the number of questions whose scores were added varied, ranging from 3 to 5. This makes it more than usually desirable to base the principal components calculation on the correlation matrix.

Here now is a summary of the steps that have been followed, to obtain the results that will be described:

1. Following a principal components calculation on variables 9–19, we obtained scores for the first six principal components.
2. The six sets of scores were then used as six explanatory variables, in a regression analysis that had BDI as the response variable. The first run of this regression calculation identified an outlier, which was then omitted and the regression calculation repeated.
3. The regression output suggested that only the first of the variables used for the regression, i.e., only the principal component scores for the first principal component, contributed to the explanation of BDI. Compare $p = 0.00007$ for scores on the first component with p-values for later sets of scores, all of which have $p > 0.05$.
4. It was then of interest to examine the coefficients or loadings of the first principal component, to see which of the initial social support variables was involved.

Code to do the initial analysis, then presenting a scatterplot matrix of the scores on the first three principal components, is:

```
## Principal components: data frame socsupport (DAAG)
ss.pr1 <- princomp(as.matrix(na.omit(socsupport[, 9:19])), cor=TRUE)
pairs(ss.pr1$scores[, 1:3])
sort(-ss.pr1$scores[,1], decr=TRUE)[1:10]   # Note the outlier
## Alternative to pairs(), using the lattice function splom()
splom(~ss.pr1$scores[, 1:3])
```

The name given with the point that we have identified as an outlier is "36", which is the row name in the initial file. We omit this point and repeat the calculation.

```
not.na <- complete.cases(socsupport[, 9:19])
not.na[36] <- FALSE
ss.pr <- princomp(as.matrix(socsupport[not.na, 9:19]), cor=TRUE)
```

The output from `summary()` is:

```
> summary(ss.pr)     # Examine the contributions of the components
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Standard deviation      2.394  1.219  1.137 0.8448 0.7545  0.695
Proportion of Variance  0.521  0.135  0.117 0.0649 0.0517  0.044
Cumulative Proportion   0.521  0.656  0.773 0.8383 0.8901  0.934
                       Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
Standard deviation     0.4973 0.4561 0.3595 0.29555 0.23189
Proportion of Variance 0.0225 0.0189 0.0118 0.00794 0.00489
Cumulative Proportion  0.9565 0.9754 0.9872 0.99511 1.00000
```

We now regress `BDI` on the first six principal components. Because the successive columns of scores are uncorrelated, the coefficients are independent. Extraneous terms that contribute little except noise will have little effect on residual mean square, and hence to the standard errors. Thus, there is no reason to restrict the number of terms that we choose for initial examination. The coefficients in the regression output are:

```
> ss.lm <- lm(BDI[not.na] ~ ss.pr$scores[, 1:6], data=socsupport)
> summary(ss.lm)$coef
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 10.461      0.893  11.709 3.49e-19
ss.pr$scores[, 1:6]Comp.1    1.311      0.373   3.513 7.23e-04
ss.pr$scores[, 1:6]Comp.2   -0.396      0.733  -0.540 5.91e-01
ss.pr$scores[, 1:6]Comp.3    0.604      0.786   0.768 4.45e-01
ss.pr$scores[, 1:6]Comp.4    1.425      1.058   1.347 1.82e-01
ss.pr$scores[, 1:6]Comp.5    2.146      1.184   1.812 7.36e-02
ss.pr$scores[, 1:6]Comp.6    1.288      1.285   1.003 3.19e-01
```

Components other than the first do not make an evident contribution to prediction of BDI. We now examine the loadings for the first component:

```
> ss.pr$loadings[, 1]
   emotional emotionalsat     tangible  tangiblesat       affect
```

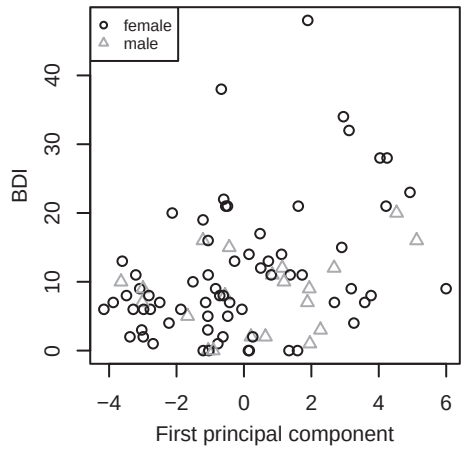

Figure 13.1 Plot of BDI against scores on the first principal component.

```
    -0.320       -0.298       -0.247       -0.289       -0.307
affectsat          psi       psisat     esupport     psupport
    -0.288       -0.363       -0.332       -0.289       -0.285
socsupport
    -0.285
```

The first component is like an average of the 11 measures.[1] A further step is then to plot `BDI` against the scores on the first principal component, using different colors and/or different symbols for females and males. This should be repeated for each of the other seven factors represented by columns 1–8 of the data frame `socsupport`. Figure 13.1 does this for the factor `gender`.[2]

Two observations seem anomalous, with BDI indices that are high given their scores on the first principal component. Both are females. We leave it as an exercise for the reader to recalculate the principal components with these points omitted, and repeat the regression.

Regression on principal component scores has made it possible to identify a clear effect from the social support variables. Because we have regressed on the principal components, it is not possible to ascribe these effects, with any confidence, to individual variables. The attempt to ascribe effects to individual social support variables, independently of other support variables, may anyway be misguided. It is unlikely to reflect the reality of the way that social support variables exercise their effects.

[1] The vector of loadings is unique up to multiplication by -1; the presence of negative signs here is due to the nature of the algorithm.

[2]
```
## Plot first principal componenet score against BDI
attach(socsupport)
plot(BDI[not.na] ~ ss.pr$scores[ ,1], col=as.numeric(gender[not.na]),
     pch=as.numeric(gender[not.na]), xlab ="1st principal component",
     ylab="BDI")
topleft <- par()$usr[c(1,4)]
legend(topleft[1], topleft[2], col=1:2, pch=1:2, legend=levels(gender))
detach(socsupport)
```

13.2* Propensity scores in regression comparisons – labor training data

A propensity is a measure, determined by explanatory variable values, of the probability that an observation will fall in the treatment rather than in the control group. Various forms of discriminant analysis may be used to determine scores. The propensity score is intended to account for between-group differences that are not due to the effect under investigation. If there is substantial overlap between propensity scores for the different groups, then comparison of observations within the approximate region of overlap may be reasonable, but using the propensity score to adjust for differences that remain. See Rosenbaum and Rubin (1983) for further comments on the methodology.

We will first describe the data, then investigate more conventional regression approaches to the analysis of these data, then investigate the use of propensity scores. The results highlight the difficulty in reaching secure conclusions from the use of observational data.

The labor training data

Data are from an experimental study, conducted under the aegis of the US National Supported Work (NSW) Demonstration Program, of individuals who had a history of employment and related difficulties. Over 1975–1977, an experiment randomly assigned individuals who met the eligibility criteria either to a treatment group that participated in a 6–18 months training program, or to a control group that did not participate.

The results for males, because they highlight methodological problems more sharply, have been studied more extensively than the corresponding results for females. Participation in the training gave an increase in male 1978 earnings, relative to those in the control group, by an average of $886 [SE $472].

Can the same results be obtained from data that matches the NSW training group with a non-experimental control group that received no such training? Lalonde (1986) and Dehejia and Wahba (1999) both investigated this question, using two different non-experimental control groups. These were:

1. The Panel Study of Income Dynamics (PSID: 2490 males, data in `psid1`, filtered data in `psid2` and `psid3`).
2. Westat's Matched Current Population Survey – Social Security Administration file (CPS: 16 289 males, data in `cps1`, filtered data in `cps2` and `cps3`).

Variables are:

```
trt (0 = control 1=treatment)
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); i.e. educ <= 11
re74 (real earnings in 1974; available for a subset of the
     experimental data only)
re75 (real earnings in 1975)
re78 (real earnings in 1978)
```

Table 13.1 *Proportion in the stated category, for each of the data sets indicated. Proportions for the experimental data are in the final two lines of the table.*

			Proportion			
	Black	Hispanic	Married	Dropout	re75 > 0	re78 > 0
psid1	0.25	0.03	0.87	0.31	0.90	0.89
psid2	0.39	0.07	0.74	0.49	0.66	0.66
psid3	0.45	0.12	0.70	0.51	0.39	0.49
cps1	0.07	0.07	0.71	0.30	0.89	0.86
cps2	0.11	0.08	0.46	0.45	0.82	0.83
cps3	0.20	0.14	0.51	0.60	0.69	0.77
nsw-ctl	0.80	0.11	0.16	0.81	0.58	0.70
nsw-trt	0.80	0.09	0.17	0.73	0.63	0.77

Observe that `trt`, `black`, `hisp`, `marr`, and `nodeg` are all binary variables. Here, they will be treated as dummy variables. In the language of Section 7.1, observations that have the value zero are the baseline, while the coefficient for observations that have the value 1 will give differences from this baseline. (For `marr`, where values are 0 or 2, the coefficient for observations that have the value 2 will be half the difference from the baseline.)

Note that `nodeg` is a categorical summary of the data in `educ`. It will not be used, additionally to `educ`, as an explanatory variable in the various analyses.

Summary information on the data

Table 13.1 has summary information on proportions on discrete categories that are of interest.[3] Information on `re74` is complete for the non-experimental sets of control data, but incomplete for the experimental data. We will examine the issue of how to handle `re74` below.

Notice the big differences, for `black`, `marr`, and `nodeg` (dropout), between the non-experimental controls (first six lines) and both sets of experimental data (final two lines). Even in the filtered data sets (`psid2`, `psid3`, `cps2`, and `cps3`), the differences are

[3]
```
showprop <-
  function(dframe=psid1, facCols=4:7, zeroCols=9:10){
    info <- numeric(length(facCols)+length(zeroCols))
    info[1:length(facCols)] <- sapply(dframe[,facCols], function(x){
      z <- table(x); z[2]/sum(z)})
    info[-(1:length(facCols))] <- sapply(dframe[,zeroCols], function(x)
      sum(x>0)/sum(!is.na(x)))
    info
  }
## Create matrix to hold resuilt
propmat <- matrix(0, ncol=6, nrow=8)
dimnames(propmat) <-
  list(c("psid1", "psid2", "psid3", "cps1", "cps2", "cps3",
         "nsw-ctl", "nsw-trt"), names(nswdemo)[c(4:7, 9:10)])
## Run function
for(k in 1:8){
  dframe <- switch(k, psid1, psid2, psid3, cps1, cps2, cps3,
                   subset(nswdemo, trt==0), subset(nswdemo, trt==1))
  propmat[k,] <- showprop(dframe)
}
```

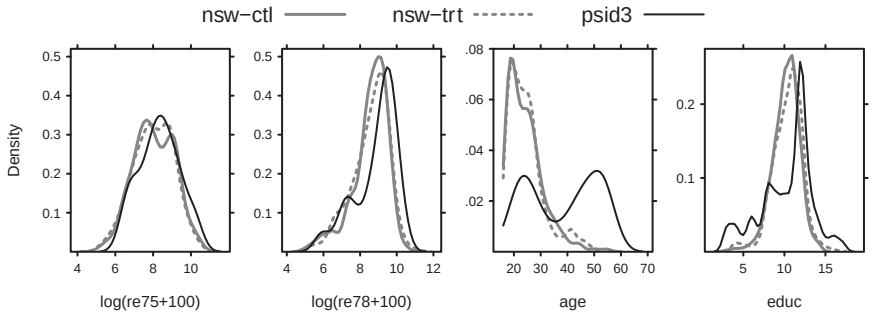

Figure 13.2 Overlaid density plots, comparing treatment groups with the experimental control data in `nswdemo` and with the non-experimental control data in `psid3`, for the variables `age` `educ`, `log(re75+30)`, and `log(re78+30)`.

substantial. The big changes that the filtering has made to the proportion with non-zero earnings is worrying. Notice particularly the huge differences between `psid3` and `psid1`, both for `re75` and `re78`.

For those who did earn an income, how do the distributions compare? The very heavy tails in the distributions of `re75` and `re78` make use of a logarithmic transformation desirable. Figure 13.2 compares the distributions of values, in the control and treatment groups, for the explanatory variables `age`, `educ`, `log(re75+30)`, and `log(re78+30)`. The offset of 30 is around half the minimum non-zero value for each of these variables, in the combined data. Plate 7 is an extended version of Figure 13.2 that has comparisons with all of the candidate sets of control data.

Examination of Figure 13.2, and of the additional comparisons in Plate 7, makes it clear that there are large differences between treatment and controls, whichever set of non-experimental controls is chosen.

The distributions of non-zero values of $\log(re78 + 30)$ are almost identical between experimental treated and control observations, just as similar as for $\log(re75 + 30)$. A more careful comparison will use QQ-plots. The comparison can be repeated with several bootstrap samples, as a check that such small differences as are apparent are not maintained under bootstrap sampling. This is pursued in the exercises at the end of the chapter. We will later check whether the differences that are apparent between non-experimental controls and treatment are maintained after a propensity score adjustment.

Plate 8 shows the scatterplot matrix, again for the data set that combines the `psid1` control data with the experimental treatment data, for the same variables as shown in Figure 13.2. Slightly simplified code is:

```
vnames <- c("trt","educ","age","re75","re78")
nsw <- rbind(psid1, subset(nswdemo, trt==1))
## Check minimum non-zero values of re75 and re78
round(sapply(nsw[,c("re75","re78")], function(x)unique(sort(x))[2]))
nsw[,c("re75","re78")] <- log(nsw[,c("re75","re78")] + 30)
lab <- c(vnames[2:3], paste("log\n", vnames[-(1:3)], "+", 30))
nsw$trt <- factor(nsw$trt, labels=c("Control (psid1)","Treatment"))
splom(~ nsw[,vnames[-1]], type=c("p","smooth"), groups=nsw$trt,
      varnames=lab, auto.key=list(columns=2))
```

13.2.1 Regression comparisons

One possibility is to use regression methods directly to compare the two groups, with variables other than `re78` used as explanatory variables. The nature of the data does however raise serious issues, for its use for this purpose.

Issues for the use of regression methods

The following points require consideration:

- Continuous variables almost certainly require some form of non-linear transformation. Regression splines may be a reasonable way to go.
- Should interaction terms be included?
- The large number of explanatory variables, and interactions if they are included, complicates the use of diagnostic checks.
- A substantial proportion of the values of `re78` are zero. The distribution of non-zero values of `re78` is highly skew, in both of the experimental groups (treatment and non-treatment), and in all of the non-experimental controls. A consequence is that the regression results will be strongly influenced by a small number of very large values. A $\log(re78 + 30)$ transformation (the choice of offset, in a range of perhaps 20–200, is not crucial) gives values that may more reasonably be used for regression, however. (In spite of the evident skewness, both Lalonde (1986) and Dehejia and Wahba (1999) used `re78` as the response variable in their analyses.)
- In the experimental data, almost 40% of values of the explanatory variable `re74` are missing. It is then necessary to ask whether these are "missing at random", or whether there is a pattern in the missingness. An indication that values of `re74` may not be missing at random is that its minimum value in the experimental data is 445 (dollars), which is close to 6 times the minimum of 74 for `re75` and almost 10 times the minimum of 45 for `re78`. Perhaps information on 1974 income was more readily available for participants who for most of 1974 held one steady job, or a small number of steady jobs. In the analysis below, a factor will be created from `re74` that has the levels: no income, some income, and income status unknown.
- Control and training groups can be made more comparable by some initial filtering of the data, on values of the explanatory variables. Inevitably, the choice of filtering mechanism and extent of filtering will be somewhat arbitrary, and filtering may introduce its own biases.
- Explanatory variables must both model within-group relationships acceptably well and model between-group differences acceptably well. These two demands can be in conflict.

Taken together, these points raise such serious issues that results from any use of regression methods have to be treated skeptically.

The complications of any use of regression analyses, and the uncertainties that remain after analysis, are in stark contrast to the relative simplicity of analysis for the experimental data. Experimental treatment and control distributions can be compared directly

and (assuming that the randomization was done properly) with confidence, without the complications that arise from the attempt to adjust for explanatory variable effects.

Regression calculations

As there may be information on whether or not `re74` is known, and on whether known values are non-zero, it seems useful to distinguish three categories – no income in 1974, some income in 1974, and details of income not known. Hence an argument for the use of a factor `fac74` that is derived thus:

```
nsw$fac74 <- with(nsw, factor(re74>0, exclude=NULL))
table(nsw$fac74)     # Check the order of the levels
levels(nsw$fac74) <- c("0","gt0","<NA>")
```

In the following analysis, two degrees of freedom have been allowed for a regression using a natural spline basis for each of $\log(re75 + 30)$, `age`, and `educ`. Here is a function that can be used for the calculations. The function has an argument that controls whether or not to apply a logarithmic transformation to `re78`.

```
nswlm <-
  function(control=psid1, df1=2, log78=TRUE, offset=30, printit=TRUE){
    nsw0 <- rbind(control, subset(nswdemo, trt==1))
    nsw0$fac74 <- factor(nsw0$re74>0, exclude=NULL)
    levels(nsw0$fac74) <- c("0","gt0","<NA>")
    if(log78) nsw.lm <- lm(log(re78+offset) ~ trt + ns(age,df1) +
                           ns(educ,df1) + black + hisp + fac74 +
                           ns(log(re75+offset),df1), data=nsw0) else
    nsw.lm <- lm(re78 ~ trt + ns(age,df1) + ns(educ,df1) + black +
                           hisp + fac74 + ns(log(re75+offset),df1),
                           data=nsw0)
    if(printit) print(summary(nsw.lm))
    trtvec <- unlist(summary(nsw.lm)$coef["trt", 1:2])
    trtEst <- c(trtvec[1], c(trtvec[1]+trtvec[2]*c(-1.96,1.96)))
    if(log78) {
      trtEst <- c(trtEst[1], exp(trtEst[1]), exp(trtEst[-1]))
      names(trtEst)=c("Est.","exp(Est.)","CIlower","CIupper")
    } else
    names(trtEst)=c("Est.","CIlower","CIupper")
    if(printit) print(trtEst)
    invisible(list(obj=nsw.lm, est=trtEst))
  }
## Try for example
library(splines)
nsw.lm1 <- nswlm(control=psid1)$nsw.lm
nswlm(control=subset(nswdemo, trt=0))
nswlm(control=psid1, log78=FALSE)
for (z in list(psid1,psid2,psid3,cps1,cps2,cps3))
  print(nswlm(control=z, printit=FALSE)$est)
```

Use of `termplot()` with the arguments `partial=TRUE` and `smooth=panel.smooth` suggests that the default numbers of degrees of freedom are adequate or more

than adequate. The coefficients of other terms in the equation are not highly sensitive to the number of degrees of freedom allowed.

The following table summarizes results, showing how they depend on the choice of control group:

Control used	Estimate of treatment effect	95% CI
`psid1`	$\exp(0.99) = 2.7$	(1.9, 3.7)
`psid2`	$\exp(0.61) = 1.8$	(1.0, 3.4)
`psid3`	$\exp(0.92) = 2.5$	(1.2, 5.2)
`cps1`	$\exp(0.85) = 2.3$	(1.7, 3.1)
`cps2`	$\exp(0.47) = 1.6$	(1.1, 2.4)
`cps3`	$\exp(0.49) = 1.6$	(0.96, 2.8)
`subset(nswdemo, trt=0)`	$\exp(0.35) = 1.4$	(1.1, 1.9)

These results vary widely, but do all point in the same direction as the experimental comparison in the final row. It is instructive to rerun the above calculations with `log78=FALSE`. The different results do not now all point in the same direction. The likely reason is that a few very large values of `re78` now have high leverage and a large influence. (Exercise 5 at the end of the chapter is designed to check this out.)

13.2.2 A strategy that uses propensity scores

A propensity "score" is a single variable whose values characterize the difference between the control and treatment groups. Importantly, the score is designed to model only between-group differences; it does not model within-group differences. Use of a single propensity score in place of many explanatory variables facilitates the use of standard checks to investigate whether the propensity score effect is plausibly linear. There is just one explanatory variable to investigate, rather than the complicated and often unfruitful task of carrying out checks on several explanatory variables.

For the analyses described here, we will start by using control observations from the data set `psid1`. Analyses that start by using control observations from one of the other data sets are left as an exercise for the reader.

Propensity scores will be derived from a discriminant analysis that uses the `randomForest()` function, from the package of the same name. Advantages of this approach are that prior transformation of variables is unnecessary, assumptions about the form of model are minimal, and there is automatic allowance for interactions. The extent of prior filtering of observations should not unduly matter.

We will however check out, for comparison, scores that arise from use of the function `lda()` (*MASS* package). It will turn out that `lda()` is similarly effective, as measured by predictive accuracy, in distinguishing the control and treatment groups. The key point is that it does no better than `randomForest()`, and thus that there is no reason to prefer the `lda` scores.

Either method yields, for each observation, an estimated probability p that the observation is from the treatment group. A convenient choice of propensity score is then $\log(p/(1-p))$.

The analysis will then replace the explanatory variables by a single propensity score. This is justifiable on theoretical grounds if the distribution of the explanatory variables is, conditional on the propensity score, the same for treatment and control observations. Checks can be performed to determine whether this assumption is plausible. If these checks fail, the analysis might still give reasonable results, but the theory does not give good grounds for confidence.

Derivation and investigation of scores

We now derive propensity scores. We convert `re74` to a factor with three levels – `0` (no income in 1974), `gt0` (income in 1974), and `<NA>` (income status in 1974 not known). The observations for which 1974 income information is available may be a biased selection, and it seems safest to use information on `re74` as a coarse indicator only.

```
## Use the dataset nsw that combines psid1 with exptl trt obsns
nsw.rf <- randomForest(trt ~ ., data=nsw[, -c(7:8,10)],
                       sampsize=c(297,297))
## NB: Use of equal bootstrap sample sizes (= 297 = number of
## treatment observations) gives the two groups equal prior weight.
```

We can check model accuracy

```
> nsw.rf
. . .
        OOB estimate of  error rate: 4.52%
Confusion matrix:
    0     1 class.error
0 2381  109      0.0438
1   17  280      0.0572
```

The random forest calculation should be rerun several times. We have found error rates that vary, over four runs, between 4.38% and 4.56%. These are the error rates that would be expected from a separate random sample from the same population.

The following fits a logistic regression model:

```
> library(MASS)
> library(splines)
> nsw.lda <- lda(trt ~ ns(age,2) + ns(educ,2) + black + hisp +
+                fac74 +  ns(log(re75+30),3),
+                CV=TRUE, prior=c(.5,.5), data=nsw)
> tab <- table(nsw.lda$class, nsw$trt)
> 1 - sum(tab[row(tab)==col(tab)])/sum(tab)
[1] 0.042
```

The `lda()` cross-validation error rate is very similar to that for `randomForest()`. The simple `lda()` model that does not allow for interaction effects may be adequate. The regression spline terms in the `lda` model seem to account for most of the non-linearity in the explanatory variables.

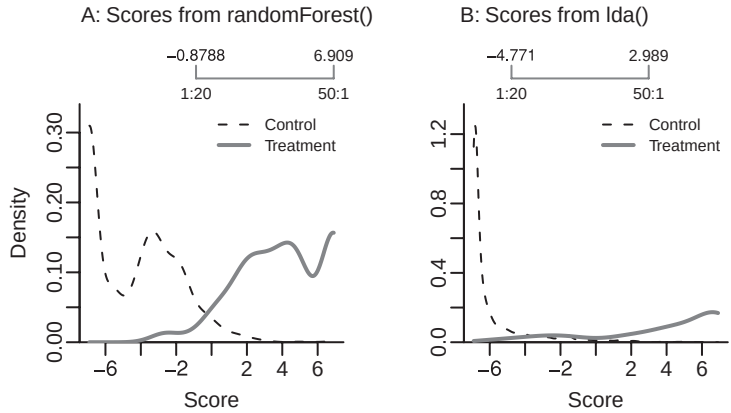

Figure 13.3 Panel A shows density plots of scores ($\log((p + 0.001)/(1 + 0.001 - p))$, where p is predicted value) from the object `sc.rf`, separately for control and treatment groups. Panel B is for scores, calculated similarly, from `sc.lda`. The ranges shown are ranges of relative numbers.

Here is code that calculates the scores and compares the densities between the control and treatment groups, as shown in Figure 13.3:

```
logit <- function(p, offset=0.001)log((p+offset)/(1+offset-p))
tnum <-  unclass(nsw$trt)
## NB: Derive scores by a logit transform of probabilities
sc.rf <- logit(predict(nsw.rf, type="prob")[,2])
overlapDensity(sc.rf[tnum==1], sc.rf[tnum==2], ratio=c(1/20, 50))
nsw.lda <- lda(trt ~ ns(age,2) + ns(educ,2) + black + hisp + fac74 +
               ns(log(re75+30),3), prior=c(.5,.5), data=nsw)
sc.lda <- logit(nsw.lda$posterior[,2])
overlapDensity(sc.lda[tnum==1], sc.lda[tnum==2], ratio=c(1/20, 50),
               compare.numbers=TRUE, plotval="Density")
```

The bulk of the control observations lie, in each instance, off to the left of the minimum score for which the ratio of treatment frequency to control frequency reached $\frac{1}{20} = 0.05$. For use of the `randomForest` scores, choosing observations with a score of more than -1.5 will retain approximately equal numbers (307/289, varying from run to run) of control and treatment scores. Without some such filtering, there may be undue leverage from the very large proportion of control observations that have large negative scores, where there are no treatment observations. Even modest filtering of observations with high scores (e.g., insist on a ratio of less than 50 treatment to one control observation) will filter out a large fraction of the treatment observations, and we keep such filtering to a minimum.

Now recalculate the propensity scores, at the same time calculating proximities between observations. The proximity between any pair of observations is the proportion of trees, out of the total number of trees (by default, 500), where the two observations appear together at the same terminal node.

```
nswa <- nsw[sc.rf>-1.5, ]
nswa.rf <- randomForest(trt ~ ., data=nswa[, -c(7:8,10)])
proba.rf <- predict(nswa.rf, type="prob")[,2]
sca.rf <- logit(proba.rf)
```

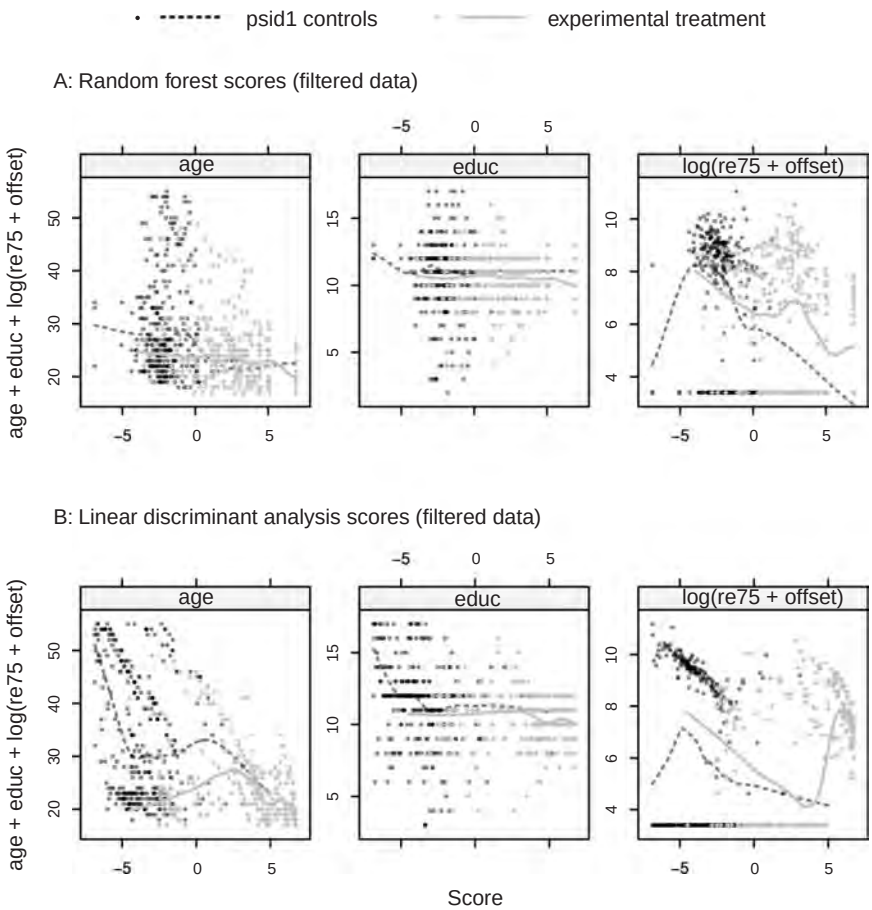

Figure 13.4 These plots are designed as a check whether, in each case, the distribution of the explanatory variable is, conditional on the score, similar for treated and controls. Panel A shows scores from `randomForest`, while panel B shows scores from `lda()`. Plate 9 is a color version that shows, also, the distribution of new `randomForest` scores obtained by refitting the model to data for which the scores shown in A were at least -1.5.

For `lda` scores, choosing observations with a score of more than -4 would retain somewhat more treatment than control scores (329/281).

Checks on the propensity scores

Is the distribution of the explanatory variables, conditional on the propensity score, the same for treatment and control? This can be checked for each individual explanatory variable. As interactions have seemed unimportant in determining the propensities, this may be enough. Figure 13.4 and and Plate 9 provide a visual check. Code that gives a close equivalent of Figure 13.4A is:

```
xyplot(age + educ + log(re75+30) ~ sca.rf, groups=trt, layout=c(3,1),
       data=nswa, type=c("p","smooth"), span=0.4, aspect=1,
```

```
par.settings=simpleTheme(lwd=c(2,1.5), col=c("gray", "black"),
  pch=c(20,3), cex=0.5, lty=c("solid","21")),
scales=list(y=list(relation="free"), tck=0.5),
auto.key=list(columns=2, points=TRUE, lines=TRUE,
  text=c("psid1 controls", "experimental treatment")),
xlab="Scores, derived using randomForest()")
```

For Figure 13.4B, replace `sca.rf` by `sca.lda`, obtained by linear discriminant calculations that use the subset of `nsw` for which `nsw.lda` is at least -4.

Conditional on the scores, both sets of panels show substantial differences for `age` and for `log(re75+30)`. The `randomForest` scores seem however preferable. In A (`randomForest()`), removal of points with very low scores (less than -1.5) has largely dealt with the most serious differences. In B (`lda()`) there is, for both of these variables, a large cluster of control points on the right of the plot. For `educ`, differences seem minor, for both sets of scores.

The graphs suggest that the formal requirements of the propensity score theory are in doubt. There are not good grounds for confidence that propensity scores will work well in making the necessary adjustment.

Use of proximities to give a two-dimensional representation

A more global graphical comparison is available by using the proximities from a `randomForest()` discriminant analysis as the basis for an ordination, i.e., for a two-dimensional representation as described in Subsection 12.1.3. Plots are shown for three ranges of scores – low, medium, and high. The text in the panel is labeled according to the equivalent range of probabilities.

Figure 13.5 shows the result. The code is:

```
## Repeat randomForest calculation, now with proximities
nswa.rf <- randomForest(trt ~ ., data=nswa[, -c(7:8,10)],
                        proximity=TRUE)
## Subtract proximities from 1.0, add 0.001, and use as "distances"
# NB: Use of isoMDS() will require all "distances" to be +ve
dmat <- 1-nswa.rf$proximity +0.001
proba.rf <- predict(nswa.rf, type="probability")[,2]
## From "distances", derive an ordination.
pts <- cmdscale(dmat)
ordScores <- isoMDS(dmat, pts)$points
cutpts <- c(0, round(quantile(proba.rf, c(1/3,2/3)), 2), 1).
cutp <- cut(proba.rf, breaks=cutpts, include.lowest=TRUE)
xyplot(ordScores[,2] ~ ordScores[,1]|cutp, groups=nswa$trt,
       xlab="Co-ordinate 1", ylab="Co-ordinate 2",
       auto.key=list(columns=2), aspect=1, layout=c(3,1),
       par.settings=simpleTheme(col=c("black","gray"), pch=c(1,3)))
```

In the sequel, the `randomForest` scores will be used. They do at least as well as the `lda` scores in accounting for differences between the two groups. Conditional on the propensity score, the distribution of the explanatory variables may be rather more similar

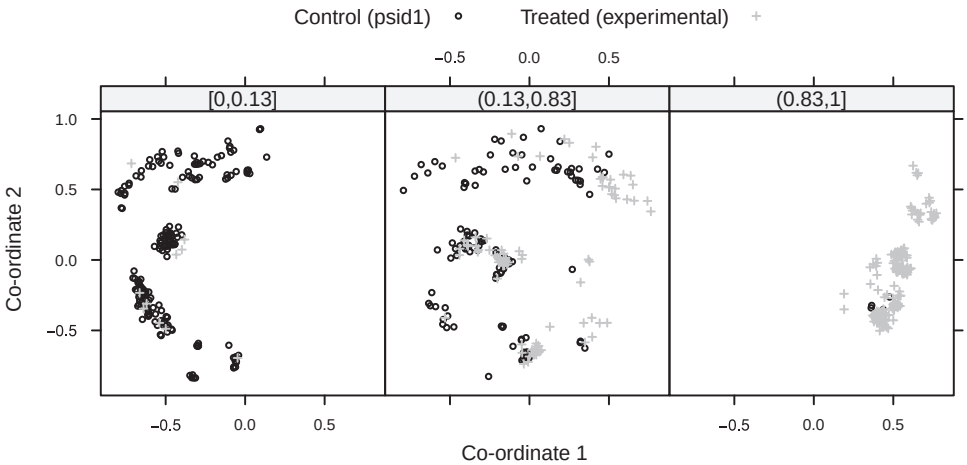

Figure 13.5 These plots are designed as a check whether, in each case, the distribution of the explanatory variables is, conditional on the score from `randomForest()`, similar for treated and controls. They examine a two-dimensional representation that is derived from the propensities. The ranges shown are for the probabilities before use of the logit transformation to give scores. Cutpoints have been chosen so that the three ranges contain an approximately equal number of observations. Note that results will differ somewhat from one run to the next.

between treatment and control than for the `lda` scores. They minimize opportunities for bias such as arise from the assumption, in the `lda` analysis, of a specific form of additive model.

Probability of non-zero earnings – analysis using the scores

The following checks whether there is a detectable training effect on the probability of non-zero earnings:

```
> sca.rf <- logit(sca.rf)
>  rf.glm <- glm(I(re78>0) ~ ns(sca.rf,2)+trt, data=nswa,
+                family=binomial)
>  summary(rf.glm)
. . . .
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.926  -1.363   0.705   0.831   1.313

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.158      0.555    0.28    0.776
ns(sca.rf, 2)1               1.001      1.313    0.76    0.446
ns(sca.rf, 2)2              -1.026      0.492   -2.09    0.037
trttreated (experimental)    0.740      0.305    2.43    0.015
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 691.39  on 595  degrees of freedom
Residual deviance: 676.63  on 592  degrees of freedom
AIC: 684.6

Number of Fisher Scoring iterations: 4
```

The estimate is in line with that from comparing experimental treatment data with experimental controls. Use of the linear discriminant scores yields a result that is even more clearcut.

Distribution of non-zero earnings – analysis using the scores

```
> rf.lm <- lm(log(re78+30) ~ ns(sca.rf,2)+trt, data=nswa,
+              subset = re78>0)
> summary(rf.lm)
. . . .
Residuals:
   Min     1Q Median     3Q    Max
-3.639 -0.441  0.153  0.660  2.695

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.699      0.329   29.47   <2e-16
ns(sca.rf, 2)1              -1.488      0.768   -1.94    0.053
ns(sca.rf, 2)2              -0.432      0.263   -1.64    0.101
trttreated (experimental)   -0.373      0.151   -2.47    0.014

Residual standard error: 0.987 on 433 degrees of freedom
Multiple R-squared: 0.0931,        Adjusted R-squared: 0.0868
F-statistic: 14.8 on 3 and 433 DF,  p-value: 3.37e-09
```

The negative (and statistically significant) treatment estimate contrasts with the result from the experimental data, where the estimated treatment effect is well below the threshold of statistical detectability.

```
> round(summary(lm(log(re78+30) ~ trt, data=nswdemo,
+       subset=re78>0))$coef, 4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   8.5601     0.0578 148.1486   0.0000
trt           0.0021     0.0874   0.0245   0.9804
```

In the absence of the check that the experimental data provides, it would be necessary to treat any of these results with extreme caution. Use of `psid2` or `psid3` (or `cps2` or `cps3`) is not an adequate answer. There are large elements of arbitrariness in the choice of observations to be removed, the filtering leaves data sets that still differ from the experimental treatment data in important respects, and results vary depending on which of these data sets is used as a control.

13.3 Further reading

Streiner and Norman (2003) discuss important issues that relate to the collection and analysis of multivariate data in medicine, in the health social sciences, and in psychology. On the use of propensity scores, see Rosenbaum and Rubin (1983), Rosenbaum (2002). On wider issues with respect to the analysis of observational data, see Rosenbaum (1999, 2002, 2005).

References for further reading

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Rosenbaum, P. R. 1999. Choice as an alternative to control in observational studies. *Statistical Science* 14: 259–78. With following discussion, pp. 279–304.

Rosenbaum, P. R. 2002. *Observational Studies*, 2nd edn.

Rosenbaum, P. R. 2005. Reasons for effects. *Chance* 18: 5–10.

Streiner, D. L. and Norman, G. R. 2003. *Health Measurement Scales. A Practical Guide to their Development and Use*, 3rd edn.

13.4 Exercises

1. Repeat the principal components calculation omitting the points that appear as outliers in Figure 13.1, and redo the regression calculation. What differences are apparent, in loadings for the first two principal components and/or in the regression results?

2. Examine the implications that the use of the logarithms of the income variables in the analysis of the data set `nswpsid1` has for the interpretation of the results. Determine predicted values for each observation. Then exp(predicted values) gives predicted incomes in 1978. Take exp(estimated treatment effect) to get an estimate of the factor by which a predicted income for the control group must, after adding the offset, be multiplied to get a predicted (income+offset) for the treatment group, if explanatory variable values are the same.

3. Investigate the sensitivity of the regression results in Subsection 13.2.2 to the range of values of the scores that are used in filtering the data. Try the effect of including data where: (a) the ratio of treatment to control numbers, as estimated from the density curve, is at least 1:40; (b) the ratio lies between 1:40 and 40; (c) the ratio is at least 1:10.

4. Modify the function `nswlm()` so that use of `fac74` as an explanatory factor is optional. With the `psid3` controls, is use of `fac74` as an explanatory factor justified? What is the effect on the confidence interval for the treatment effect?

5. Subsection 13.2.1 defined a function `nswlm()`, then using it to create a table that gives treatment effect estimates. Rerun the calculations that generated the table entries, now supplying the argument `log 78=FALSE`. Comment on changes in the treatment effect estimates.