ARE 213 PS 1b

S. Sung, H. Husain, T. Woolley, A. Watt

2021-10-18

Contents

Packages

```
# install.packages("pacman")
# install.packages("plm")
# install.packages('foreign')
# install.packages('stargazer')
# install.packages("finalfit")
# install.packages("glmnet")
# install.packages("jtools")
# install.packages("Hmisc")
library(tidyverse)
library(foreign)
library(xtable)
library(stargazer)
library(finalfit)
library(glmnet)
library(jtools)
```

Problem 1

In Problem Set 1a, you used linear regression to relate infant health outcomes and maternal smoking during pregnancy.

Data cleaning from PS 1a

```
data = read.dta('ps1.dta')
missing_codes = read.csv('missing_codes.csv')
mvars = as.character(missing_codes$varname)
missing_codes$num_missing = as.integer(0)
for (row in 1:nrow(missing_codes)) {
  var = as.character(missing_codes[row, "varname"])
  code  = as.numeric(missing_codes[row, "missing_code"])
  nmissing = as.integer(sum(data[, var] == code))
  missing_codes$num_missing[missing_codes$varname==var] = nmissing
  data[, var] = na_if(data[, var], code)
}
# Convert all variables with <7 unique values to factor (and 3 additional variables)
factor_vars = c("isllb10", "birmon", "weekday")
for (var in colnames(data)) {
  if (length(unique(data[!is.na(data[, var]), var])) < 7 || var %in% factor_vars) {
    data[, var] = factor(data[, var])
  }
}
# label data
variable_labels_df = read.csv('variable_labels.csv')
variable_labels <- setNames(as.character(variable_labels_df$label), variable_labels_df$varname)
data <- Hmisc::upData(data, labels = variable_labels)

# Dataframe with missing dropped
df = data[complete.cases(data), ]
```

Part (a)

Under the assumption of random assignment conditional on the observables, what are the sources of misspecification bias in the estimates generated by the linear model estimated in Problem Set 1a?

Part (b)

Now, consider a series estimator. Estimate the smoking effects using a flexible functional form for the control variables (e.g., higher order terms and interactions). What are the benefits and drawbacks to this approach?

```
df1b = df %>%
  select(dbrwt, tobacco, rectype, pldel3, birattnd, cntocpop, stresfip, dmage,
         ormoth, mrace3, dmeduc, dmar, adequacy, nlbnl, dlivord, dtotord,
         totord9, nprevist, disllb, isllb10, dfage, orfath, dfeduc,
         weekday, csex, delmeth5, cardiac, diabetes, herpes, chyper, preterm)


vartypes = sapply(df1b, class)
print(vartypes)
```

```r
# indicator vars (no higher order terms)
vars1 = names(Filter(is.factor, select(df1b, -dbrwt)))

# quantitative var (create higher order terms)
vars2 = names(Filter(is.integer, select(df1b, -dbrwt)))



df1b = df %>%
  select(dbrwt, mrace3, csex, cardiac, diabetes, herpes, chyper)
reg1b <- lm(
  dbrwt ~ tobacco, polym(mrace3, csex, cardiac, diabetes, herpes, chyper, degree=2, raw=TRUE),
  data = df)
reg1b
summ(reg1b)
```

Part (c)

Use the LASSO to determine which covariates (and higher order terms) to include in your regression from part (b). Do you end up dropping some covariates that you had thought might be necessary to include?

```r
X = df %>%
  select(mrace3, csex, cardiac, diabetes, herpes, chyper)
y = df$dbrwt
reg1c = glmnet(X, y, family="gaussian", alpha=1)
reg1c$beta[,3]


m2 <- do.call(polym,c(as.list(X),degree=2, raw=TRUE))
ncol(m2)
```

Problem 2

Describe the propensity score approach to the problem of estimating the average causal effect of smoking when the treatment is randomly assigned conditional on the observables. How does it reduce the dimensionality problem of multivariate matching? Try a few ways to estimate the effects of maternal smoking on birthweight:

Part (a)

First create the propensity score. For our purposes let's use a logit specification. First specify the logit using all of the "predetermined" covariates (don't include interactions). Next, include only those "predetermined" covariates that enter significantly in the first logit specification. How comparable are the propensity scores? If they are similar does this imply that we have the "correct" set of covariates in the logit specification used for our propensity score?

```r
#   create the propensity score using logit
#   using all of the "predetermined" covariates


# then try logit with only the significant covariates
```

Part (b)

Control directly for the estimated propensity scores using a regression analysis, and estimate an average treatment effect. State clearly the assumptions under which your estimate is correct.

```
# Control for p-score in regression analysis

# Estimate ATE
```

PART (c) . . . . . .

Reweight data using p-score to weight

Estimate ATE

Estimate TOT with rewieghted data

PART (d) . . . . . .

Estimate the counterfactual birthweight densities with a kernel density estimator

See Joel's notes for kernel density estimator

Play around with a bandwidth starting with half the default Stata bandwidth

For stata bandwidht, see rkdensity.pdf page 9 in this ps1b github folder.

You can also run on stata with no bandwidth specified, then print the

default bandwidth used using `display r(bwidth)`

Choose the same bandwidth for all the pictures

Graph both kernel densities over range of birthweight in the same plot

calculate the kernel estimator at birthweight equals 3,000 grams

PART (e) . . . . . .

Plot one of your densities with multiple bandwidths in addition to the one used above

PART (f) . . . . . .

Might need to display summary stats for people with high and low p-scores

PART (g) . . . . . .

Create tables to present relationship between birthweight and smoking

#===
PROBLEM 3 #===

estimate the “non-parametric” conditional mean of birth weight as a function # of the estimated probability of smoking, separately for smokers and non-smokers

divide the data from smokers/ non-smokers into 100 approximately equally spaced

bins based on the estimated propensity score.

Use blocking estimator discussed in class

#==========
PROBLEM 4 #==========
Create indicator for <2500 grams

Redo question 3 using indicator instead of birthweight

#==========
PROBLEM 5 #==========

PART (a) . . . . . .

Select variables

df5a = df %>% select(birthweight, rectype, pldel3, cntocpop, stresfip, dmage, mrace3, dmar, adequacy, csex, dplural)

Run regression

PART (b) . . . . . .

Select vars and smoking indicator

df5b = df %>% select(birthweight, rectype, pldel3, cntocpop, stresfip, dmage, mrace3, dmar, adequacy, csex, dplural)

Run regression

PART (c-d) . . . . . .

weighted version of the exact matching estimator that estimates the

same thing as the regression above

Compare to regression from (b)

PART (f) . . . . . .

Compute a standard error for your matching estimator using the formula from Imbens (2015).

compute the conditional variance of estimator from (d)

```
#=========================================================================
# PROBLEM 6 #=============================================================
# Summarize and give intuition # Is our best estimate of the effects of smoking credibly identified?
```