Problem Set 3: Matching and Weighting

Alejandro Sarria

Part 1

1.

a.

```
left_c <- 6/10
ambi_c <- 4/2
right_c <- 90/88

cat("Weight for control left handed:", left_c, "\n",
    "Weight for control ambidextrous:", ambi_c, "\n",
    "Weight for control right handed:", right_c)
```

```
Weight for control left handed: 0.6
 Weight for control ambidextrous: 2
 Weight for control right handed: 1.022727
```

b.

```
left_t <- 10/6
ambi_t <- 2/4
right_t <- 88/90

cat("Weight for treatment left handed:", left_t, "\n",
    "Weight for treatment ambidextrous:", ambi_t, "\n",
    "Weight for treatment right handed:", right_t)
```

```
Weight for treatment left handed: 1.666667
 Weight for treatment ambidextrous: 0.5
 Weight for treatment right handed: 0.9777778
```

c.

```
left_p <- (left_t * 6) / (left_t*6 + ambi_t*4 + right_t*90)
ambi_p <- (ambi_t * 4) / (left_t*6 + ambi_t*4 + right_t*90)
right_p <- (right_t * 90) / (left_t*6 + ambi_t*4 + right_t*90)

cat("Proportion for treatment left handed:", round(left_p*100), "%", "\n",
    "Proportion for treatment ambidextrous:", round(ambi_p*100), "%", "\n",
    "Proportion for treatment right handed:", round(right_p*100), "%")
```

```
Proportion for treatment left handed: 10 %
 Proportion for treatment ambidextrous: 2 %
 Proportion for treatment right handed: 88 %
```

d.

```
pensmanship <- (left_t*6*7 + ambi_t*4*4 + right_t*90*6) /
  (left_t*6 + ambi_t*4 + right_t*90)
cat("The weighted average pensmanship score in the treated group is", pensmanship)
```

```
The weighted average pensmanship score in the treated group is 6.06
```

2.

a.

Selecting multiple control matches for each treatment introduces several sub-optimal matches, this makes it so that observations in the control group that not as similar to treated observations enter the sample, making so that back door effected are not completely accounted for.

b.

Increasing the bandwidth allows for worse matches to be added to the sub-sample, this introduces more bias into the calculation.

c.

Selecting matches without replacement creates more bias as, once again, sub-optimal are introduced in the sub-sample, making it that the possible back door effects are not fully covered.

d.

Using only one exact match introduces more bias. This is because applying weights that decay with distance uses more information and creates a sample in which the differences in the relevant variables is reduced.

3.

Exact matching (or coarsened exact matching) quickly runs into the problem that as more matching variables are added, the sample splits into many small strata. This makes it much more likely that treated observations will not find suitable controls, leaving large portions of the data unmatched. Because of this, the approach only works well when the sample is very large, or when only a few matching variables are used, so that backdoor paths can still be blocked without discarding most of the data.

4.

d. propensity score requires a regression model to be specified, bringing over the limitations inherent in regression and specific to the chosen model.

5.

a.

The common support assumption on the group of treated business fails in the subgroup of retail businesses with 1 to 5 employees, as there are no such businesses in the control group, so there cannot be any comparison between and control for that strata.

b.

It is not a concern. We would be looking to create a sample with matches for cases in the treatment group, if there is a strata that has no cases in the treatment group, we don't need to look for matches anyways, so the lack of common support is not an issue in this particular case.

c.

The single untreated case in the strata of service businesses with 11 to 20 employees is going to have to carry a lot of weight in the matching schema. Either its propensity score is going to be significantly higher compared to others or it will have to be used multiple times on a matching with replacement scheme. Either way, it will introduce a significant amount of variance.

d.

Dropping the observations without common support would change the meaning of the analysis. Whatever the result of our analysis is, we could not say that it is applicable to retail businesses with less than 5 employees.

6.

When we select untreated observations to match the treated, we are estimating the effect of taking away treatment from those who actually received it. The matches serve as the counterfactuals for the treated group, so this gives us the average treatment effect on the treated (ATT).

When we select treated observations to match the untreated, we are estimating the effect of adding treatment to those who did not receive it. The matches serve as the counterfactuals for the untreated group, so this gives us the average treatment effect on the control (ATC).

Part 2

1.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
```

```
v purrr     1.1.0
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
here() starts at C:/Users/Alejandro/Documents/PhD Sociology Duke/2025-2/Stats/stats2-homework
```

2.

a.

```
d <- d |>
  mutate(weight = 1)
```

b.

```
model <- lm(re78 ~ treat,
            data = d,
            family = binomial,
            weights = weight)
```

```
Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
    ...) :
 extra argument 'family' will be disregarded
```

```
summary(model)
```

```

Call:
lm(formula = re78 ~ treat, data = d, weights = weight, family = binomial)

Residuals:
   Min     1Q Median     3Q    Max
 -6349  -4555  -1829   2917  53959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   4554.8      408.0  11.162  < 2e-16 ***
treat         1794.3      632.9   2.835  0.00479 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6580 on 443 degrees of freedom
Multiple R-squared:  0.01782,   Adjusted R-squared:  0.01561
F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

c.

```r
vars <- setdiff(names(d), c("data_id", "weight", "treat"))

balance_table <- data.frame(variable = character(),
                            treated = numeric(),
                            control = numeric(),
                            diff = numeric())

for (v in vars) {
  treated_mean <- weighted.mean(d[[v]][d$treat == 1], d$weight[d$treat == 1])
  control_mean <- weighted.mean(d[[v]][d$treat == 0], d$weight[d$treat == 0])
  diff <- abs(treated_mean - control_mean)

  row <- data.frame(variable = v,
                    treated = treated_mean,
                    control = control_mean,
                    diff = diff)

  balance_table <- rbind(balance_table, row)
}

balance_table[ , 2:4] <- lapply(balance_table[ , 2:4], round, 2)

balance_table
```

```
  variable treated control    diff
1      age   25.82   25.05    0.76
2     educ   10.35   10.09    0.26
3    black    0.84    0.83    0.02
4     hisp    0.06    0.11    0.05
```

```
5     marr    0.19    0.15    0.04
6 nodegree    0.71    0.83    0.13
7     re74 2095.57 2107.03   11.45
8     re75 1532.06 1266.91  265.15
9     re78 6349.14 4554.80 1794.34
```

d.

There is an issue in the difference between treated and control in the re74 and re75 variables. This are both elements that exist pre-treatment, meaning that with careful treatment assignation the difference between the two should be close to 0.

3.

a.

```
library(Matching)
```

```
Loading required package: MASS
```

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
##
##  Matching (Version 4.10-15, Build Date: 2024-10-14)
##  See https://www.jsekhon.com for additional documentation.
##  Please cite software as:
##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
##   Software with Automated Balance Optimization: The Matching package for R.''
##   Journal of Statistical Software, 42(7): 1-52.
##
```

```
Y <- d |> pull(re78)
Tr <- d |>  pull(treat)

X <- d |>
  dplyr::select("age", "educ", "black", "hisp", "marr",
                "nodegree", "re74", "re75") |>
  as.matrix()

M <- Match(Y, Tr, X, Weight = 2, M = 3)

matched_treated <- tibble(id = M$index.treated,
                          weight = M$weights)
matched_control <- tibble(id = M$index.control,
                          weight = M$weights)
matched_sets <- bind_rows(matched_treated, matched_control)

matched_sets <- matched_sets |>
  group_by(id) |>
  summarize(weight = sum(weight))

matched_d <- d[, -which(names(d) == "weight")] |>
  mutate(id = row_number()) |>
  left_join(matched_sets,
            by = 'id')
```

b.

```
library(cobalt)
```

```
 cobalt (Version 4.6.1, Build Date: 2025-08-20)
```

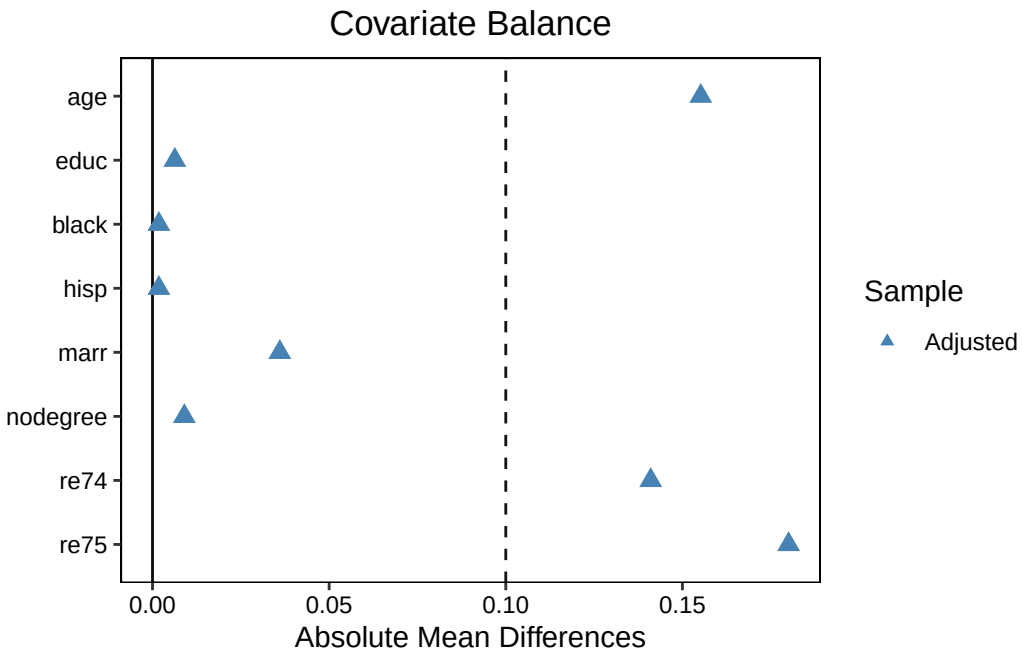
```
bal <- bal.tab(M, data = d,
        treat = "treat",
        covs = c("age", "educ", "black", "hisp", "marr",
                 "nodegree", "re74", "re75"))
bal
```

```
Balance Measures
            Type Diff.Adj
```

```
age      Contin.   0.1552
educ     Contin.  -0.0063
black     Binary   0.0018
hisp      Binary  -0.0018
marr      Binary   0.0360
nodegree  Binary  -0.0090
re74     Contin.   0.1410
re75     Contin.   0.1801

Sample sizes
                     Control Treated
All                   260.       185
Matched (ESS)         146.45     185
Matched (Unweighted)  222.       185
Unmatched              38.         0
```

```
love.plot(bal,
          binary = "std",
          drop.distance = TRUE,
          var.order = "unadjusted",
          colors = c("steelblue"),
          thresholds = .1,
          abs = TRUE)
```

```
Warning: Unadjusted values are missing. This can occur when `un = FALSE` and
`quick = TRUE` in the original call to `bal.tab()`.

Warning: `var.order` was set to "unadjusted", but no unadjusted mean differences
were calculated. Ignoring `var.order`.

Warning: Standardized mean differences and raw mean differences are present in
the same plot. Use the `stars` argument to distinguish between them and
appropriately label the x-axis. See `?love.plot` for details.
```



The balance of the matching results is mixed. education, black, hispanic, marriage status and degree are balanced (below the threshold) but age and both real earnings measure are still unbalanced in the sample.

c.

```
summary(M)
```

```
Estimate...  2027.8
AI SE......  785.12
T-stat.....  2.5827
p.val......  0.009802

Original number of observations..............  445
Original number of treated obs...............  185
Matched number of observations...............  185
Matched number of observations  (unweighted).  662
```

```
cat("Average effect of treatment on the treated:", M$est)
```

```
Average effect of treatment on the treated: 2027.764
```

4.

a.

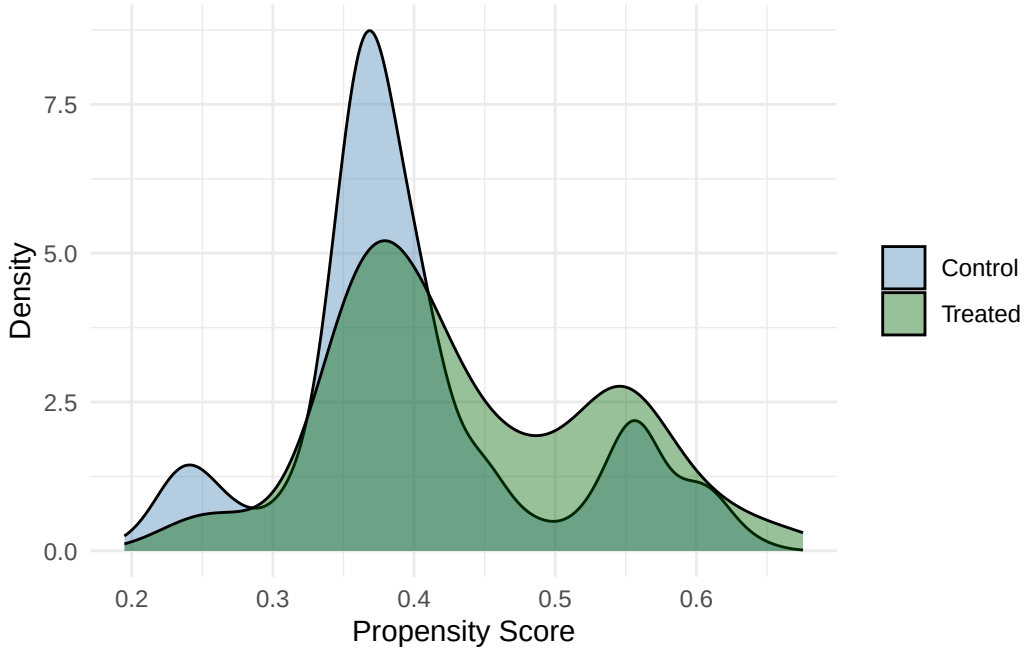
```
ps_model <- glm(treat ~ age + educ + black + hisp + marr +
                  nodegree + re74 + re75,
                data = d,
                family = binomial)

d <- d |>
  mutate(propensity = predict(ps_model, type = "response"))
```

b.

```
d <- d |>
  mutate(ipw = ifelse(treat == 1,
                      1,
                      propensity/(1-propensity)))
```

c.

```
ggplot(d,
       aes(x = propensity, fill = factor(treat))) +
  geom_density(alpha=0.4) +
  labs(x = "Propensity Score",
       y = "Density",
       fill = NULL) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "darkgreen"),
                    labels = c("Control", "Treated")) +
  theme_minimal()
```



The common support looks good, There is a bit of a distance around the 0.3 and 0.5 ranges but there is no area that indicates the need to cut a significant portion of the data.

d.

```
ipw_model <- lm(re78 ~ treat,
                data = d,
                weights = ipw)
cat("The treatment effect is estimated to be:", round(ipw_model$coefficients[2], 2))
```

```
The treatment effect is estimated to be: 1806.42
```