Problem Set 4. Statistical modeling.

Relevant material will be covered by **Oct 5**. Problem set is due **Oct 19**.

To complete the problem set, **Download the .Rmd** and complete the homework. Omit your name so we can have anonymous peer feedback. Compile to a PDF and submit the PDF on Canvas.

The learning goals of completing this problem set are

- explain the role of statistical modeling
  - with respect to causal claims
  - with respect to data sparsity
- estimate average treatment effects by
  - exact matching (in a setting with few confounders)
  - learning an outcome model
  - learning a treatment model
  - a matching method of your choosing

The reason for practicing many statistical modeling estimators is so you can see how the ideas of this class apply to all those estimators—and to future estimators you will encounter that are not part of this class!

This problem set uses data from the following paper:

> Dehejia, R. H. and Wahba, S. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94(448):1053–1062.

The paper compares methods for observational causal inference to recover an average causal effect that was already known from a randomized experiment. You do not need to read the paper; we will just use the study's data as an illustration.

The following lines will load these data into R.

```
library(tidyverse)
```

```
FALSE Warning: package 'tidyverse' was built under R version 4.2.3

FALSE Warning: package 'ggplot2' was built under R version 4.2.3

FALSE Warning: package 'tibble' was built under R version 4.2.3

FALSE Warning: package 'purrr' was built under R version 4.2.3

FALSE Warning: package 'dplyr' was built under R version 4.2.3

FALSE Warning: package 'stringr' was built under R version 4.2.3
```

```
library(MatchIt)
```

```
FALSE Warning: package 'MatchIt' was built under R version 4.2.3
```

```r
data("lalonde")
```

To learn about the data, type `?lalonde` in your R console.

1. Conceptual questions

1.1. (5 points) Statistical modeling and causal claims

Imagine that someone who has not taken our class tells you they don't need DAGs or causal assumptions because they know a really good matching method. In no more than 3 sentences, explain to them why causal assumptions are necessary for matching to yield causal conclusions.

Answer.

Because they support the reliability of the matching process, causal assumptions are necessary for matching to produce causal findings. Without making causal assumptions, matching may not produce truly comparable groups, and any differences that are found cannot be securely attributed to the treatment or intervention under study. By ensuring that any confounding variables are effectively controlled for during the matching process, causal assumptions help to produce conclusions about the cause of events that are more trustworthy.

2. Nonparametric estimation

Our goal is to estimate the effect of job training `treat` on future earnings `re78` (real earnings in 1978), among those who received job training (the average treatment effect on the treated, ATT).

2.1. (4 points) Exact matching with low-dimensional confounding

For this part, assume that three variables comprise a sufficient adjustment set: `race`, `married`, and `nodegree`. Use `matchit` with the argument `method = "exact"` to conduct exact matching, which matches two units only if they are identical along `race`, `married`, and `nodegree`.

> Note: Here we are calling this **exact matching**. This is the same thing we previously called **nonparametric estimation**: make subgroups of units identical along confounders, estimate the treatment effect within those subgroups, and aggregate over the sample. We are using the language of matching to be parallel with what comes in Question 4.

How many control units were matched? How many treated units?

Answer.

```r
m_out_data<-matchit(treat~race+married+nodegree,data = lalonde,method = 'exact')

summary(m_out_data)
```

```
##
## Call:
## matchit(formula = treat ~ race + married + nodegree, data = lalonde,
##     method = "exact")
##
## Summary of Balance for All Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
```

```
## raceblack         0.8432        0.2028          1.7615          .    0.6404
## racehispan        0.0595        0.1422         -0.3498          .    0.0827
## racewhite         0.0973        0.6550         -1.8819          .    0.5577
## married           0.1892        0.5128         -0.8263          .    0.3236
## nodegree          0.7081        0.5967          0.2450          .    0.1114
##            eCDF Max
## raceblack    0.6404
## racehispan   0.0827
## racewhite    0.5577
## married      0.3236
## nodegree     0.1114
##
## Summary of Balance for Matched Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## raceblack         0.8432        0.8432               0          .         0
## racehispan        0.0595        0.0595               0          .         0
## racewhite         0.0973        0.0973               0          .         0
## married           0.1892        0.1892               0          .         0
## nodegree          0.7081        0.7081               0          .         0
##            eCDF Max Std. Pair Dist.
## raceblack         0               0
## racehispan        0               0
## racewhite         0               0
## married           0               0
## nodegree          0               0
##
## Sample Sizes:
##               Control Treated
## All            429.       185
## Matched (ESS)  111.53     185
## Matched        429.       185
## Unmatched        0.         0
## Discarded        0.         0
```

Control=429 Treated=185

2.2. (4 points) Effect estimate

Estimate a linear regression model using your match data from 2.1. Include the treatment and all confounders from 2.1 in a linear, additive specification. Weight by the weights from matching.

What is the estimated effect of job training on earnings?

Answer.

```
m_out_data_get<-match.data(m_out_data)
lm(re78~race+married+nodegree+treat, data = m_out_data_get, weights = weights)
```

```
##
## Call:
## lm(formula = re78 ~ race + married + nodegree + treat, data = m_out_data_get,
##     weights = weights)
##
## Coefficients:
```

```
## (Intercept)   racehispan    racewhite      married     nodegree        treat
##      5858.9       2015.3       1120.5        685.1      -1663.8       1309.9
```

This means that, on average, individuals who received job training (treat=1) had higher earnings by approximately $1309.9 compared to those who did not receive job training (treat=0).

2.3. (4 points) Exact matching with high-dimensional confounding

Now suppose the adjustment set needs to also include 1974 earnings, `re74`. The adjustment set for this part is `race`, `married`, `nodegree`, and `re74`. Repeat exact matching as above.

How many control units were matched? How many treated units?

Answer.

```
m_out_new<-matchit(treat~race+married+nodegree+re74,data = lalonde,method = 'exact')
summary(m_out_new)
```

```
##
## Call:
## matchit(formula = treat ~ race + married + nodegree + re74, data = lalonde,
##     method = "exact")
##
## Summary of Balance for All Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## raceblack         0.8432        0.2028          1.7615          .    0.6404
## racehispan        0.0595        0.1422         -0.3498          .    0.0827
## racewhite         0.0973        0.6550         -1.8819          .    0.5577
## married           0.1892        0.5128         -0.8263          .    0.3236
## nodegree          0.7081        0.5967          0.2450          .    0.1114
## re74           2095.5737     5619.2365         -0.7211     0.5181    0.2248
##            eCDF Max
## raceblack    0.6404
## racehispan   0.0827
## racewhite    0.5577
## married      0.3236
## nodegree     0.1114
## re74         0.4470
##
## Summary of Balance for Matched Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## raceblack         0.8473        0.8473              -0          .         0
## racehispan        0.0458        0.0458               0          .         0
## racewhite         0.1069        0.1069              -0          .         0
## married           0.1603        0.1603               0          .         0
## nodegree          0.7405        0.7405               0          .         0
## re74              0.0000        0.0000               0          .         0
##            eCDF Max Std. Pair Dist.
## raceblack         0               0
## racehispan        0               0
## racewhite         0               0
## married           0               0
## nodegree          0               0
## re74              0               0
```

```
##
## Sample Sizes:
##                Control Treated
## All             429.       185
## Matched (ESS)    48.73     131
## Matched         108.       131
## Unmatched       321.        54
## Discarded         0.         0
```

From this new adjustment control units had 108 matches while treated units had 131 matches

2.4. (4 points) Examining matched units

Look at the `re74` values in the full data and among the matched units (no need to print this in your output). Explain what happened: what is different about the 1974 earnings of the matched vs the unmatched cases?

Here is one way to do this:

- Using the function `summary`, look at descriptive statistics of the `re74` values in the full data.
- Using the function `summary`, look at descriptive statistics of the `re74` values in the matched data. You can get the matched data using the `match.data` function.
- You can learn about how to use the `summary` function to look at descriptive statistics of R data **here**.

What do you notice? What is different about the values of `re74` in the full data versus the matched data? Explain what happened and why it happened.

Answer.

```
matched <- match.data(m_out_new)
summary(matched)
```

```
##      treat             age             educ           race        married
##  Min.   :0.0000   Min.   :16.00   Min.   : 1.00   black :148   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:18.00   1st Qu.: 9.00   hispan: 18   1st Qu.:0.0000
##  Median :1.0000   Median :24.00   Median :10.00   white : 73   Median :0.0000
##  Mean   :0.5481   Mean   :26.05   Mean   :10.14                Mean   :0.2134
##  3rd Qu.:1.0000   3rd Qu.:30.00   3rd Qu.:12.00                3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :55.00   Max.   :18.00                Max.   :1.0000
##
##     nodegree           re74        re75             re78
##  Min.   :0.0000   Min.   :0   Min.   :   0.0   Min.   :    0
##  1st Qu.:0.0000   1st Qu.:0   1st Qu.:   0.0   1st Qu.:    0
##  Median :1.0000   Median :0   Median :   0.0   Median : 3228
##  Mean   :0.7322   Mean   :0   Mean   : 452.7   Mean   : 5870
##  3rd Qu.:1.0000   3rd Qu.:0   3rd Qu.: 177.2   3rd Qu.: 9432
##  Max.   :1.0000   Max.   :0   Max.   :8853.7   Max.   :60308
##
##     weights           subclass
##  Min.   :0.07495   4      :90
##  1st Qu.:0.34351   7      :34
##  Median :1.00000   3      :33
##  Mean   :1.00000   1      :21
##  3rd Qu.:1.00000   6      :17
##  Max.   :3.06216   2      :14
##                    (Other):30
```

2.5. (4 points) Drawbacks of exact matching

Briefly interpret the result from 2.4: what is the drawback of using exact matching in this setting?

Answer.

From 2.4 we see the min, max, and median age and educ. The minimum age is 16 while its max is 55, while educ the minimum is 1 and the max is 18. Marriage and treat have values of 1 as their max and 0 as their minimum because both of these are binary values. The race heading is broken up into 3 races and shows the amount of time each one appeared. When we look at re74 it has 0s in every category. Now when looking as to why exact matching is a drawback in this situation is simply because there would have been a great loss of data in order to do this matching. Age and educ have such a large range that finding an exact match would mean discarding a large portion of the data. Earlier on we found that out of the 429 control units, 321 were unmatched and out of the 185 for the treated, 54 of them were unmatched both numbers being large.portions of the overall total

3. Parametric estimation

3.1. (5 points) Outcome modeling

In the code below, we use `lm()` to estimate an Ordinary Least Squares regression of future earnings `re78` on treatment `treat`, interacted with confounders: `race`, `married`, `nodegree`, and `re74`.

```
outcome_model <- lm(re78 ~ treat * (race + married + nodegree + re74),
                    data = lalonde)
```

Use the model above to estimate the average treatment effect among the *treated*.

To do this, you should

1. Create two data frames
   - The first should contain the *treated* individuals (with their factual treatment of `1`)
   - The second should contain the same *treated* individuals, but with `treat` set to the value `0`
2. Using the model above, predict the expected outcomes for the two data frames you created in step 1.
3. Report the average treatment effect among the treated.

Answer.

```
treated_d1<-lalonde|>
  mutate(treat=1)

treated_d0<-lalonde|>
  mutate(treat=0)

outcome_model <- lm(re78 ~ treat * (race + married + nodegree + re74),
                    data = lalonde)


pred_outcome<-lalonde|>
  mutate(
    yhat1=predict(outcome_model,treated_d1),
    yhat0=predict(outcome_model,treated_d0)
  )
head(pred_outcome)
```

```
##      treat age educ   race married nodegree re74 re75       re78    yhat1
## NSW1     1  37   11  black       1        1    0    0  9930.0460 6722.725
## NSW2     1  22    9 hispan       0        1    0    0  3595.8940 6090.326
## NSW3     1  30   12  black       0        0    0    0 24909.4500 7304.078
## NSW4     1  27   11  black       0        1    0    0  7506.1460 5064.535
## NSW5     1  33    8  black       0        1    0    0   289.7899 5064.535
## NSW6     1  22    9  black       0        1    0    0  4056.4940 5064.535
##         yhat0
## NSW1 3050.788
## NSW2 4784.800
## NSW3 4023.670
## NSW4 3168.427
## NSW5 3168.427
## NSW6 3168.427
```

```
pred_display<-pred_outcome|>
  summarize(average_yhat1 = mean(yhat1),
            average_yhat0 = mean(yhat0),
            average_effect = mean(yhat1 - yhat0))
pred_display
```

```
##   average_yhat1 average_yhat0 average_effect
## 1      7581.317      6252.046       1329.271
```

From the table, we can see that the average effect amongst the treated units is approximately 1329.2708096

3.2. (5 points) Treatment modeling: Creating weights

> Note: This part has much help from us. You should read what we have provided to understand, and you will do a small part at the end. We are doing this to maximize the learning-value-to-workload ratio of the problem.

Using the `glm()` below, we estimate the probability of treatment given confounders.

```
treatment_model <- glm(treat ~ race + married + nodegree + re74,
                       data = lalonde,
                       family = binomial)
```

Then, using the code below, we

- predict the probability that `treat = 1`
- generate the propensity score for each unit
- create a weight for estimating the Average Treatment Effect on the Treated, by the formula

$$w_i = \frac{P(A = 1 \mid \vec{L} = \vec{\ell}_i)}{P(A = a_i \mid \vec{L} = \vec{\ell}_i)}$$

> Note: For treated units, this weight is 1. For untreated units, the value varies.

```
with_weight <- lalonde %>%
  # Create the propensity score
  mutate(p_a_1 = predict(treatment_model, type = "response"),
         pscore = case_when(treat == 1 ~ p_a_1,
                            treat == 0 ~ 1 - p_a_1),
         weight = p_a_1 / pscore)
```

How many treated units does the most-heavily-weighted *untreated* unit represent? To answer this, you will want to determine the maximum weight amongst untreated individuals in `with_weight`.

Answer.

```
max_weight_untreated <- with_weight|>
  filter(treat == 0)|>
  summarize(max_weight = max(weight))

max_weight_untreated
```

```
##   max_weight
## 1   2.535103
```

The number of mostly heavily weighted untreated units is 2.5351028

3.3. (5 points) Treatment modeling: Estimating outcomes

Using the `with_weight` object, take *weighted means* of the observed outcomes `re78` weighted by `weight` to estimate the average outcome of treated units, and the weighted average outcome of control units (weighted to be comparable to the treated units).

> Hint: You will want to take a *weighted mean*, but *grouped by* treatment status.

Answer.

```
with_weight|>
  group_by(treat)|>
  summarise(
    weight_mean=weighted.mean(re78,weight)
  )
```

```
## # A tibble: 2 x 2
##   treat weight_mean
##   <int>       <dbl>
## 1     0       4825.
## 2     1       6349.
```

4. Matching without requiring exact matches

We hope that from this class you are prepared to learn new causal estimators, apply them in R, and explain what you have done. This question is a chance to practice! In class we discussed many matching approaches. For this question, you will choose your own approach. There are many correct answers, and you will be evaluated by the clarity of your code and explanations.

Task: Using `matchit`, conduct matching to estimate the ATT where `treat` is the treatment and the sufficient adjustment set is `race`, `married`, `nodegree`, and `re74`.

1. Use `matchit`, setting `method`, `distance`, and any other arguments to any values of your choosing. The only requirements are
   - `formula = treat ~ race + married + nodegree + re74`
   - `estimand = "ATT"`
2. Create matched dataset using `match.data()`
3. Estimate a linear regression model using `lm()` with the formula `re78 ~ treat + race + married + nodegree + re74` using your matched data, weighted by the `weights` that are produced by `match.data()`.

4.1. (4 points) Conduct the matching

This is space to conduct the matching. We expect this part to be an R code chunk.

Answer.

```
m_out_new2<-matchit(treat~race + married + nodegree + re74,data = lalonde,method = 'optimal',distance =

summary(m_out_new2)
```

```
##
## Call:
## matchit(formula = treat ~ race + married + nodegree + re74, data = lalonde,
##     method = "optimal", distance = "mahalanobis", estimand = "ATT")
##
## Summary of Balance for All Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## raceblack         0.8432        0.2028          1.7615          .    0.6404
## racehispan        0.0595        0.1422         -0.3498          .    0.0827
## racewhite         0.0973        0.6550         -1.8819          .    0.5577
## married           0.1892        0.5128         -0.8263          .    0.3236
## nodegree          0.7081        0.5967          0.2450          .    0.1114
## re74           2095.5737     5619.2365         -0.7211     0.5181    0.2248
##            eCDF Max
## raceblack    0.6404
## racehispan   0.0827
## racewhite    0.5577
## married      0.3236
## nodegree     0.1114
## re74         0.4470
##
## Summary of Balance for Matched Data:
##            Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## raceblack         0.8432        0.4703          1.0259          .    0.3730
## racehispan        0.0595        0.0595          0.0000          .    0.0000
## racewhite         0.0973        0.4703         -1.2585          .    0.3730
## married           0.1892        0.2054         -0.0414          .    0.0162
## nodegree          0.7081        0.6865          0.0476          .    0.0216
## re74           2095.5737     2602.7324         -0.1038     1.0924    0.0586
##            eCDF Max Std. Pair Dist.
```

```
## raceblack    0.3730          1.0259
## racehispan   0.0000          0.0000
## racewhite    0.3730          1.2585
## married      0.0162          0.0414
## nodegree     0.0216          0.0476
## re74         0.3135          0.1749
##
## Sample Sizes:
##           Control Treated
## All           429     185
## Matched       185     185
## Unmatched     244       0
## Discarded       0       0
```

```
matched2 <- match.data(m_out_new2)
summary(matched2)
```

```
##      treat          age             educ          race        married
##  Min.   :0.0   Min.   :16.00   Min.   : 1.0   black :243   Min.   :0.0000
##  1st Qu.:0.0   1st Qu.:18.00   1st Qu.: 9.0   hispan: 22   1st Qu.:0.0000
##  Median :0.5   Median :23.00   Median :10.0   white :105   Median :0.0000
##  Mean   :0.5   Mean   :25.47   Mean   :10.2                Mean   :0.1973
##  3rd Qu.:1.0   3rd Qu.:29.00   3rd Qu.:12.0                3rd Qu.:0.0000
##  Max.   :1.0   Max.   :55.00   Max.   :18.0                Max.   :1.0000
##
##     nodegree           re74            re75           re78          weights
##  Min.   :0.0000   Min.   :    0   Min.   :    0   Min.   :    0   Min.   :1
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:    0   1st Qu.:    0   1st Qu.:1
##  Median :1.0000   Median :    0   Median :    0   Median : 3529   Median :1
##  Mean   :0.6973   Mean   : 2349   Mean   : 1465   Mean   : 5682   Mean   :1
##  3rd Qu.:1.0000   3rd Qu.: 1950   3rd Qu.: 1656   3rd Qu.: 8896   3rd Qu.:1
##  Max.   :1.0000   Max.   :35040   Max.   :25142   Max.   :60308   Max.   :1
##
##     subclass
##  1      :  2
##  2      :  2
##  3      :  2
##  4      :  2
##  5      :  2
##  6      :  2
##  (Other):358
```

4.2. (2 points) Explain your choices

In a few sentences, tell us about the matching approach you have chosen.

Answer.

When doing my own matching I decided to change both the method as well as the distance variable values. I chose optimal for my method variable, it was either that or nearest. Optimal was chosen as it performs better compared to nearest as long as the data isn't too long. Another reason for this is simply because by using optimal, matches are selected with the smallest possible pairwise distance. For distance, I chose Mahalanobis because the variables chosen in the formula would need to be on the same scaling.Married,and nodegree both have a similar scaling, but re74 has such a wide scaling that could throw things off, so mahalanobis was used to make it scaled down.

4.3. (2 points) How many units did you keep?

Report the number of treated and control units in the original data, and how many were kept by your matching procedure.

Answer.

In total I had 429 Control units and 185 Treated units. In the end I was able to keep all 185 for Treated units but I was sadly only able to keep 185 out of the 429 Control units

4.4. (2 points) Report your causal estimate

What do you estimate for the average treatment effect on the treated? This is the coefficient on `treat` in the linear regression you fit on the matched data.

Answer.

```
lm(re78 ~ treat + race + married + nodegree + re74,
                    data =matched2,weights = weights)
```

```
##
## Call:
## lm(formula = re78 ~ treat + race + married + nodegree + re74,
##     data = matched2, weights = weights)
##
## Coefficients:
## (Intercept)        treat   racehispan    racewhite      married     nodegree
##   5056.0477    1747.6294    1613.5360     729.9154     768.8367   -1629.2968
##        re74
##      0.1846
```

1747.6294 is the estimated average treatment effect on the treated