Balance Checks

Annie Chen

1/15/2020

Admin

- Wednesday 5:30pm-6:30pm lab
- Problem set 1 + GitHub/markdown issues?
- Final project

Balance in Randomized Experiments

- Randomization balances both observed and unobserved pre-treatment covariates between the treated and untreated in large samples.
- Why check for balance between groups?
- Review of Hypothesis Testing:
- Test-statistic?
- P-value?

Balance Checks

- Conduct balance checks with respect to observed pre-treatment covariates
- Compare means, standard deviations etc. between the treated and untreated; can also regress treatment indicator on covariates
- Statistical tests for the difference between groups.
- Visualizations

An example: *Ethnic quotas and Political Mobilization*

- Dunning and Nilekani (2013) investigate the effect of ethnic quotas on redistribution in India.
- Are quotas for council presidencies an effective means of channeling benefits to marginalized groups?
- Comparing reserved (treated) and unreserved (untreated) council presidencies for Scheduled Castes (SCs) and Scheduled Tribes (STs) in three Indian states (Karnataka, Rajasthan, and Bihar).
- Unit of analysis is the village council constituency
- Regression Discontinuity identification strategy – but let's set that aside for now.

Exercise: *Ethnic quotas and Political Mobilization*

```
library(foreign)
library(dplyr)

data = read.dta(file.path(path, "/data/dunning_bal.dta"))
#glimpse(data)
```

- Try it yourself! Create a balance table for the following covariates, given the treatment variable `scst_reserved_current`.
- `P_ILL`: Mean number of illiterates
- `MARGWORK_P`: Mean number of marginal workers
- `No_HH`: Number of households
- `MAIN_AL_P`: Mean agricultural laborers
- `MAIN_CL_P`: Mean cultivators
- `NON_WORK_F`: Mean female nonworkers
- Null hypothesis of usual balance tests…
- H_0: treatment and control groups are the same

Ethnic quotas and Political Mobilization

- Choose the covariates to balance on…

```
vars = data %>%
  dplyr::select(P_ILL, MARGWORK_P, No_HH, MAIN_AL_P, MAIN_CL_P, NON_WORK_F)
```

- Calculate the mean and SD by treatment status for each covariate

```
bal.mean = aggregate(vars, by=list(data$scst_reserved_current),
                     function(x) mean(x, na.rm =T))

bal.sd = aggregate(vars, by=list(data$scst_reserved_current),
                   function(x) sd(x, na.rm =T))
```

Ethnic quotas and Political Mobilization

```
# calculate difference in means for each covariate (w/ for loop)
diff.means <- vector()
for (i in 1:6) {
  diff.means[i] <- mean(vars[data$scst_reserved_current==1, i], na.rm = T) -
    mean(vars[data$scst_reserved_current==0, i],na.rm = T)
}

diff.means
```

```
## [1] -257.58130  -12.47508  -99.01548  -14.13658  -57.94011 -198.13824
```

```
# Test the difference in means between conditions for each covariate (using apply)
# keep the p-values
diff_means_pval <- function(x) {t.test(vars[data$scst_reserved_current==1, x],
                                       vars[data$scst_reserved_current==0, x])$p.value}

bal.pv = sapply(1:length(vars), diff_means_pval)

bal.pv
```

```
## [1] 0.20694762 0.77969243 0.09224547 0.79410448 0.23259976 0.11403539
```

Ethnic quotas and Political Mobilization

```
# Stack
bal = rbind(bal.mean, bal.sd, c(NA, diff.means), c(NA, bal.pv))

# Rearrange
bal = t(bal)
bal = bal[-1, 1:6]

# and label the balance table
colnames(bal) = c("Control_Mean", "Control_SD", "Treat_Mean",
                  "Treat_SD", "Diff_Means", "ttest_p-val" )
```

Ethnic quotas and Political Mobilization

```
as_tibble(bal) %>%
  mutate(Covariate = c("P_ILL", "MARGWORK_P", "No_HH",
                       "MAIN_AL_P", "MAIN_CL_P", "NON_WORK_F")) %>%
  dplyr::select(Covariate, everything()) %>%
  kable(type = "text")
```

Covariate	Control_Mean	Control_SD	Treat_Mean	Treat_SD	Diff_Means	ttest_p-val
P_ILL	3928.7519	3671.1706	2462.1699	2132.9770	-257.58130	0.2069476
MARGWORK_P	729.5465	717.0714	539.9272	464.7068	-12.47508	0.7796924
No_HH	1404.1978	1305.1823	597.6406	517.1067	-99.01548	0.0922455
MAIN_AL_P	571.5969	557.4603	622.5571	600.0217	-14.13658	0.7941045
MAIN_CL_P	933.6822	875.7421	600.0342	490.4952	-57.94011	0.2325998
NON_WORK_F	2391.3605	2193.2222	1515.2815	1305.7212	-198.13824	0.1140354

- Why might this procedure be problematic?

You can also visualize Treatment v. Control groups

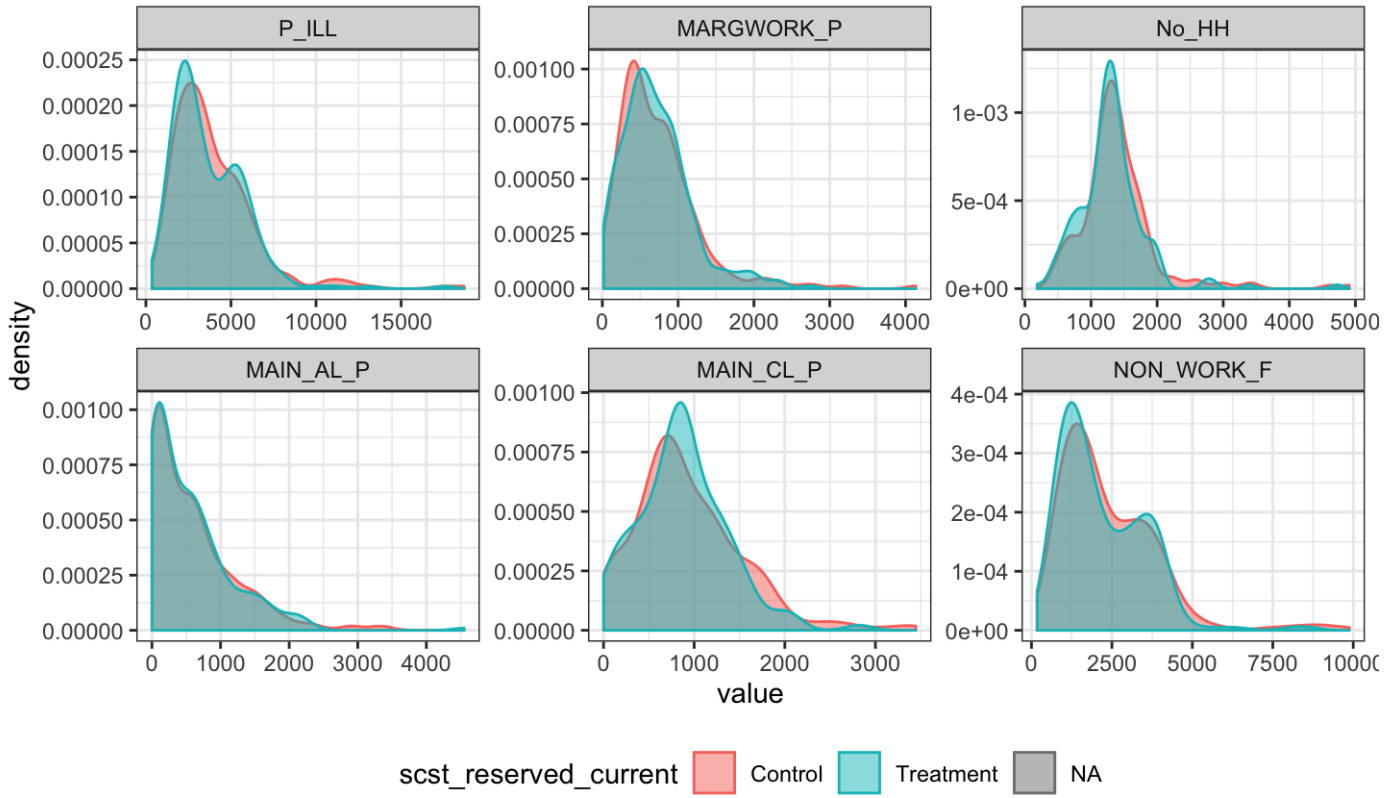
```
library(ggplot2)
library(reshape2)

# select the relevant variables
data_plot = data %>%
  dplyr::select(P_ILL, MARGWORK_P, No_HH,
                MAIN_AL_P, MAIN_CL_P, NON_WORK_F, scst_reserved_current) %>%
  # add id and factor treatment variables
  mutate(id = row_number(),
         scst_reserved_current = factor(scst_reserved_current,
                                        labels = c("Control", "Treatment")))

# Melt the data for easy plotting with facets in ggplot
data.melt = melt(data_plot, id.vars = c("id", "scst_reserved_current"))

# plot the densities
bal_plot <- ggplot(data.melt, aes(x = value, fill = scst_reserved_current,
                                  color = scst_reserved_current)) +
  geom_density(alpha=.5) +
  facet_wrap(~ variable, scales="free") +
  theme_bw() +
  theme(legend.position="bottom")
```

You can also visualize Treatment v. Control groups



F-tests

- Testing the joint significance of the difference in means between treated and untreated groups across all covariates.
- The test-statistic for an F-test is given by (the F-statistic):

$$F = \frac{\left(\frac{RSS_{res} - RSS_{unres}}{q}\right)}{\left(\frac{RSS_{unres}}{n-p-1}\right)}$$

  where q is the number of restrictions p is the number of independent variables.
- What is the null?
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

```
pre_bal <- lm(scst_reserved_current ~ P_ILL + MARGWORK_P + No_HH + MAIN_AL_P + MAIN_CL_P + NON_WORK_F,
   data = data)
summary(pre_bal)
```

```
##
## Call:
## lm(formula = scst_reserved_current ~ P_ILL + MARGWORK_P + No_HH +
##     MAIN_AL_P + MAIN_CL_P + NON_WORK_F, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6208 -0.4950 -0.2230  0.4944  0.7311
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.264e-01  7.839e-02   7.992 1.88e-14 ***
## P_ILL       -1.292e-05  4.254e-05  -0.304    0.761
## MARGWORK_P   5.888e-05  8.803e-05   0.669    0.504
## No_HH       -1.768e-04  1.216e-04  -1.453    0.147
## MAIN_AL_P    9.022e-05  8.523e-05   1.059    0.291
## MAIN_CL_P    5.793e-06  6.001e-05   0.097    0.923
## NON_WORK_F   1.791e-05  5.551e-05   0.323    0.747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5018 on 356 degrees of freedom
##   (151 observations deleted due to missingness)
## Multiple R-squared:  0.01213,    Adjusted R-squared:  -0.00452
## F-statistic: 0.7285 on 6 and 356 DF,  p-value: 0.6269
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

Average Treatment Effect

- Outcome variable (`ave_jobbenefit01`): "…asked citizens whether they had received a job or benefit from the village council in the previous year."

```
# Apply the difference in means estimator
ybar <- tapply(dat$ave_jobbenefit01,
               list('treated'= dat$scst_reserved_current),
               function(x) mean(x, na.rm = T))
ybar['1'] - ybar['0']
```

```
##          1
## 0.01190874
```

```
# Estimate the standard error of the difference in means
seDiffMeans <- function(y, tx){
  y1 = y[tx == 1]
  y0 = y[tx == 0]
  n1 = length(y1)
  n0 = length(y0)
  sqrt(((var(y1)/n1 + var(y0)/n0)))
}

seDiffMeans(dat$ave_jobbenefit01, dat$scst_reserved_current)
```

```
## [1] 0.02935379
```

Try computing the difference using bivariate OLS.

- How does this compare to the difference-in-means estimator? What about the SEs?

Try computing the difference using bivariate OLS.

- How does this compare to the difference-in-means estimator? What about the SEs?

```
mod.bivariate = lm(ave_jobbenefit01 ~ scst_reserved_current, data=data)
summary(mod.bivariate)$coefficients
```

```
##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)           0.25299959 0.02089318 12.1091944 1.553798e-29
## scst_reserved_current 0.01190874 0.02935982  0.4056135 6.852128e-01
```

- Now, re-estimate using robust standard errors.

Re-estimate using robust standard errors.

```
library(sandwich)

se.bivariate = sqrt(diag(vcovHC(mod.bivariate, type='HC2')))
stargazer(mod.bivariate, se=list(se.bivariate),
          keep.stat=c('n'), digits=8, notes = "HC2 Robust SEs", type="text")
```

```
##
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                             ave_jobbenefit01
## ------------------------------------------------
## scst_reserved_current         0.01190874
##                              (0.02935379)
##
## Constant                     0.25299960***
##                              (0.02072343)
##
## ------------------------------------------------
## Observations                        468
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
##                                     HC2 Robust SEs
```