Rosenbaum, P. R., 1999. Choice as an alternative to control in observational studies. *Statistical Science*, 14:259–278. With following discussion, pp.279–304.
Rosenbaum, P. R., 2002. *Observational Studies*. Springer-Verlag, 2 edition.
Streiner, D.L. and Norman, G.R. 1995. *Health Measurement Scales. A Practical Guide to Their Development and Use,* 2nd edn. Oxford University Press.

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 dataset `nsw74psid1` 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+offet) for the treatment group, if covariate values are the same.

3. Investigate the sensitivity of the regression results in Section 13.2.1 to the range of values of the scores that are used in flitering 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.

13.2* Propensity Scores in Regression Comparisons – Labor Training Data

A propensity is a measure, determined by covariate 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 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 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 by 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 study (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)

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.)

Summary information on the data

Table 13.1 has summary information on proportions on discrete categories that are of interest.[2] Information on re74 is complete for the non-experimental sets of control data, but incomplete for the experimental data. We will take us the issue of how to handle re74 below.

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

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).

[2]
```
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)
}
```



```
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+100) ~ ns(pred.rf,2)+trt, data=nsw,
+             subset = pred.rf>-1.5 & re78>0)
. . . .
> round(summary(rf.lm)$coef, 4) %$
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      12.2136     0.8768 13.9301   0.0000
ns(pred.rf, 2)1  -5.7043     1.6550 -3.4468   0.0006
ns(pred.rf, 2)2  -2.1673     0.5039 -4.3012   0.0000
trt              -0.3125     0.1166 -2.6792   0.0076
```

The negative (and statistically significant) treatment estimate contrasts with the result from the experimental data, where the estimated treatment effect is positive, but not statistically significant.

```
> round(summary(lm(log(re78+100) ~ 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 textttcps2 or cps3) is not an adequate answer. There are large elements of arbitrariness in the choice of observations to be removed, the filtering leaves datasets that still differ from the experimental treatment data in important respects, and results vary depending on which of these datasets 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 (2002, 1999).

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.

```
+       test="Chisq")
Analysis of Deviance Table

Model 1: marr ~ ns(pred.rf, 3) * trt
Model 2: marr ~ ns(pred.rf, 3) + trt
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      2779    1400.66
2      2782    1401.81   -3    -1.15      0.77
Warning message:
In glm.fit(x = X, y = Y, weights = weights, start = start,
  etastart = etastart,  :
  fitted probabilities numerically 0 or 1 occurred
```

The chief issue with the fitted probabilities that are numerically 0 or 1 is that the approximations of the asymptotic theory may be more than otherwise in doubt.

The scores from `lda()`

The conclusion is that the formal requirents of the propensity score theory are not satisfied. There are not good grounds for confidence that propensity scores will wotk well in making the necessary adjustment.

Probability of non-zero earnings – analysis using the scores

Here then is the analysis that checks for a training effect on the probability of non-zero earnings:

```
> rf.glm <- glm(I(re78>0) ~ ns(pred.rf,2)+trt, data=nsw,
+               subset=pred.rf>-1.5, family=binomial)
> summary(rf.glm)

Call:
glm(formula = I(re78 > 0) ~ ns(pred.rf, 2) + trt, family = binomial,
    data = nsw, subset = pred.rf > -1.5)

Deviance Residuals:
-1.9479   0.5932   0.6522   0.7062   1.0263

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.1547     2.0356   1.059    0.290
ns(pred.rf, 2)1  -1.7081     3.8243  -0.447    0.655
ns(pred.rf, 2)2  -1.5562     1.1859  -1.312    0.189
trt               0.4332     0.2778   1.559    0.119

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 746.14  on 711  degrees of freedom
Residual deviance: 737.81  on 708  degrees of freedom
AIC: 745.8
```


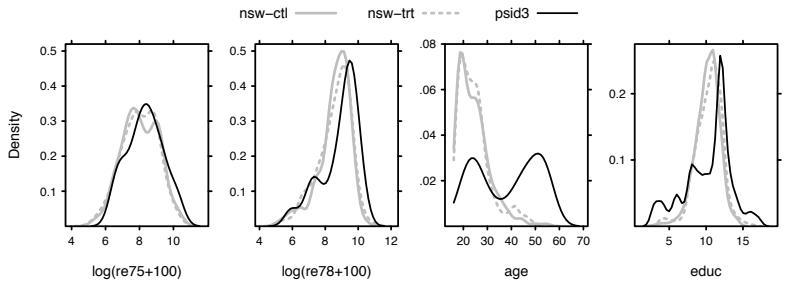

Figure 13.2: Overlaid density plots, comparing treatment groups with the experimental control data in `nswdemo` and with the non-experimental control data in `psid3`, for the variables `age` `educ`,`log(re75+100)`, and `log(re78+100)`.

Even in the filtered data sets (`psid2`, `psid3`, `cps2` and `cps3`), the differences are 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? Figure 13.2 compares the distributions of values, in the control and treatment groups, for the covariates `age`, `educ`, `re74` and `re75`. (Plate 9) 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 9, makes it clear that there are large differences between treatment and controls, whichever set of non-experimental controls is chosen. It does seem necessary to insist that the ratio of the density estimates should stay within some reasonable range, for each of these covariates.

The distributions of non-zero values of $\log(re78 + 100)$ are almost identical between experimental treated and control observations, just as similar as for $\log(re75 + 100)$. 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.

Issues to consider

One possibility is to use regression methods directly to compare the two groups, with variables other than `re78` used as explanatory covariates. Issues to consider are:

- Continuous variables then almost certainly require some form of non-linear transformation. Regression splines may be a reasonable way to go.
- Should interaction terms should 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-treatrment), 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 $\log(re78 + 100)$ transformation (the choice of offset, in a range of perhaps 50 – 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 dependent variable in their analyses.)

- The large number of explanatory variables, and interactions if they are included, complicates the use of diagnostic checks.
- 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 to an extent arbitrary, and filtering may introduce its own biases.
- Covariates 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 has to be treated sceptically.

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, without the complications that arise from the attempt to adjust for covariate effects.

Use of regression

The distributions for `re74`, `re75` and `re78` have heavy tails. In the analyses that now follow we will use a logarithmic scale for these income measures, again with an offset of 100.[3] Plate 10 shows the scatterplot matrix, for the experimental treatment data.[4]

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`. Thus, there may be some information in whether or not `re74` is known. Hence the use of the factor `fac74 = factor(is.na(re74))`.[5])

In the following analysis, 3 degrees of freedom have been allowed for regression spline functions for each of $\log(re74 + 100)$ and $\log(re75 + 100)$, and 2 degrees of freedom for

[3] One alternative is to take half the minimum non-zero value; this suggests a somewhat smaller offset).

[4]
```
linecols <- c(rgb(0,0.5,0.5), rgb(0.65,0,0.65))
trellis.par.set(theme=simpleTheme(cex=0.25, lwd=2, col.line=linecols))
vnames <- c("educ","age","re74","re75","re78")
dframe <- nsw[, vnames]
dframe[,-(1:2)] <- log(dframe[,-(1:2)] + 100)
lab <- c(vnames[1:2], paste("log\n", vnames[-(1:2)], "+", 100))
trt <- factor(nsw74psid1$trt, labels=c("Control","Treatment"))
splom(~ dframe, type=c("p","smooth"), groups=trt, varnames=lab,
      auto.key=list(columns=2))
```

[5]
```
## The following compares the distribution of log(re75+100), between
## cases for which re74 is known and cases for which it is not.
qqmath(~log(re75+22)|trt, groups=is.na(re74), data=nswdemo,
auto.key=list(columns=2))
```


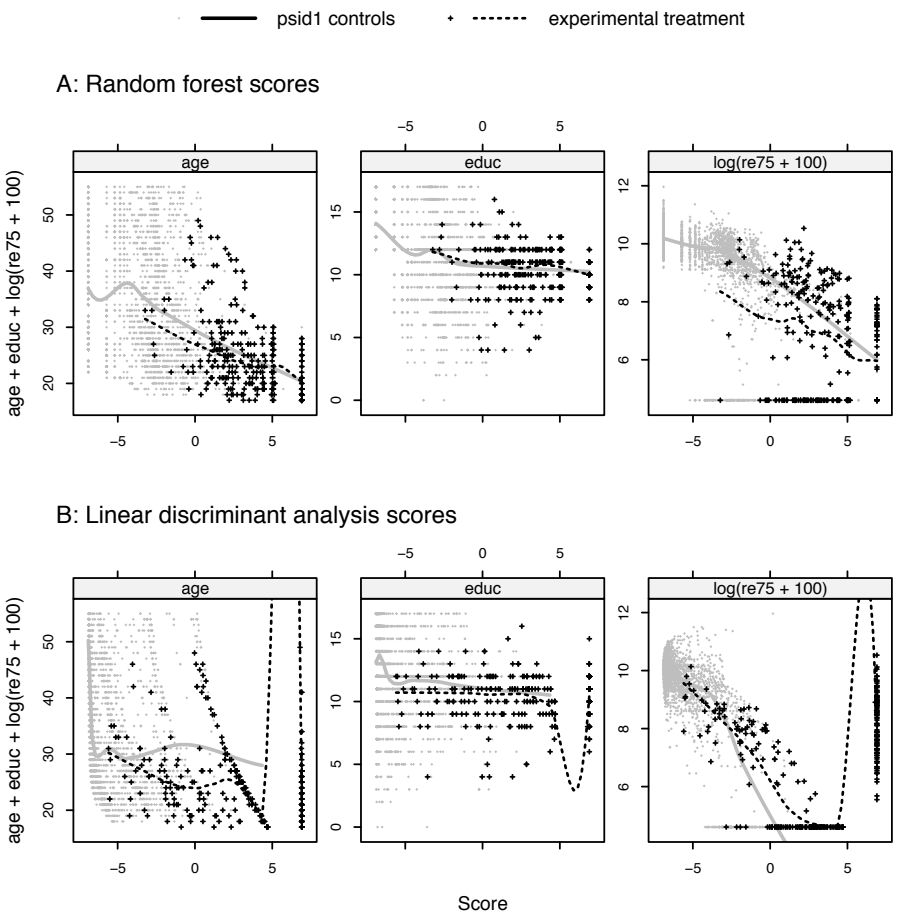

Figure 13.4: These plots are designed as a check whether, in each case, the distribution of the covariate is, conditional on the score, similar for treated and controls. Panel A shows scores from `randomForest`, while Panel B shows scores from `lda()`.

```
Model 1: log(re75 + 100) ~ poly(pred.rf, 2) * trt
Model 2: log(re75 + 100) ~ poly(pred.rf, 2) + trt
  Res.Df    RSS   Df Sum of Sq      F    Pr(>F)
1   2781 6085.2
2   2783 6109.5   -2     -24.3 5.5494 0.003933
```

Similar checks can be performed for the factors, e.g.:

```
> anova(glm(marr ~ ns(pred.rf,3)*trt, data=nsw, family=binomial),
+       glm(marr ~ ns(pred.rf,3)+trt, data=nsw, family=binomial),
```

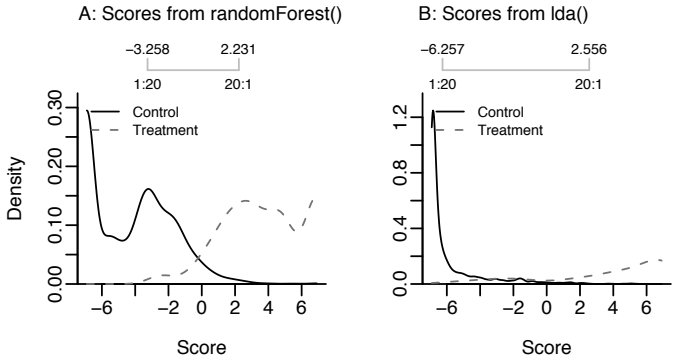

Figure 13.3: Panel A shows density plots of scores ($\log((p+0.001)/(1+0.001-p))$, where p is predicted value) from the object `pred.rf`, separately for control and treatment groups. Panel B is for scores, calculated similarly, from `pred.lda`.

the very large proportion of control observations that have scores between -6.9 and 0, where treatment observations are sparse.

Checks on the propensity scores

Is the distribution of the covariates is, conditional on the propensity score, the same for treatment and control? This can be checked for each individual covariate. Especially as interactions seems unimportant in determining the propensities, this may be enough. Figure 13.4 provides a visual check. Code that gives a close equivalent of Figure 13.4A is:

```
xyplot(age + educ + log(re75+100) ~ pred.rf, groups=trt, layout=c(3,1),
       data=nsw, 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 `pred.rf` by `pred.lda`.

The `randomForest` scores seem much preferable to `lda` scores for `age`. Differences between treatment and control are mostly for scores less than about 1, where treatment points are relatively sparse. Removal of points with very low scores will largely deal with such difference as there is. For `educ`, differences seem minor, for both sets of scores. For `log(re75+100)`, both sets of scores show substantial differences. Here is a formal check, using analysis of variance:

```
> anova(lm(log(re75+100) ~ poly(pred.rf,2)*trt, data=nsw),
+       lm(log(re75+100) ~ poly(pred.rf,2)+trt, data=nsw),test="F")
Analysis of Variance Table
```


each of `age` and `educ`. Here is a function that can be used for the calculations. It is written so that the logarithmic transformation for `re78` is optional.

```
'trylm' <-
  function(control=psid1, df1=2, df2=3, log78=TRUE, offset=100){
    nsw <- rbind(control, subset(nswdemo, trt==1))
    nsw$fac74 <- factor(is.na(nsw$re74), labels=c("has74","no74"))
    nsw$re74[is.na(nsw$re74)] <- 0
    if(log78) nsw.lm <- lm(log(re78+offset) ~ trt + ns(age,df1) +
                           ns(educ,df1) + black + hisp + fac74 +
                           ns(log(re74+offset),df2) +
                           ns(log(re75+offset),df2), data=nsw) else
    nsw.lm <- lm(re78 ~ trt + ns(age,df1) + ns(educ,df1) + black +
                           hisp + fac74 + ns(log(re74+offset),df2) +
                           ns(log(re75+offset),df2), data=nsw)
    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")
    print(trtEst)
    invisible(nsw.lm)
  }
## Try for example
nsw.lm1 <- trylm(control=psdi1)    ## Save object in nsw.lm1
trylm(control=subset(nswdemo, trt=0))
trylm(control=psdi1, log78=FALSE)  ## Regress re78 on covariates
```

Use of `termplot()` 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`	$\exp(0.75) = 2.1$	$(1.6, 2.8)$
`psid2`	$\exp(0.44) = 1.6$	$(0.96, 2.5)$
`psid3`	$\exp(0.76) = 2.1$	$(1.2, 3.8)$
`cps1`	$\exp(0.59) = 1.8$	$(1.4, 2.3)$
`cps2`	$\exp(0.41) = 1.5$	$(1.1, 2.1)$
`cps3`	$\exp(0.40) = 1.5$	$(0.98, 2.3)$
`subset(nswdemo, trt=0)`	$\exp(0.26) = 1.3$	$(1.0, 1.6)$

These results, although they vary widely, do at least point in the same direction as the experimental comparison in the final row.

It is instructive to re-run the above calculations with `log78=FALSE`. The different

results do not now all point in the same direction, presumably because a few very large values of re78 now have high leverage and a large influence. (An exercise at the end of the chapter is designed to check this out.)

13.2.1 A strategy that uses propensity scores

Propensity scores offer an alternative strategy. 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 covariates facilitates the use of standard checks to investigate whether the propensity score effect is plausibly linear. There is just one covariate to investigate, rather than the complicated and often unfruitful task of carrying out checks several covariates.

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 the data set cps1 are left as an exercise for the reader.

Propensity scores will be derived from a discriminant analysis. As a general highly automated approach for classification and consequently deriving propensity scores, use of randomForests() is attractive, providing each group has an adequate number of observations (e.g., at least 50, and preferably 100). Prior transformation of variables is unnecessary. There is automatic allowance for interactions. Prior filtering of observations is unnecessary. If however a simple-minded use of the function textttlda() (*MASS* package) gives a similar or smaller classification error, it may be preferred on the grounds of simplicity.

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 $\log((p+\epsilon)/(1+\epsilon-p))$, where ϵ is a small number (e.g., 0.001) that ensures that the argument supplied to the logarithmic function is always positive.

The analysis that then replaces the covariates by a single propensity score will be valid if the distribution of the covariates is, conditional on the propensity score, the same for treatment and control observations. Various 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.

```
nsw <- rbind(psid1, subset(nswdemo, trt==1))
nsw$fac74 <- factor(nsw$re74>0, exclude=NULL)
table(nsw$fac74)     # Check the order of the levels
```



```
levels(nsw$fac74) <- c("0","gt0","<NA>")
nsw.rf <- randomForest(as.factor(trt) ~ ., data=nsw[, -c(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: 4.27%
Confusion matrix:
     0   1 class.error
0 2394  96      0.0414
1   17 280      0.0539
```

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+100),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 random forest calculation should be re-run several times. We have found error rates that vary, over 4 runs, between 4.23% and 4.45%. These are the error rates that would be expected from a separate random sample from the same population. 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 covariates.

Here is code that calculates and plots the two sets of scores, as shown in Figure 13.3:3pt

```
logit <- function(p, offset=0.001)log((p+offset)/(1+offset-p))
trt <-  nsw$trt
pred.rf <- logit(predict(nsw.rf, type="prob")[,2])
overlapDensity(pred.rf[trt==0], pred.rf[trt==1], ratio=c(1/20, 20))
nsw.lda <- lda(trt ~ ns(age,2) + ns(educ,2) + black + hisp + fac74 +
               ns(log(re75+100),3), prior=c(.5,.5), data=nsw)
pred.lda <- logit(nsw.lda$posterior[,2])
overlapDensity(pred.lda[trt==0], pred.lda[trt==1], ratio=c(1/20, 20))
```

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 (285/289) of control and treatment scores. For lda scores, choosing observations with a score of more than -4 will retain approximately equal numbers (301/280). Without some such filtering, there may be undue leverage from