

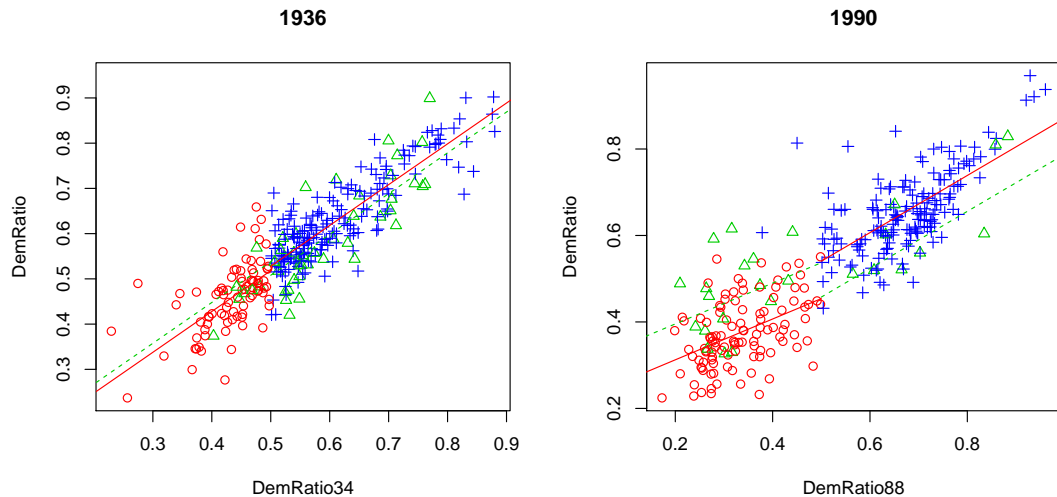
Stat 505 Assignment 10 Solutions

1. Elections

- (a) Take data from a particular year not ending in 2 and pull out districts where the election was contested AND the election 2 years before was contested. Estimate the incumbency effect.

I'm using both 1936 and 1990

- (b) Plot fitted model and data. Discuss political interpretation of coefficient estimates.



After accounting for democratic share of the vote in the previous election, it seems that incumbency effect in 1936 was 0.02 ($SE = .009$) which is much smaller than in 1990 when it's estimated as 0.083 ($SE=0.013$). I did include party as another predictor, so we are seeing a “pure” incumbent effect after adjusting for which party last held the seat.

- (c) What have we assumed?

To make causal inferences about incumbency we must assume that we have attained ignorability – that is that we have adjusted away all lurking variables. Certainly there are lots of variables which could plausibly affect the Democratic vote share in a particular year. However, looking across the last 120 years, it makes some sense to me to say that prior democratic vote share, party of incumbent, and whether it was an open seat or not are a good batch of variables to adjust for. You might also say that we have extrapolated across the .50 line to use counterfactuals on each side. So we must assume that relationships between incumbency and prior Dem.Share on current Demshare are the same across the .50 boundary.

2. Load the arm package and type data(lalonde). This is the data used by Dehejia and Wahba. Gelman has more data on his website, but the 185 treated cases are the same, and there are big samples of data from CPS and PSID. Lalonde and subsequent authors are comparing results of an experiment to results we might get if we use controls

from general survey data. The underlying question is "How well do propensity scores and other such techniques work to answer causal questions?" In particular, are these methods biased compared to the actual treatment effects we compute in part (a)? See Smith & Todd for more comparison and interpretation.

(a) Using the experimental data

Using treatment as sole predictor we get an estimated treatment effect of \$1794 with standard error \$633.

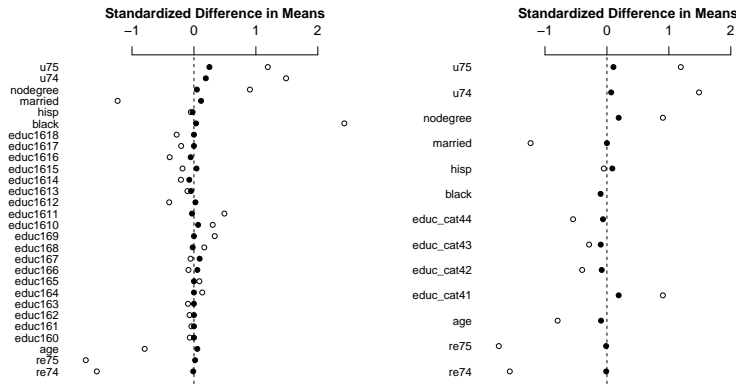
Using all the available predictors (education coded as a 4-level factor), the treatment effect estimate is \$1528 with SE = \$640. The two estimates differ by less than half an SE. Precision is basically the same. I would say, on general principles, that we should adjust for the other variables before evaluating the treatment effect, so I like the second estimate.

(b) Using all the CPS data as controls.

Simply comparing treatment to controls gives a treatment effect of -\$8497 with SE = \$712, way different from that in (a).

Adjusting for the other variables makes a huge difference: changing the treatment effect to \$773 with SE = \$548.

(c) Propensity scoring on the CPS data.



My first propensity scores uses education as a 12 level factor, the second as 4 levels. Based on the balance plots, I like using the 4 level factor for education rather than a 12 level factor because the black dots are slightly closer to the zero line.

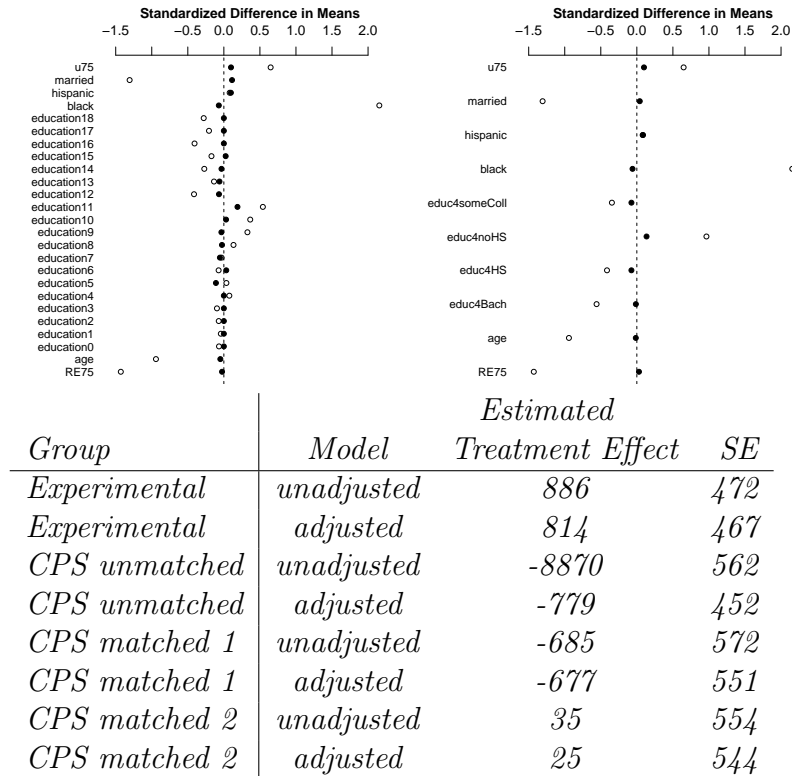
Just the raw treatment effect is estimated as \$1509 with SE \$724 using matches from the first propensity score and \$1776 (SE = \$751). When using all the adjusting variables, the first propensity score matched data gives an estimated treatment effect of \$1500 with SE \$713, the second of \$1787 with SE \$706.

(d) What did we estimate in (b) and (c)?

In (b) we compared the NWS treatment group to the entire CPS dataset and found that they were quite different. For example, a much higher proportion of the NWS folks were unemployed in 1974 and 1975 than for the CPS folks. The plots in (c) illustrate how different they were – looking at the unmatched open circles we see lots of strong differences before matching.

In (c) we are comparing the NWS treatment group against a specially selected subsample of the CPS group. Each treated “case” is matched to one untreated “control” according to proximity of propensity score. Since we have 16000 in the CPS data, it’s not surprising that we can find matches which reduce the differences dramatically between the two groups.

- (e) Redo, excluding earnings in 1974. When we drop that variable, more data is available for the experimental subjects, so use these data. What does this say about ignorability?



I note that experimental treatment effect sizes are half what they were in (a), with slightly smaller SE. The unmatched CPS data gives an unadjusted estimate that’s way off, as was the case in (b), and the adjusted value has similar magnitude to that of (b), but the wrong sign. We then use propensity scores to try to remedy the lack of control, and are not very successful. Adjusted estimates are -677 and 25, depending on which model for education I use. These are more than 3 SE from the desired value \$814 in the second line. This shows that ignorability of treatment does not hold (that we get much different answers) when we don’t control for income in 1974.

R Code

```
options(continue="+" , width=120)
#require(xtable)
require(arm)
congress ← read.csv("../data/congress.csv")
congress[congress== -9] ← NA
```

```

congress$DemRatio ← with(congress, Dem/(Dem + Rep))
congress$DemRatio ← ifelse(congress$DemRatio < 1.0e-10 | congress$DemRatio > 1 - 1.0e-10, NA
, congress$DemRatio)
elect36 ← subset(congress, year==1936)
elect34 ← subset(congress, year==1934)
names(elect34)[7] ← "DemRatio34"
elect1 ← merge(elect36, elect34[, c(2,3,7)])
elect1$party ← factor(c("Rep", NA, "Dem")[2+elect1$incumbent])
elect1$party[is.na(elect1$party)] ← factor(ifelse(elect1$DemRatio34[is.na(elect1$party)]
<1/2, "Rep", "Dem"))
elect36.fit ← lm(DemRatio ~ DemRatio34 + incumbent + party, elect1)
elect90 ← subset(congress, year==1990)
elect88 ← subset(congress, year==1988)
names(elect88)[7] ← "DemRatio88"
elect2 ← merge(elect90, elect88[, c(2,3,7)])
elect2$party ← factor(c("Rep", NA, "Dem")[2+elect2$incumbent])
elect2$party[is.na(elect2$party)] ← factor(ifelse(elect2$DemRatio88[is.na(elect2$party)]
<1/2, "Rep", "Dem"))
elect90.fit ← lm(DemRatio ~ DemRatio88*party + incumbent, elect2)
par(mfrow=c(1,2))
plot(DemRatio ~ DemRatio34, elect1, pch = incumbent+2, col = incumbent+3, main = "1936")
curve( 0.057+.03+ 0.902*x, add=T, from=0, to=.5, lty=2, col=3) #openseat Rep
curve( 0.057+.03-.0199 + 0.902*x, add=T, from=0, to=.5, lty=1, col=2) #inc Rep
curve( 0.057+ 0.902*x, add=T, from=.5, to=1, lty=2, col=3) #openseat Dem
curve( 0.057+.0199 + 0.902*x, add=T, from=.5, to=1, lty=1, col=2) #inc Dem
plot(DemRatio ~ DemRatio88, elect2, pch = incumbent+2, col = incumbent+3, main = "1990")
curve( 0.127+.174 + 0.472*x, add=T, from=0, to=.5, lty=2, col=3) #openseat Rep
curve( 0.127+.174 -.083 + 0.472*x, add=T, from=0, to=.5, lty=1, col=2) #inc Rep
curve( 0.127+ 0.661*x, add=T, from=.5, to=1, lty=2, col=3) #openseat Dem
curve( 0.127+.083 + 0.661*x, add=T, from=.5, to=1, lty=1, col=2) #inc Dem
display(elect36.fit)
display(elect90.fit)
data(lalonde)
lalond.fit1 ← lm(re78 ~ treat, lalonde)
lalonde$educ4 ← factor(ifelse(lalonde$educ<13, 1, ifelse(lalonde$educ<16, 3, 4)), label=c("noHS", "HS", "someColl", "Bach"))
lalond.fit2 ← lm(re78~treat+age+educ4+black+hispanic+married+re74+re75+u74+u75, lalonde)
gelman ← foreign::read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/lalonde/
NSW.dw.obs.dta")
gelman.fit1 ← lm(re78 ~ treat, gelman, subset= sample < 3)
gelman.fit2 ← lm(re78 ~ treat + age+ factor(educ_cat4)+ black+ hispanic+ married+ re74+ re75,
data = gelman, subset = sample < 3)
gelman$u74 ← ifelse(gelman$re74 > 0, 0, 1)
gelman$u75 ← ifelse(gelman$re75 > 0, 0, 1)
gelman$educ16 ← factor(gelman$educ)
propensity.fit1 ← glm(treat ~ re74 + re75 + age + educ16 + black + hispanic
+ married + nodegree + u74 + u75, family=binomial, data=gelman[1:16179,])
pscores ← predict(propensity.fit1, type="response")
matches ← matching(z=gelman$treat[1:16179], score=pscores)
matched ← gelman[matches$matched,]
balance1 ← balance(gelman[1:16179,], matched, propensity.fit1)
# print(balance1)
par(mfrow=c(1,2))
plot(balance1)
gelman$educ_cat4 ← factor(gelman$educ_cat4)
propensity.fit2 ← glm(treat ~ re74 + re75 + age + educ_cat4 + black + hispanic
+ married + nodegree + u74 + u75, family=binomial, data=gelman[1:16179,])
pscore2 ← predict(propensity.fit2, type="response")
matches2 ← matching(z=gelman$treat[1:16179], score=pscore2)
matched2 ← gelman[matches2$matched,]
balance2 ← balance(gelman[1:16179,], matched2, propensity.fit2)
# print(balance2)
plot(balance2)
summary(lm(re78 ~ treat, data=gelman[1:16179,], subset=matches$matched))
gelman.propfit1 ← lm(re78 ~ re74 + re75 + age + educ + black + hispanic+ married + u74 + u75 +
treat, data=matched)
summary(gelman.propfit1)
gelman.propfit2 ← lm(re78 ~ re74 + re75 + age + educ_cat4 + black + hispanic+ married + u74 +

```

```

    u75 +treat ,data=matched2)
summary(gelman.propfit2)
nswAll ← read.csv("http://www.math.montana.edu/~jimrc/classes/stat505/data/nswAll.csv")
nswAll.fit1 ← lm(RE78~treatment, nswAll)
nswAll$educ4 ← factor( ifelse( nswAll$educ<12,1, ifelse( nswAll$educ<13,2,
    ifelse( nswAll$educ<16,3,4) ) ),label=c("noHS","HS","someColl","Bach"))
nswAll$u75 ← ifelse( nswAll$RE75>0,0,1)
nswAll.fit2 ← lm(RE78~treatment + age+educ4+black+hispanic+married+RE75+u75, nswAll)
summary( nswAll.fit2)
## names(gelman)[c(11,1,2,3,9,4,5,7,8,15,13)] ← names(nswAll)
## partE ← rbind(gelman[gelman$sample==2,c(11,1,2,3,9,4,5,7,8,15,13)], nswAll[nswAll$
    treatment==1,])
partE ← read.csv("../data/combinedNSWall-CPS.csv")
partE$education ← factor(partE$education)
summary(lm(RE78~treatment, partE))
summary(lm(RE78~-educ4, partE))
propensity.fitE1 ← glm(treatment ~ RE75 + age + education + black + hispanic+ married +
    u75, family=binomial, data=partE)
pscoresE1 ← predict(propensity.fitE1, type="response")
matchesE1 ← matching(z=partE$treatment, score=pscoresE1)
matchedE1 ← partE[matchesE1$matched,]
balanceE1 ← balance(partE, matchedE1, propensity.fitE1)
print(balanceE1)
par(mfrow=c(1,2))
plot(balanceE1)
propensity.fitE2 ← glm(treatment ~ RE75 + age + educ4 + black + hispanic+ married + u75,
    family=binomial, data=partE)
pscoresE2 ← predict(propensity.fitE2, type="response")
matchesE2 ← matching(z=partE$treatment, score=pscoresE2)
matchedE2 ← partE[matchesE2$matched,]
balanceE2 ← balance(partE, matchedE2, propensity.fitE2)
print(balanceE2)
plot(balanceE2)
summary( lm(RE78 ~ treatment, data=partE, subset=matchesE1$matched))
summary(partE.propfit1 ← lm(RE78 ~ RE75 + age + education + black + hispanic+ married +
    u75 +treatment, data=partE, subset=matchesE1$matched))
summary( lm(RE78 ~ treatment, data=partE, subset=matchesE2$matched))
summary(partE.propfit2 ← lm(RE78 ~ RE75 + age + educ4 + black + hispanic+ married + u75 +
    treatment, data=partE, subset=matchesE2$matched))

```