Statistical Adjustment and Assessment of Adjustment in Observational Studies

ICPSR 2019 Session 2

August 05, 2019

Today

1. Agenda: The problem of covariance adjustment to reduce "bias"/ confounding. **How can we answer the question about whether we have adjusted enough.** A simple approach: stratification on one categorical variable (and interaction effects). A more complex approach: find sets that are as similar as possible in terms of a continuous variable (bipartite matching). Balance assessment after stratification.
2. Reading for tomorrow and next week: DOS 8–9, 13 and Gelman and Hill, 2006, § 9.5, and Ho et al., 2007
3. Questions arising from the reading or assignments or life?

1. Strategies for Causal Inference
2. But first, how to assess the randomization process in an experiment.
3. Did we control for enough?

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies and Workflow for randomized studies

Before fielding the study

- Plan the design for interpretability and power.
- Pre-register your analysis plan and experimental design.

After outcome data have been collected

- Make the case that randomization worked.
- Increase precision with design and test statistics (and even models of effects)
- Make sure that your estimator is estimating the right thing (i.e. is not biased or at least is consistent)
- Make sure that your test has a controlled false positive rate and is as powerful as possible (i.e. the coverage rate of a confidence interval is nominal, the Type I error rate is controlled).

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

Strategies for Adjustment of Observational Studies

- If you have randomized Z but not D, then IV. (Not really an observational study. **What additional arguments do you have to make?**)
- Find a discontinuity/A Natural Experiment (RDD: either natural experiment or continuous forcing function)
- Multiple controls (i.e. "Choice as an alternative to control" Rosenbaum, 1999)
- "Controlling For"
- Difference in Differences
- Matched Stratification (approximating a block-randomized experiment)
- Best matched subset selection (approximating a completely or simply randomized experiment)
- Weighting
- Direct theoretical modeling / Directed Acyclic Graphs (DAGS) to guide other data modeling choices.

1. Strategies for Causal Inference
2. But first, how to assess the randomization process in an experiment.
3. Did we control for enough?

The Neyman-Rubin Model for (simple) experiments

This is what randomization ensures:

$$(y_t, y_c, X) \perp Z$$

I.e., each of X, y_c and y_t is balanced between treatment and control groups (in expectation; given variability from randomization).

- In controlled experiments, random assignment justifies this argument.
- In natural experiments, justified otherwise, this is an article of faith.
- In an experiment, the *x*es aren't necessary for inference (although they can be used, carefully, to increase precision in both the design and analysis phases of a project).
- **However, the part with the *x*es has testable consequences.**

The Neyman-Rubin Model for (simple) experiments

This is what randomization ensures:

$$(y_t, y_c, X) \perp Z$$

I.e., each of X, y_c and y_t is balanced between treatment and control groups (in expectation; given variability from randomization).

- In controlled experiments, random assignment justifies this argument.
- In natural experiments, justified otherwise, this is an article of faith.
- In an experiment, the *x*es aren't necessary for inference (although they can be used, carefully, to increase precision in both the design and analysis phases of a project).
- **However, the part with the *x*es has testable consequences.**

The Neyman-Rubin Model for (simple) experiments

This is what randomization ensures:

$$(y_t, y_c, X) \perp Z$$

I.e., each of X, y_c and y_t is balanced between treatment and control groups (in expectation; given variability from randomization).

- In controlled experiments, random assignment justifies this argument.
- In natural experiments, justified otherwise, this is an article of faith.
- In an experiment, the *x*es aren't necessary for inference (although they can be used, carefully, to increase precision in both the design and analysis phases of a project).
- **However, the part with the *x*es has testable consequences.**

The Neyman-Rubin Model for (simple) experiments

This is what randomization ensures:

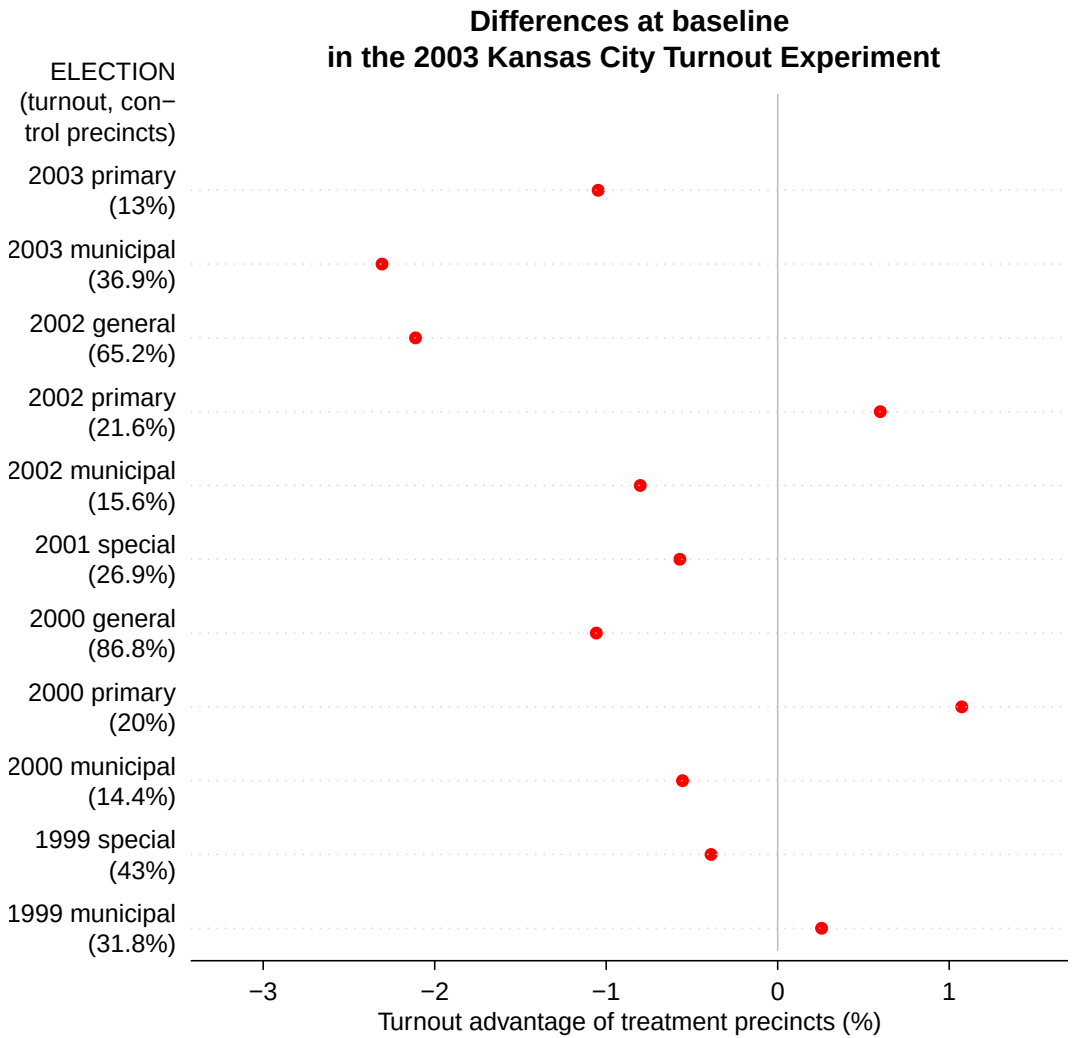
$$(y_t, y_c, X) \perp Z$$

I.e., each of X, y_c and y_t is balanced between treatment and control groups (in expectation; given variability from randomization).

- In controlled experiments, random assignment justifies this argument.
- In natural experiments, justified otherwise, this is an article of faith.
- In an experiment, the *x*es aren't necessary for inference (although they can be used, carefully, to increase precision in both the design and analysis phases of a project).
- **However, the part with the *x*es has testable consequences.**

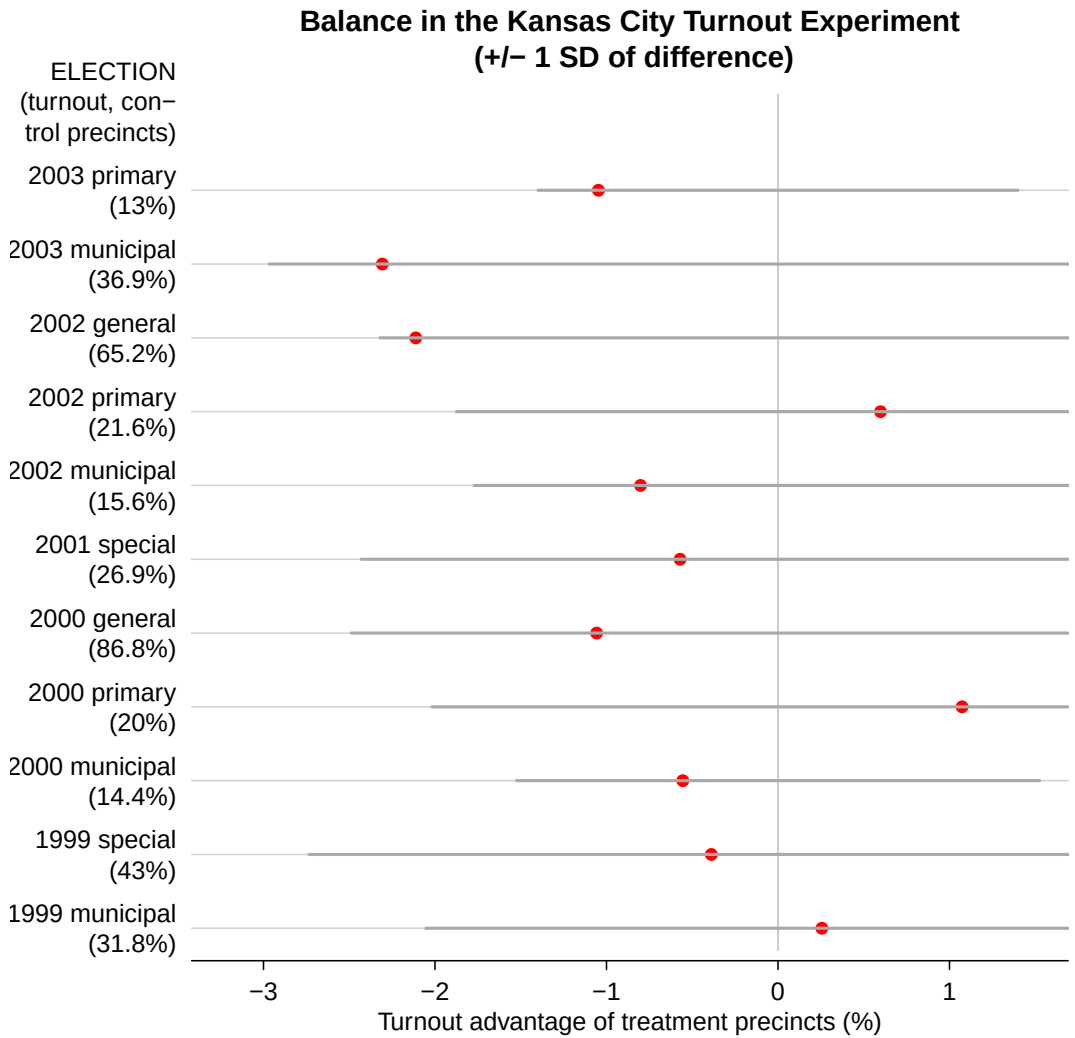
Covariate balance in experiments

- Arceneaux, 2005
- Kansas City, November 2003
- Completely randomized design: 14 precincts $\rightarrow$ Tx; 14 $\rightarrow$ Control.
- Substantively large baseline differences
- Differences not large compared to other possible assignments from same design; compared to other possible experiments with the same design.
- $\mathbf{Pr}(\chi^2 > x) = .91$ (Hansen and Bowers, 2008).



Covariate balance in experiments

- Arceneaux, 2005
- Kansas City, November 2003
- Completely randomized design: 14 precincts $\rightarrow$ Tx; 14 $\rightarrow$ Control.
- Substantively large baseline differences
- Differences not large compared to other possible assignments from same design; compared to other possible experiments with the same design.
- $\mathbf{Pr}(\chi^2 > x) = .91$ (Hansen and Bowers, 2008).



How did we do this?

```
acorn <- read.csv("data/acorn03.csv", row.names=1)
xb1 <- xBalance(z ~ v_p2003 + v_m2003 + v_g2002 + v_p2002 + v_m2002 + v_s2001 +
         v_g2000 + v_p2000 + v_m2000 + v_s1999 + v_m1999 + v_g1998 +
         v_m1998 + v_s1998 + v_m1997 + v_s1997 + v_g1996 + v_p1996 +
         v_m1996 + v_s1996 + size, data=acorn,
         report = 'all')

xb1$results
```

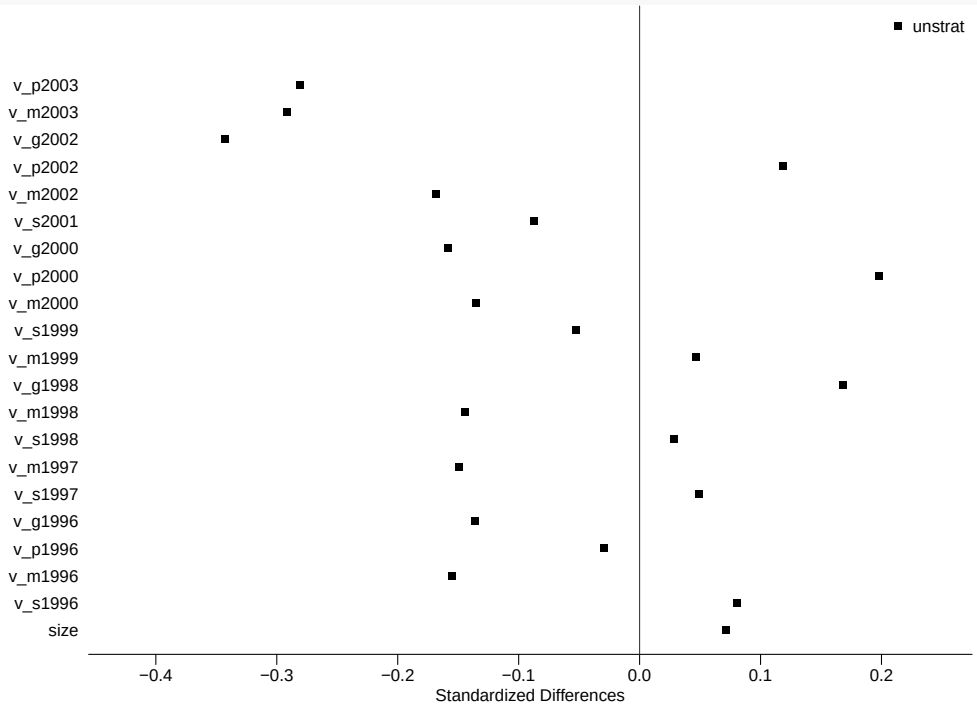
```
, , strata = unstrat

         stat
vars        Control Treatment  adj.diff adj.diff.null.sd std.diff        z      p
  v_p2003   0.12996   0.11949 -0.010469         0.013994 -0.28041 -0.74815 0.4544
  v_m2003   0.36868   0.34561 -0.023072         0.029669 -0.29171 -0.77763 0.4368
  v_g2002   0.65195   0.63083 -0.021119         0.023207 -0.34283 -0.91003 0.3628
  v_p2002   0.21589   0.22187  0.005980         0.018757  0.11848  0.31883 0.7499
  v_m2002   0.15567   0.14766 -0.008016         0.017721 -0.16841 -0.45233 0.6510
  v_s2001   0.26898   0.26327 -0.005709         0.024302 -0.08722 -0.23492 0.8143
  v_g2000   0.86812   0.85755 -0.010576         0.024878 -0.15821 -0.42512 0.6707
  v_p2000   0.20033   0.21106  0.010735         0.020187  0.19827  0.53177 0.5949
  v_m2000   0.14404   0.13849 -0.005552         0.015249 -0.13536 -0.36405 0.7158
  v_s1999   0.42957   0.42569 -0.003887         0.027343 -0.05275 -0.14216 0.8870
  v_m1999   0.31756   0.32012  0.002559         0.020532  0.04624  0.12463 0.9008
  v_g1998   0.43718   0.44900  0.011824         0.026134  0.16844  0.45243 0.6510
  v_m1998   0.18136   0.17315 -0.008210         0.021114 -0.14463 -0.38886 0.6974
  v_s1998   0.24104   0.24254  0.001496         0.019385  0.02863  0.07719 0.9385
  v_m1997   0.13388   0.12856 -0.005326         0.013300 -0.14898 -0.40047 0.6888
  v_s1997   0.15490   0.15659  0.001684         0.012635  0.04945  0.13327 0.8940
  v_g1996   0.59515   0.58192 -0.013232         0.036083 -0.13635 -0.36671 0.7138
```

How did we do this?

```
xb1$overall
```

```
         chisquare df p.value
unstrat      12.93 21   0.911
```

```
plot(xb1)
```



DeMystifying xBalance

```
d.stat<-function(zz, mm, ss){
  ## this is the d statistic (harmonic mean weighted diff of means statistic)
  ## from Hansen and Bowers 2008
  h.fn<-function(n, m){(m*(n-m))/n}
  myssn<-apply(mm, 2, function(x){sum((zz-unsplit(tapply(zz, ss, mean), ss))*x)})
  hs<-tapply(zz, ss, function(z){h.fn(m=sum(z), n=length(z))})
  mywtsum<-sum(hs)
  myadjdiff<-myssn/mywtsum
  return(myadjdiff)
}
```

Calculate the reference distribution of the d-stat and the d^2 stat

Does d^2 follow a χ^2 distribution in this case?

For all vectors $z \in \Omega$ get adj.diffs. This is the distribution of the d statistic

```
acorncovs<-c("v_p2003","v_m2003","v_g2002","v_p2002","v_m2002","v_s2001","v_g2000","v_p2000","v_
d.dist<-replicate(10000, d.stat(sample(acorn$z), acorn[,acorncovs], ss=rep(1,nrow(acorn))))
```

Get the randomization-based p-values:

```
xb1ds <- xb1$results[,"adj.diff",]
xb1ps <- xb1$results[,"p",]
obs.d<-d.stat(acorn$z, acorn[, acorncovs], rep(1,nrow(acorn)))
dps <- matrix(NA,nrow=length(obs.d),ncol=1)
for(i in 1:length(obs.d)){
  dps[i,] <- 2*min( mean(d.dist[i,] >= obs.d[i]),mean(d.dist[i,] <= obs.d[i]))
}
## You can compare this to the results from xBalance
round(cbind(randinfps=dps[,1],xbps=xb1ps,obsdstats=obs.d,xbdstats=xb1ds),3)
```

```
         randinfps  xbps obsdstats xbdstats
v_p2003      0.454 0.454    -0.010   -0.010
v_m2003      0.444 0.437    -0.023   -0.023
v_g2002      0.366 0.363    -0.021   -0.021
v_p2002      0.756 0.750     0.006    0.006
v_m2002      0.643 0.651    -0.008   -0.008
v_s2001      0.819 0.814    -0.006   -0.006
v_g2000      0.695 0.671    -0.011   -0.011
v_p2000      0.614 0.595     0.011    0.011
v_m2000      0.705 0.716    -0.006   -0.006
v_s1999      0.870 0.887    -0.004   -0.004
```

Calculate the reference distribution of the d-stat and the d^2 stat

The d^2 statistic is a linear function of the d-statistics that accounts for the covariance between those statistics (across the possible assignments under the null hypothesis of no effects).

```
d2.stat <- function(dstats,ddist=NULL,theinvcov=NULL){
  ## d is the vector of d statistics
  ## ddist is the matrix of the null reference distributions of the d statistics
  if(is.null(theinvcov) & !is.null(ddist)){
    as.numeric( t(dstats) %*% solve(cov(t(ddist))) %*% dstats)
  } else {
    as.numeric( t(dstats) %*% theinvcov %*% dstats)
  }
}
```

Calculate the reference distribution of the d-stat and the d^2 stat

The distribution of the d^2 statistic arises from the distribution of the d statistics — for each draw from the set of treatment assignments we can collapse the d-statistics into one d^2. And so we can calculate the p-value for the d^2.

```
## Here we have the inverse of the covariance/variance matrix of the d statistics
invCovDDist <- solve(cov(t(d.dist)))
obs.d2<- d2.stat(obs.d,d.dist,invCovDDist)

d2.dist<-apply(d.dist, 2, function(thed){
                  d2.stat(thed,theinvcov=invCovDDist)
          })
## The chi-squared reference distribution only uses a one-sided p-value going in the positive dir
d2p<-mean(d2.dist>=obs.d2)
cbind(obs.d2,d2p)
```

```
     obs.d2   d2p
[1,]  12.72 0.993
```

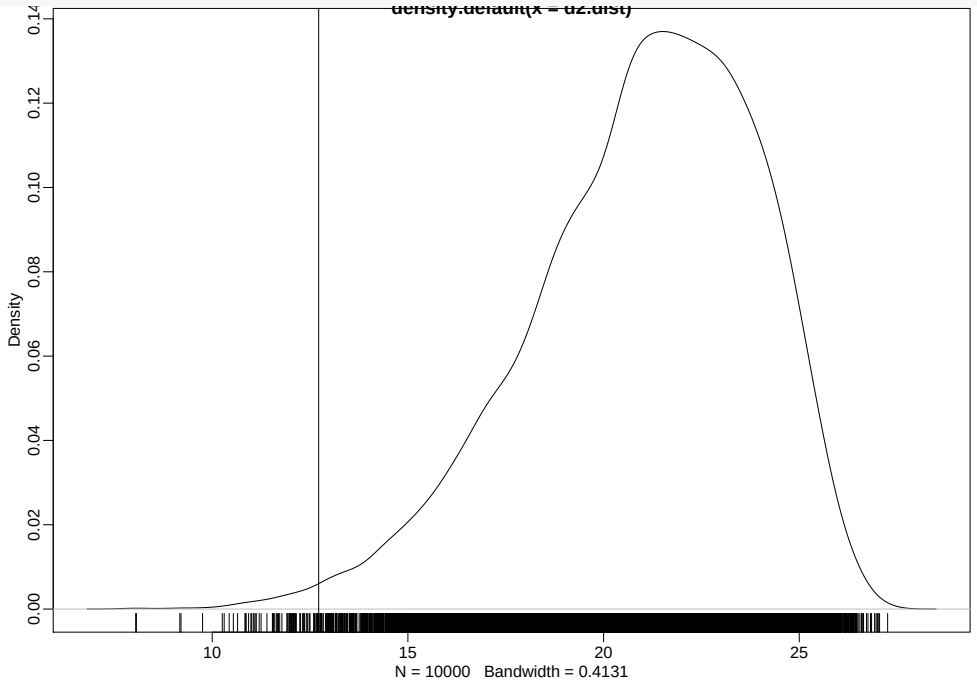
```
xb1$overall
```

```
        chisquare df p.value
unstrat     12.93 21   0.911
```

Calculate the reference distribution of the d-stat and the d^2 stat

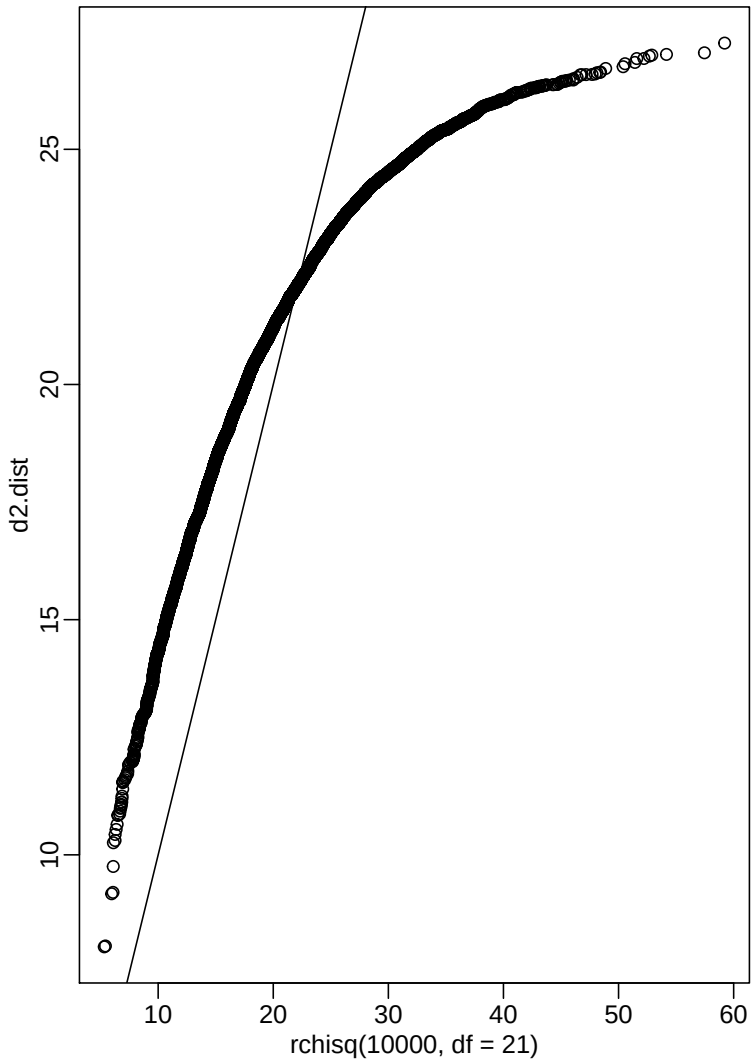
The distribution of the d^2 statistic arises from the distribution of the d statistics — for each draw from the set of treatment assignments we can collapse the d-statistics into one d^2. And so we can calculate the p-value for the d^2.

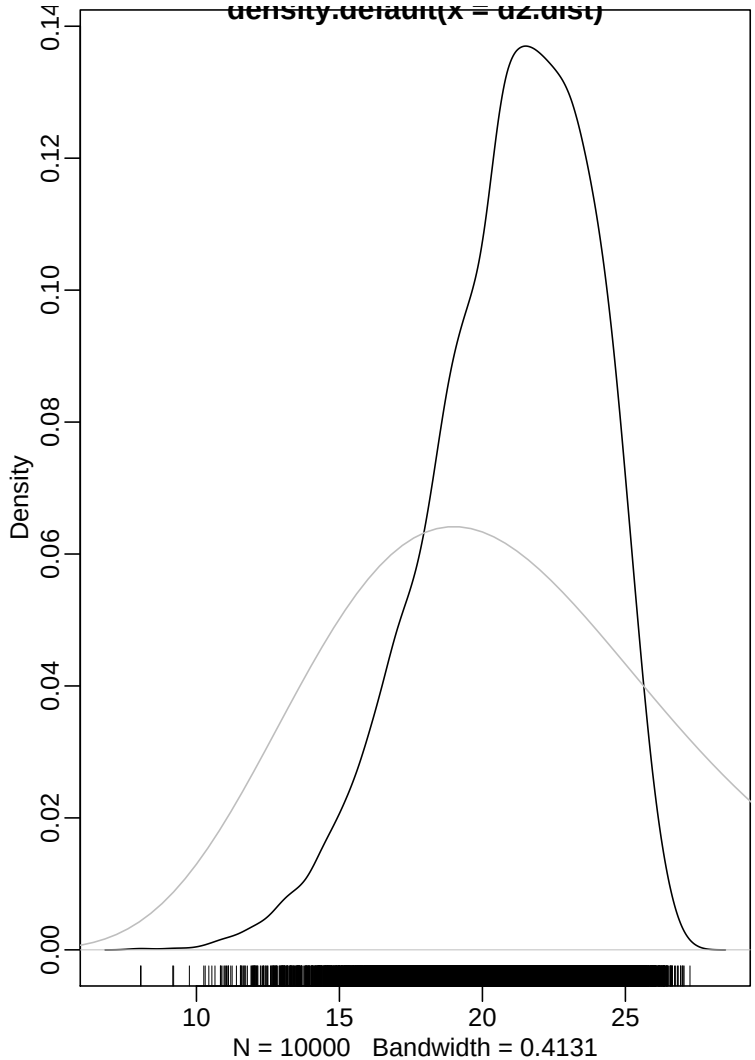
```
plot(density(d2.dist))
rug(d2.dist)
abline(v=obs.d2)
```



Why differences between xBalance and d2?

I suspect that $N = 28$ is too small. `xBalance` uses an asymptotic approximation to the randomization distribution.





Does xBalance have a controlled false positive rate here?

```
xbfn <- function(){
    acorn$newz <- sample(acorn$z)
    xb1 <- xBalance(newz ~ v_p2003 + v_m2003 + v_g2002 + v_p2002 + v_m2002 + v_s2001 +
            v_g2000 + v_p2000 + v_m2000 + v_s1999 + v_m1999 + v_g1998 +
            v_m1998 + v_s1998 + v_m1997 + v_s1997 + v_g1996 + v_p1996 +
            v_m1996 + v_s1996 + size, data=acorn,
        report = 'chisquare.test')
    return(xb1$overall[["p.value"]])
}

res <- replicate(1000,xbfn())
```

```
summary(res)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.184   0.333   0.443   0.472   0.599   0.962
```

```
mean(res <= .05)
```

```
[1] 0
```

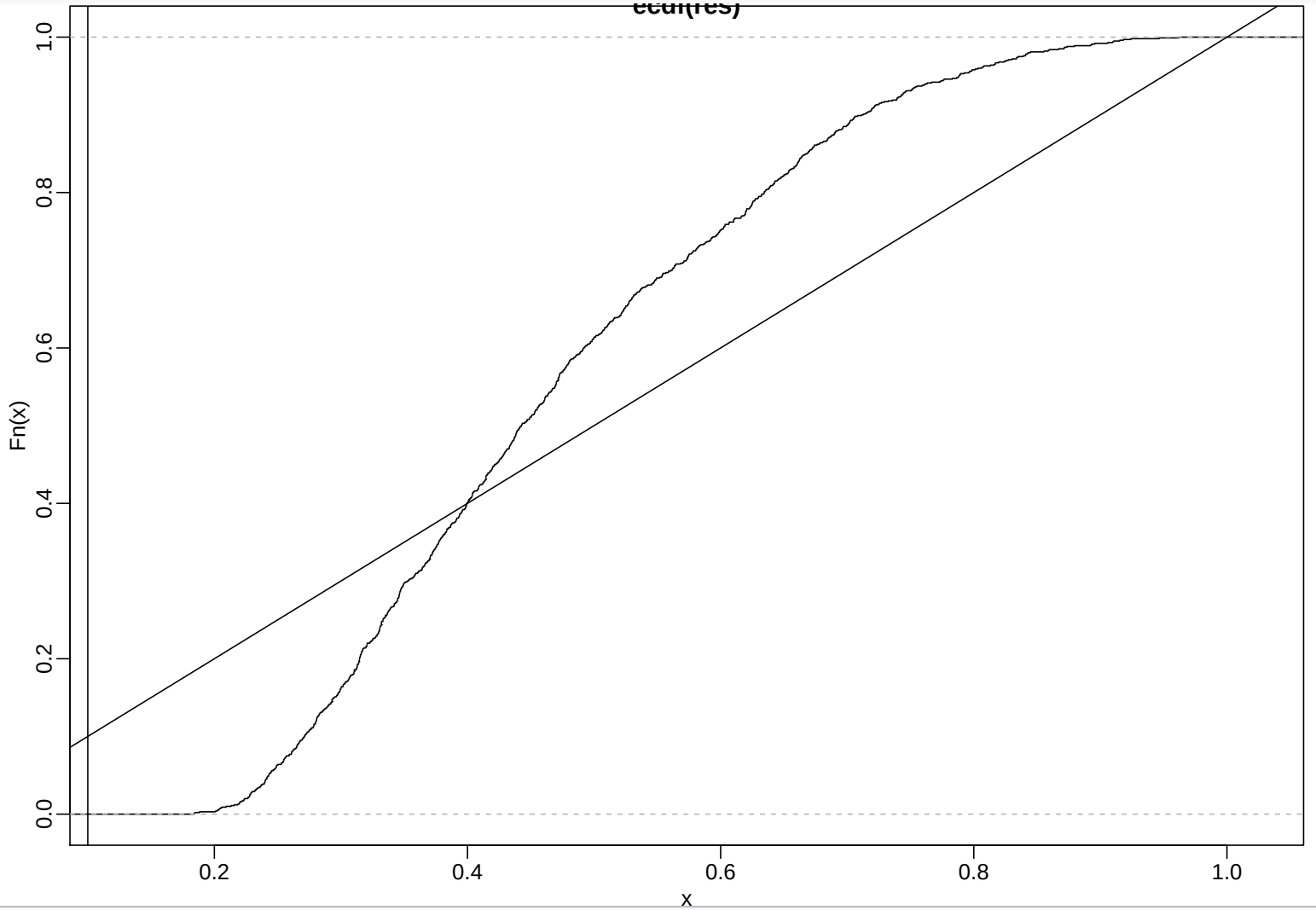
```
mean(res <= .2)
```

```
[1] 0.003
```

Does xBalance have a controlled false positive rate here?

Ex. are fewer than 5% of the p-values less than .05?

```
plot(ecdf(res))
abline(0,1)
abline(v=c(.01,.05,.1))
```



Does the simulation based approach have a controlled false positive rate here?

```
d2pfn <- function(z,X){
    newz <- sample(z)

    d.dist<-replicate(1000, d.stat(sample(newz), X, ss=rep(1,nrow(X))))

    obs.d<-d.stat(newz, X, rep(1,nrow(X)))

    dps <- matrix(NA,nrow=length(obs.d),ncol=1)
    for(i in 1:length(obs.d)){
        dps[i,] <- 2*min( mean(d.dist[i,] >= obs.d[i]),mean(d.dist[i,] <= obs.d[i]))
    }

    invCovDDist <- solve(cov(t(d.dist)))
    obs.d2<- d2.stat(obs.d,d.dist,invCovDDist)

    d2.dist<-apply(d.dist, 2, function(thed){
                    d2.stat(thed,theinvcov=invCovDDist)
        })

    d2p<-mean(d2.dist>=obs.d2)

    return(d2p)
}
```

```
resdirect <- replicate(1000,d2pfn(z=acorn$z,X=acorn[,acorncovs]))
```

```
library(parallel)
resdirectlst <- mclapply(1:1000,function(i){ d2pfn(z=acorn$z,X=acorn[,acorncovs]) },mc.cores=det
```

Does the simulation based approach have a controlled false positive rate here?

```
##lazyLoad("day9-AdjustmentBalance_cache/beamer/doresdirectparallel_5ec4fa8cdcbcf586138e928bc0f9f
##load("day9-resdirect.rda")
summary(resdirect)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.200   0.469   0.471   0.736   1.000
```

```
mean(resdirect <= .05)
```

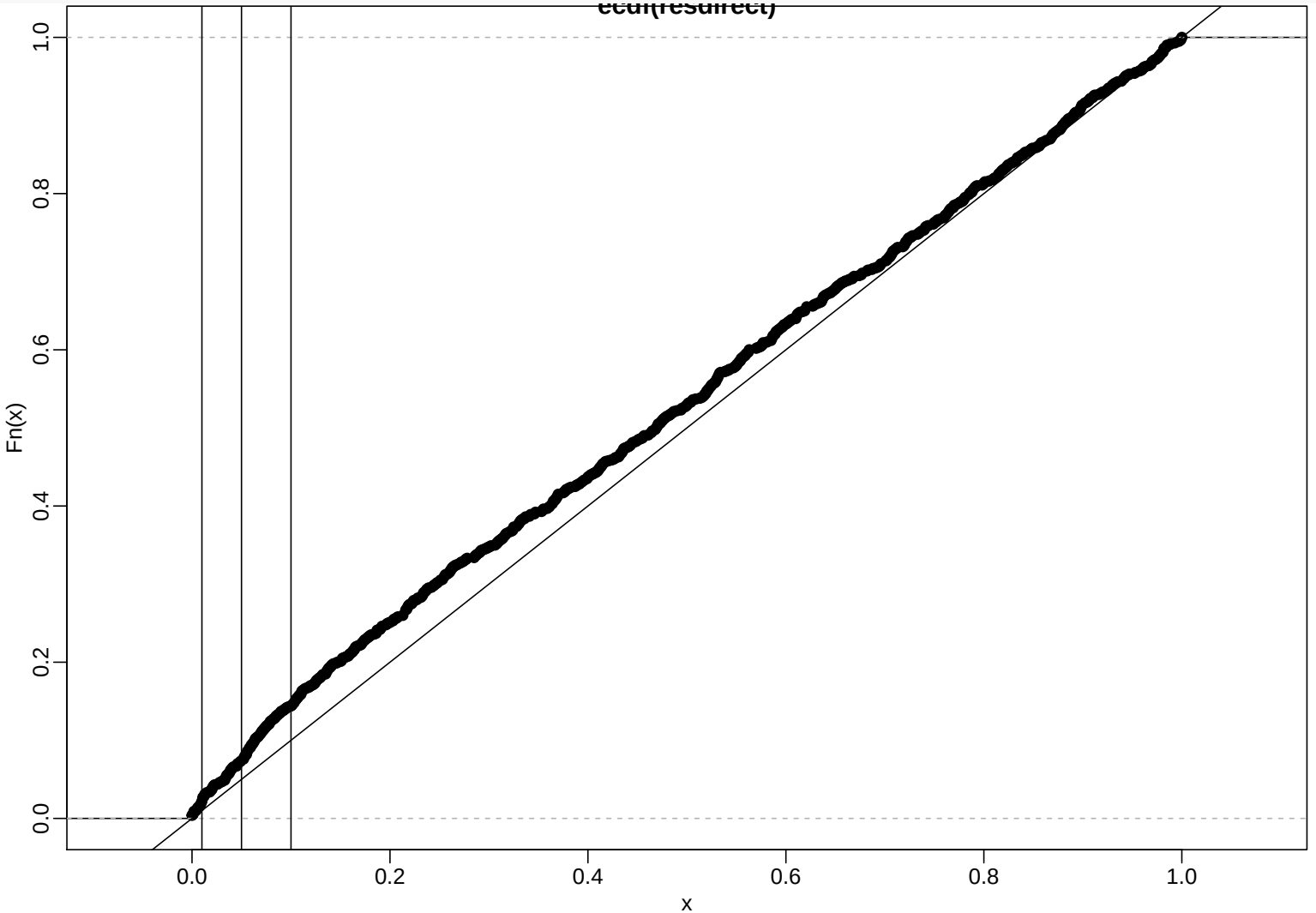
```
[1] 0.073
```

```
mean(resdirect <= .2)
```

```
[1] 0.251
```

Does xBalance have a controlled false positive rate here?

```
plot(ecdf(resdirect))
abline(0,1)
abline(v=c(.01,.05,.1))
```



1. Strategies for Causal Inference
2. But first, how to assess the randomization process in an experiment.
3. Did we control for enough?

Did we control for enough?

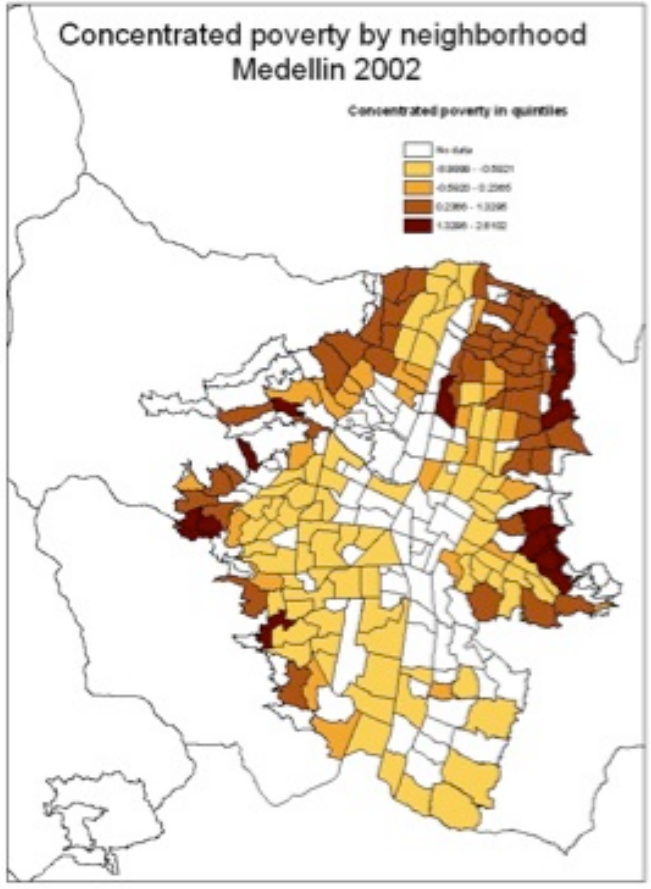
```
load(url("http://jakebowers.org/Data/meddat.rda"))
```

Introducing the Medellin Data

Cerdá et al. collected data on about roughly 45 neighborhoods in Medellin, Colombia. About 22 of had access to the new Metrocable line and 23 did not.

Introducing the Medellin Data

Cerdá et al. collected data on about roughly 45 neighborhoods in Medellin, Colombia. About 22 of had access to the new Metrocable line and 23 did not.



Introducing the Medellin Data: Variables Collected

```
## The Intervention
nhTrt         Intervention neighborhood (0=no Metrocable station, 1=Metrocable station)

## Some Covariates (there are others, see the paper itself)
nh03          Neighborhood id
nhGroup       Treatment (T) or Control (C)
nhTrt         Treatment (1) or Control (0)
nhHom         Mean homicide rate per 100,000 population in 2003
nhDistCenter Distance to city center (km)
nhLogHom      Log Homicide (i.e. log(nhHom))

## Outcomes (BE03,CE03,PV03,QP03,TP03 are baseline versions)
BE      Neighborhood amenities Score 2008
CE      Collective Efficacy Score 2008
PV      Perceived Violence Score 2008
QP      Trust in local agencies Score 2008
TP      Reliance on police Score 2008
hom     Homicide rate per 100,000 population Score 2008-2003 (in log odds)

HomCount2003 Number of homicides in 2003
Pop2003      Population in 2003
HomCount2008 Number of homicides in 2008
Pop2008      Population in 2008
```

Get rates from counts:

```
meddat<- mutate(meddat, HomRate03=(HomCount2003/Pop2003)*1000,
                HomRate08=(HomCount2008/Pop2008)*1000)
```

What is the effect of the Metrocable on Homicides?

One approach: Estimate the average treatment effect of Metrocable on Homicides after the stations were built.

```
## code here
themeans<-group_by(meddat,nhTrt) %>% summarise(ybar=mean(HomRate08))
diff(themeans$ybar)
```

```
[1] -0.2899
```

```
lmOne <- lm(HomRate08~nhTrt,meddat)
coef(lmOne)["nhTrt"]
```

```
  nhTrt
-0.2899
```

```
library(estimatr)
difference_in_means(HomRate08~nhTrt,meddat)
```

```
Warning: Assigning non-quosure objects to quosure lists is deprecated as of rlang 0.3.0.
Please coerce to a bare list beforehand with `as.list()`
This warning is displayed once per session.
Design:  Standard
      Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
nhTrt  -0.2899     0.1508  -1.922  0.06137  -0.5942  0.01445 41.87
```

Another approach, test the null of absolutely no effects:

```
xBalance(nhTrt~HomRate08,report="all",data=meddat)
```

```
          strata():   unstrat
          stat      Treatment Control adj.diff std.diff      z      p
vars
HomRate08               0.400   0.690   -0.290   -0.571 -1.858  0.063
---Overall Test---
        chisquare df p.value
```

What are alternative explanations for this effect?

We claim that the policy intervention had some effect. What are alternative explanations?

Do we have any concerns about confounding?

Sometimes people ask about "bias from observed confounding" or "bias from selection on observables".
How would we interpret the following results? (Recall how we justified the use of `xBalance` in terms of randomization above.)

```
xbMed1 <- xBalance(nhTrt~nhAboveHS,data=meddat,report="all")
xbMed1$overall
```

```
        chisquare df p.value
unstrat     4.598  1 0.03201
```

```
xbMed1$results
```

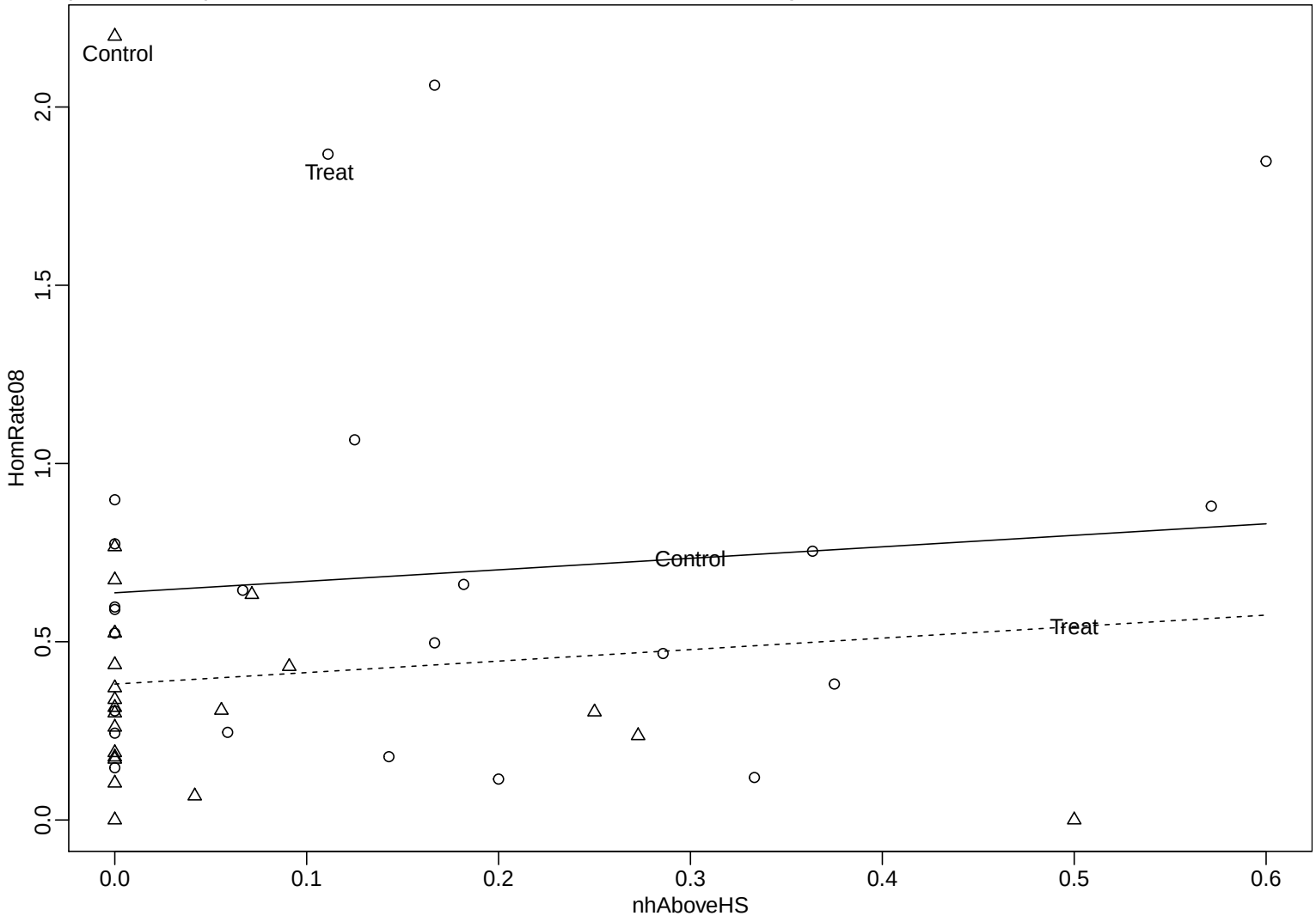
```
, , strata = unstrat

           stat
vars        Control Treatment adj.diff adj.diff.null.sd std.diff      z       p
  nhAboveHS   0.163   0.05829  -0.1047          0.04883   -0.668 -2.144 0.03201

attr(,"originals")
[1] "nhAboveHS"
```

How would you adjust for Proportion Above HS Degree?

Part of the Metrocable effect is not about Metrocable per se, but rather about the education of people in the neighborhood. How should we remove `nhAboveHS` from our estimate or test? What strategies can you think of?

One approach to this problem: model-based adjustment

Let's try to just adjust for this covariate in a very common manner:



Exactly what does this kind of adjustment do?

Notice that I can get the same coefficient (the effect of Metrocable on Homicides adjusted for HS-Education in the neighborhood) either directly (as earlier) or via **residualization**:

```
coef(lm1)["nhTrt"]
```

```
  nhTrt
-0.2561
```

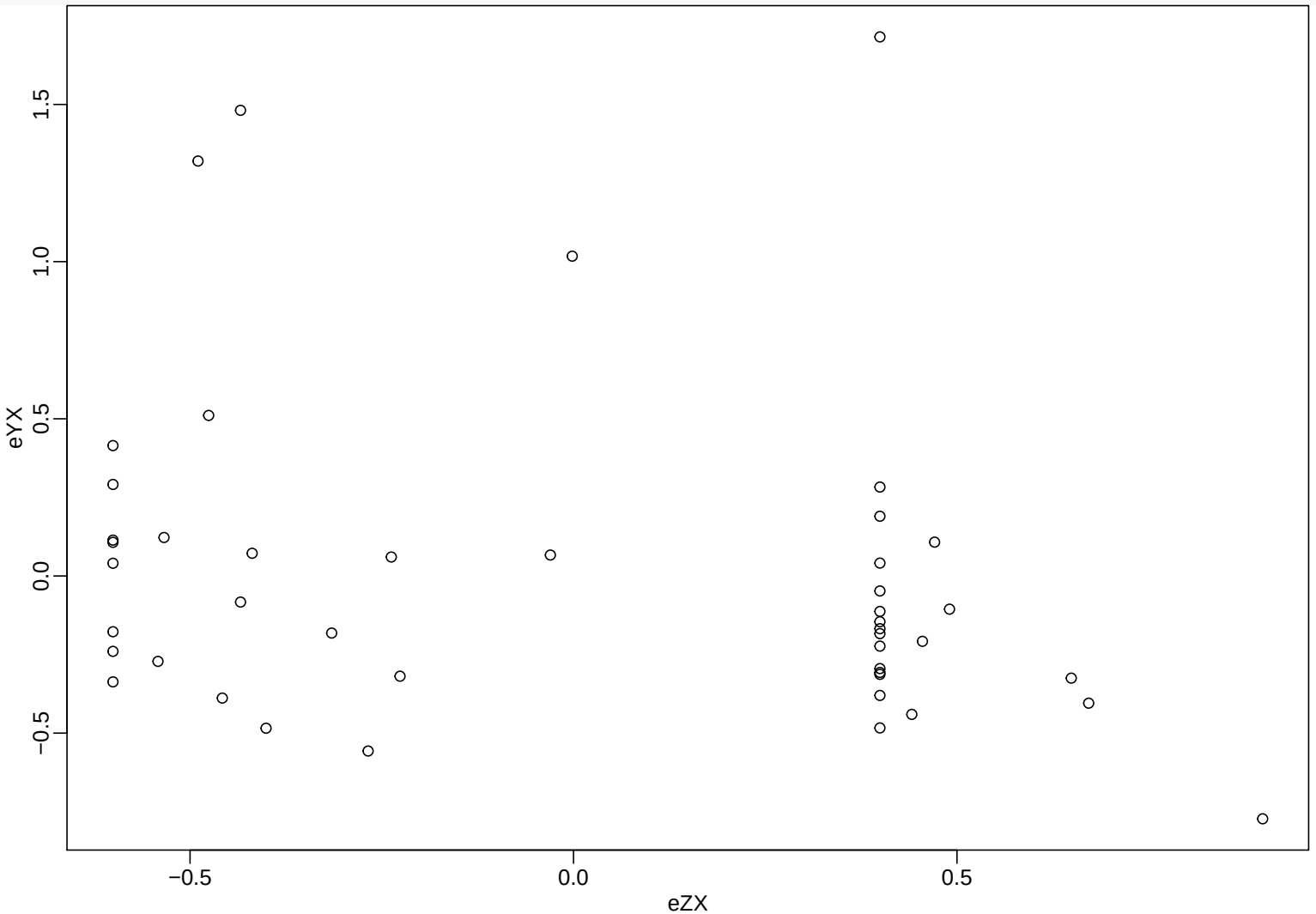
```
eYX <- residuals(lm(HomRate08~nhAboveHS,data=meddat))
eZX <- residuals(lm(nhTrt ~ nhAboveHS, data=meddat))
lm1a <- lm(eYX~eZX)
coef(lm1a)[2]
```

```
    eZX
-0.2561
```

Exactly what does this kind of adjustment do?

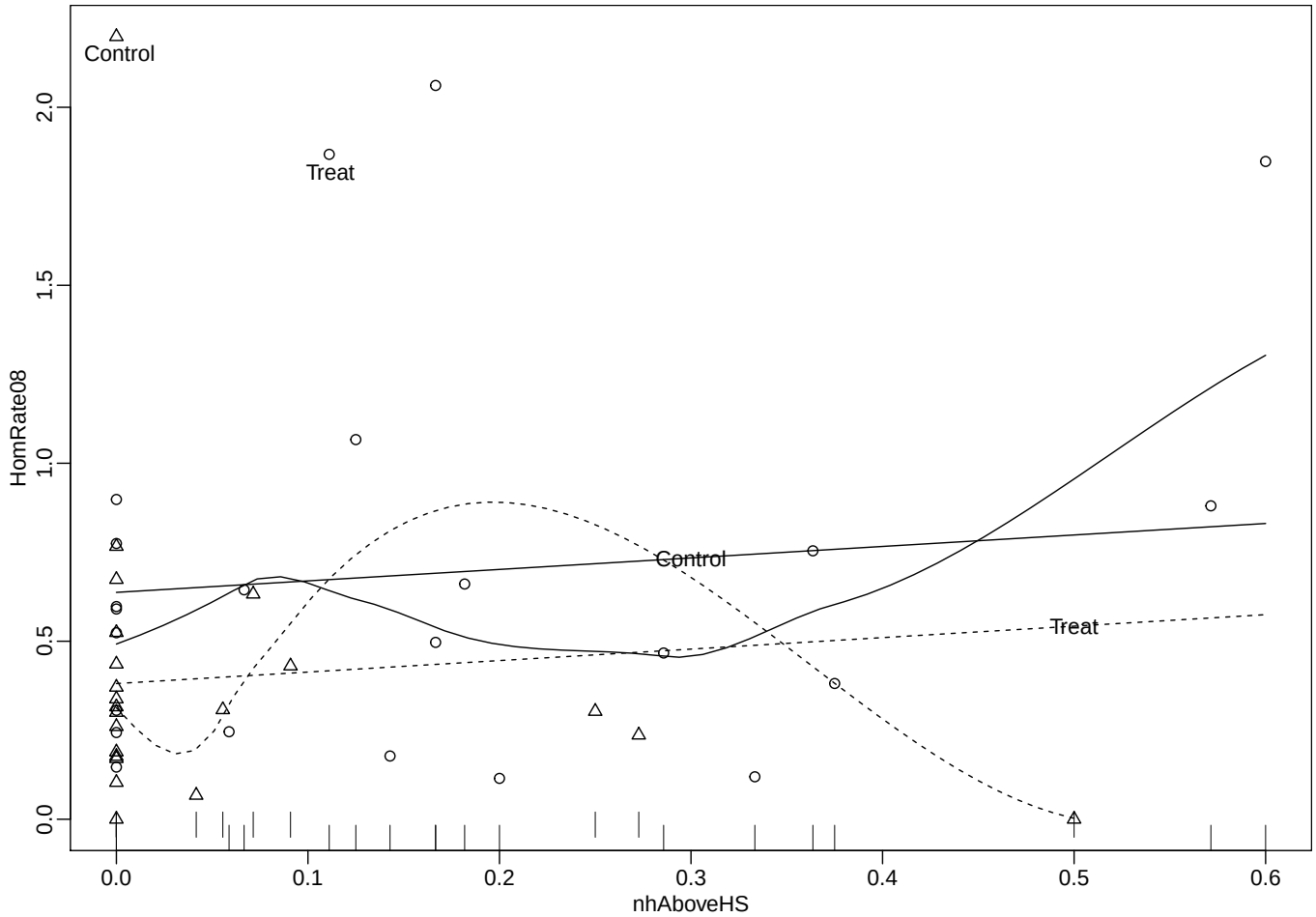
So, how would you explain what it means to "control for HS-Education" here?

```
plot(eZX,eYX)
```



Did we adjust enough?

Maybe adding some more information to the plot can help us decide whether, and to what extend, we effectively "controlled for" the proportion of the neighborhood with more than High School education. Specifically, we might be interested in assessing extrapolation/interpolation problems arising from our linear assumptions.



How should we interpret this adjustment? How should we judge the improvement

How would you adjust for Proportion Above HS Degree?

So, part of the Metrocable effect might not reflect the causal effect of Metrocable per se, but rather the education of people in the neighborhood. How should we remove `nhAboveHS` from our estimate or test? What strategies can you think of? Features of a good adjustment process:

- Blind to outcome analysis (to preserve false positive rate and deter critics). Could be pre-registered.
- Easy to interpret ("controlling for" versus "holding constant")
- Easy to diagnoses (Easy to answere the question "Did we adjust enough?")

How would you adjust for Proportion Above HS Degree?

So, part of the Metrocable effect might not reflect the causal effect of Metrocable per se, but rather the education of people in the neighborhood. How should we remove `nhAboveHS` from our estimate or test? What strategies can you think of? Features of a good adjustment process:

- Blind to outcome analysis (to preserve false positive rate and deter critics). Could be pre-registered.
- Easy to interpret ("controlling for" versus "holding constant")
- Easy to diagnoses (Easy to answere the question "Did we adjust enough?")

How would you adjust for Proportion Above HS Degree?

So, part of the Metrocable effect might not reflect the causal effect of Metrocable per se, but rather the education of people in the neighborhood. How should we remove `nhAboveHS` from our estimate or test? What strategies can you think of? Features of a good adjustment process:

- Blind to outcome analysis (to preserve false positive rate and deter critics). Could be pre-registered.
- Easy to interpret ("controlling for" versus "holding constant")
- Easy to diagnoses (Easy to answere the question "Did we adjust enough?")

Stratification V 1.0

```
lm1a <- lm(HomRate08~nhTrt,data=meddat,subset=nhAboveHS>=.1)
lm1b <- lm(HomRate08~nhTrt,data=meddat,subset=nhAboveHS<.1)
res_strat <- c(hiEd_Effect=coef(lm1a)["nhTrt"],loEd_Effect= coef(lm1b)["nhTrt"])
res_strat
```

```
hiEd_Effect.nhTrt loEd_Effect.nhTrt
         -0.65828          -0.06237
```

```
n_strat <- table(meddat$nhAboveHS>=.1)
n_strat
```

```
FALSE  TRUE
   29    16
```

```
stopifnot(sum(n_strat)==nrow(meddat)) ## A test of code
sum(res_strat * rev(n_strat)/45) ## What is happening here?
```

```
[1] -0.2743
```

But, standard errors? p-values? confidence intervals?

Stratified adjustment V 2.0

One-step stratified estimation.

```
## Weight by block size
ate1c <- difference_in_means(HomRate08~nhTrt, blocks = I(nhAboveHS>=.1),data=meddat)
ate1c
```

```
Design:  Blocked
      Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
nhTrt  -0.2743     0.1146  -2.393   0.0214  -0.5057 -0.04276 41
```

```
## Weight by both block size and  proportion in treatment vs control ("harmonic weight")
lm1c <- lm_robust(HomRate08~nhTrt, fixed_effects = ~I(nhAboveHS>=.1),data=meddat)
coef(lm1c)["nhTrt"]
```

```
 nhTrt
-0.224
```

```
lm1d <- lm(HomRate08~nhTrt+I(nhAboveHS>=.1),data=meddat)
coef(lm1d)["nhTrt"]
```

```
 nhTrt
-0.224
```

```
xbate1 <- xBalance(nhTrt~HomRate08,strata=list(hs=~I(nhAboveHS>=.1)),data=meddat,report="all")
xbate1$results[1,c("Control","Treatment","adj.diff"),]
```

```
  Control Treatment  adj.diff
   0.5896    0.3656   -0.2240
```

Balance assessment after stratification

Did we adjust enough? What would enough mean?

```
xbHS1 <- xBalance(nhTrt~nhAboveHS,strata=list(hs=~I(nhAboveHS>=.1)),data=meddat,report="all")
xbHS1$overall
```

```
   chisquare df p.value
hs    0.3912  1  0.5317
```

```
xbHS1$results[1,c("Control","Treatment","adj.diff","std.diff","z","p"),]  ## the covariate speci
```

```
  Control Treatment  adj.diff  std.diff         z         p
  0.08472   0.10240   0.01768   0.11278   0.62547   0.53166
```

Disadvantages and Advantages of Simple Stratification

- (+) Easy to explain what "controlling for" or "adjustment" means.
- (-) Hard to justify any particular cut-point
- (-) We could probably adjust more — comparing neighborhoods similar in education rather than just within big strata

Disadvantages and Advantages of Simple Stratification

- (+) Easy to explain what "controlling for" or "adjustment" means.
- (-) Hard to justify any particular cut-point
- (-) We could probably adjust more — comparing neighborhoods similar in education rather than just within big strata

Disadvantages and Advantages of Simple Stratification

- (+) Easy to explain what "controlling for" or "adjustment" means.
- (-) Hard to justify any particular cut-point
- (-) We could probably adjust more — comparing neighborhoods similar in education rather than just within big strata

The Curse of Dimensionality and linear adjustment for one more variable.

What about more than one variable? Have we controlled for both population density and educational attainment enough? How would we know?

```
lm2x <- lm(HomRate08 ~ nhTrt + nhPopD + nhAboveHS, data=meddat)
coef(lm2x)["nhTrt"]
```

```
  nhTrt
-0.1224
```

Maybe another plot?

```
meddat$nhTrtF <- factor(meddat$nhTrt)
library(car)
scatter3d(HomRate08~nhAboveHS+nhPopD,
      groups=meddat$nhTrtF,
      data=meddat,surface=TRUE,
    fit=c("linear")) #additive"))
```

The Problem of Using the Linear Model for Adjustment

- Problem of Interepretability: "Controlling for" is "removing (additive) linear relationships" it is not "holding constant"
- Problem of Diagnosis and Assessment: What is the standard against which we can compare a given linear covariance adjustment specification?
- Problem of extrapolation and interpolation: Often known as "common support", too.
- Problems of overly influential points and curse of dimensionality: As dimensions increase, odds of influential point increase (ex. bell curve in one dimension, one very influential point in 2 dimensions); also real limits on number of covariates (roughly $\sqrt{n}$ for OLS).
- Problems of bias:

$$Y_i = \beta_0 + \beta_1 Z_i + e_i (\#eq : olsbiv) \tag{1}$$

This is a common practice because, we know that the formula to estimate β_1 in equation @ref(eq:olsbiv) is the same as the difference of means in Y between treatment and control groups:

$$\hat{\beta}_1 = \overline{Y|Z=1} - \overline{Y|Z=0} = \frac{cov(Y,Z)}{var(Z)}. \tag{2}$$

The Problem of Using the Linear Model for Adjustment

- Problem of Interepretability: "Controlling for" is "removing (additive) linear relationships" it is not "holding constant"
- Problem of Diagnosis and Assessment: What is the standard against which we can compare a given linear covariance adjustment specification?
- Problem of extrapolation and interpolation: Often known as "common support", too.
- Problems of overly influential points and curse of dimensionality: As dimensions increase, odds of influential point increase (ex. bell curve in one dimension, one very influential point in 2 dimensions); also real limits on number of covariates (roughly $\sqrt{n}$ for OLS).
- Problems of bias:

$$Y_i = \beta_0 + \beta_1 Z_i + e_i (\#eq : olsbiv) \tag{1}$$

This is a common practice because, we know that the formula to estimate β_1 in equation @ref(eq:olsbiv) is the same as the difference of means in Y between treatment and control groups:

$$\hat{\beta}_1 = \overline{Y|Z=1} - \overline{Y|Z=0} = \frac{cov(Y,Z)}{var(Z)}. \tag{2}$$

The Problem of Using the Linear Model for Adjustment

- Problem of Interepretability: "Controlling for" is "removing (additive) linear relationships" it is not "holding constant"
- Problem of Diagnosis and Assessment: What is the standard against which we can compare a given linear covariance adjustment specification?
- Problem of extrapolation and interpolation: Often known as "common support", too.
- Problems of overly influential points and curse of dimensionality: As dimensions increase, odds of influential point increase (ex. bell curve in one dimension, one very influential point in 2 dimensions); also real limits on number of covariates (roughly $\sqrt{n}$ for OLS).
- Problems of bias:

$$Y_i = \beta_0 + \beta_1 Z_i + e_i (\#eq : olsbiv) \tag{1}$$

This is a common practice because, we know that the formula to estimate β_1 in equation @ref(eq:olsbiv) is the same as the difference of means in Y between treatment and control groups:

$$\hat{\beta}_1 = \overline{Y|Z=1} - \overline{Y|Z=0} = \frac{cov(Y,Z)}{var(Z)}. \tag{2}$$

The Problem of Using the Linear Model for Adjustment

- Problem of Interepretability: "Controlling for" is "removing (additive) linear relationships" it is not "holding constant"
- Problem of Diagnosis and Assessment: What is the standard against which we can compare a given linear covariance adjustment specification?
- Problem of extrapolation and interpolation: Often known as "common support", too.
- Problems of overly influential points and curse of dimensionality: As dimensions increase, odds of influential point increase (ex. bell curve in one dimension, one very influential point in 2 dimensions); also real limits on number of covariates (roughly $\sqrt{n}$ for OLS).
- Problems of bias:

$$Y_i = \beta_0 + \beta_1 Z_i + e_i (\#eq : olsbiv) \tag{1}$$

This is a common practice because, we know that the formula to estimate β_1 in equation @ref(eq:olsbiv) is the same as the difference of means in Y between treatment and control groups:

$$\hat{\beta}_1 = \overline{Y|Z=1} - \overline{Y|Z=0} = \frac{cov(Y,Z)}{var(Z)}. \tag{2}$$

The Problem of Using the Linear Model for Adjustment

- Problem of Interepretability: "Controlling for" is "removing (additive) linear relationships" it is not "holding constant"
- Problem of Diagnosis and Assessment: What is the standard against which we can compare a given linear covariance adjustment specification?
- Problem of extrapolation and interpolation: Often known as "common support", too.
- Problems of overly influential points and curse of dimensionality: As dimensions increase, odds of influential point increase (ex. bell curve in one dimension, one very influential point in 2 dimensions); also real limits on number of covariates (roughly $\sqrt{n}$ for OLS).
- Problems of bias:

$$Y_i = \beta_0 + \beta_1 Z_i + e_i (\#eq : olsbiv) \tag{1}$$

This is a common practice because, we know that the formula to estimate β_1 in equation @ref(eq:olsbiv) is the same as the difference of means in Y between treatment and control groups:

$$\hat{\beta}_1 = \overline{Y|Z=1} - \overline{Y|Z=0} = \frac{cov(Y,Z)}{var(Z)}. \tag{2}$$

Can we improve stratified adjustment?

Rather than two strata, why not three?

```
lm1cut3 <- lm(HomRate08~nhTrt+cut(nhAboveHS,3),data=meddat)
coef(lm1cut3)["nhTrt"]
```

```
  nhTrt
-0.3161
```

But why those cuts? And why not 4? Why not...?

One idea: collect observations into strata such that the sum of the differences in means of nhAboveHS within strata is smallest? This is the idea behind `optmatch` and other matching approaches.

The optmatch workflow: The distance matrix

Introduction to optmatch workflow. To minimize differences requires a matrix of those differences (in general terms, a matrix of distances between the treated and control units)

```
tmp <- meddat$nhAboveHS
names(tmp) <- rownames(meddat)
absdist <- match_on(tmp, z = meddat$nhTrt,data=meddat)
absdist[1:3,1:3]
```

```
         control
treatment     23      24     25
        1 0.1429 0.00000 0.1667
        2 0.1429 0.00000 0.1667
        3 0.0873 0.05556 0.1111
```

```
abs(meddat$nhAboveHS[meddat$nhTrt==1][1] - meddat$nhAboveHS[meddat$nhTrt==0][1] )
```

```
[1] 0.1429
```

Created a Stratified Research Design

```
fm1 <- fullmatch(absdist,data=meddat)
summary(fm1, min.controls=0, max.controls=Inf )
```

```
Structure of matched sets:
7:1 2:1 1:1 1:2 1:4 1:5
  1   2   7   2   1   1
Effective Sample Size:  17.4
(equivalent number of matched pairs).
```

```
table(meddat$nhTrt,fm1)
```

```
   fm1
    1.1 1.10 1.15 1.16 1.17 1.19 1.2 1.20 1.3 1.4 1.5 1.6 1.7 1.9
  0   1    1    5    1    1    4   1    1   1   1   1   2   2   1
  1   2    1    1    2    7    1   1    1   1   1   1   1   1   1
```

```
pm1 <- pairmatch(absdist,data=meddat)
summary(pm1, min.controls=0, max.controls=Inf )
```

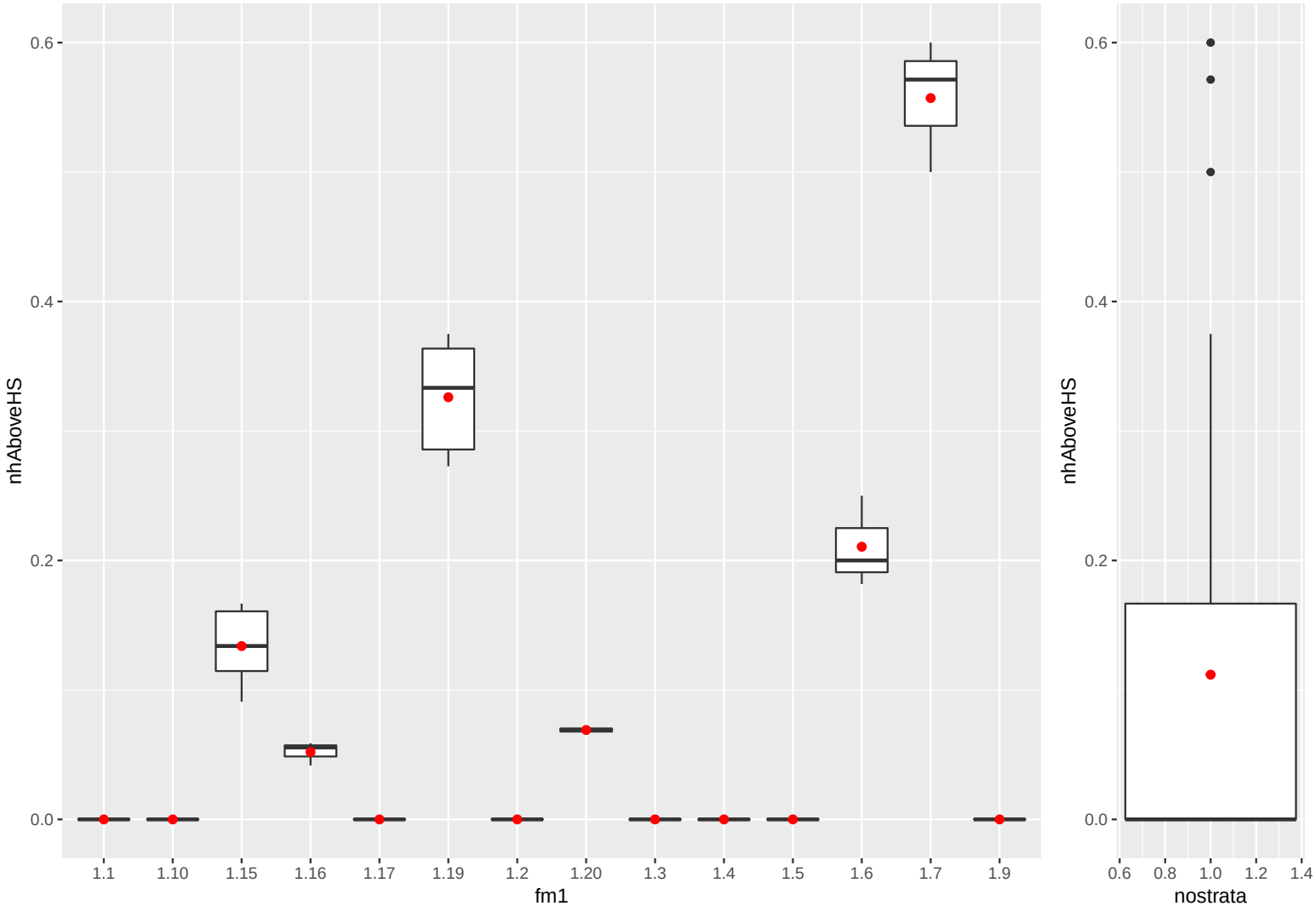
```
Structure of matched sets:
1:1 0:1
 22   1
Effective Sample Size:  22
(equivalent number of matched pairs).
```

```
table(meddat$nhTrt,pm1,exclude=c())
```

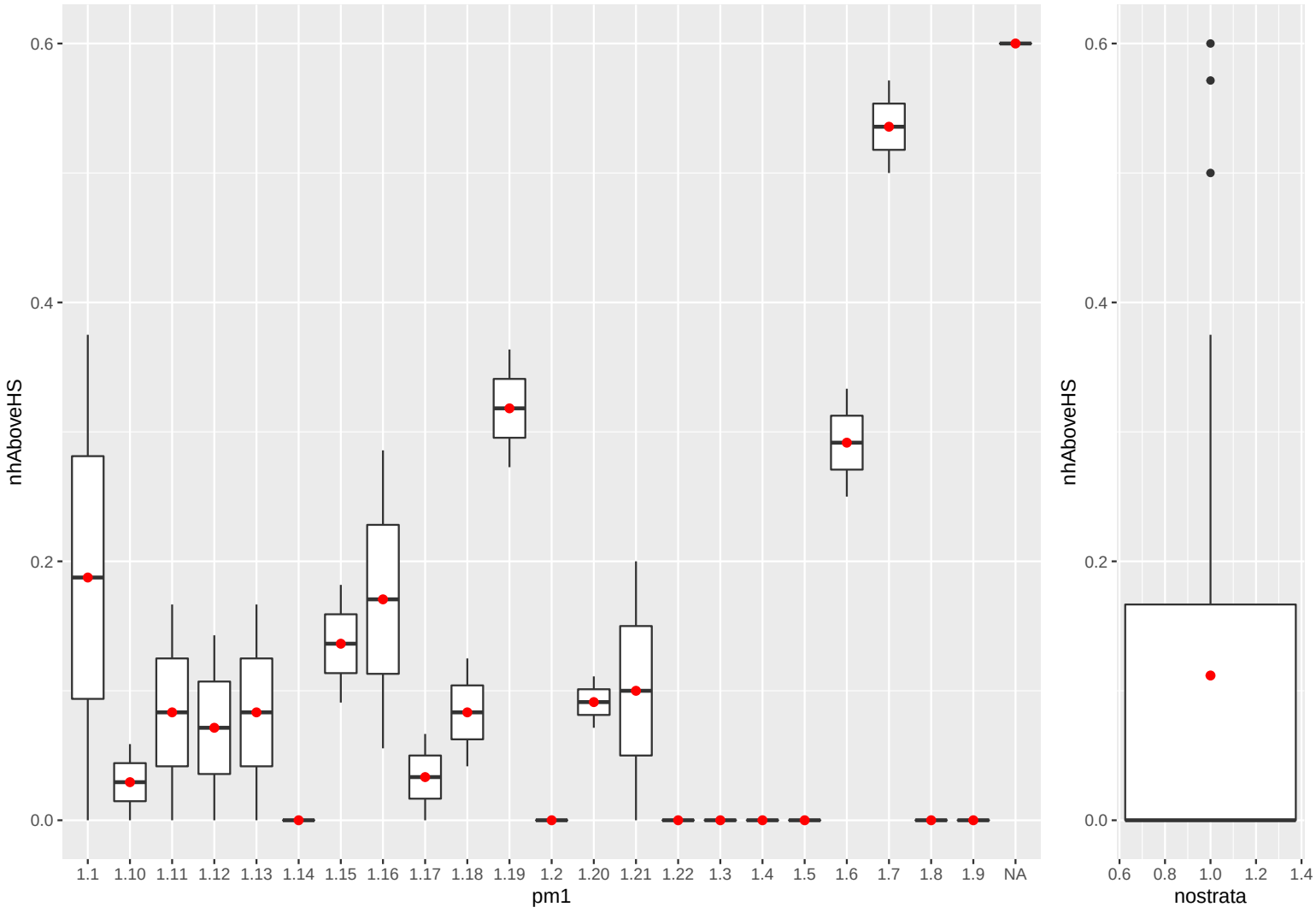
```
   pm1
    1.1 1.10 1.11 1.12 1.13 1.14 1.15 1.16 1.17 1.18 1.19 1.2 1.20 1.21 1.22 1.3 1.4 1.5 1.6 1.7
  0   1    1    1    1    1    1    1    1    1    1    1   1    1    1    1   1   1   1   1   1
  1   1    1    1    1    1    1    1    1    1    1    1   1    1    1    1   1   1   1   1   1
```

Evaluate the design: Within set differences

Differences within sets versus raw differences.



Evaluate the design: Within set differences



Evaluate the design: Inspecet within set differences

```
# A tibble: 14 x 5
   fm1   mneddiffs mnAboveHS minAboveHS maxAboveHS
   <fct>     <dbl>     <dbl>      <dbl>      <dbl>
 1 1.1     0          0          0          0
 2 1.10    0          0          0          0
 3 1.15   -0.0516     0.134      0.0909     0.167
 4 1.16   -0.0102     0.0520     0.0417     0.0588
 5 1.17    0          0          0          0
 6 1.19   -0.0667     0.326      0.273      0.375
 7 1.2     0          0          0          0
 8 1.20    0.00476    0.0690     0.0667     0.0714
 9 1.3     0          0          0          0
10 1.4     0          0          0          0
11 1.5     0          0          0          0
12 1.6     0.0591     0.211      0.182      0.25
13 1.7    -0.0857     0.557      0.5        0.6
14 1.9     0          0          0          0
```

Evaluate the design: Inspect within set differences

```
Warning: Factor `pm1` contains implicit NA, consider using `forcats::fct_explicit_na`
# A tibble: 23 x 5
   pm1   mneddiffs mnAboveHS minAboveHS maxAboveHS
   <fct>     <dbl>     <dbl>      <dbl>      <dbl>
 1 1.1     -0.375     0.188      0          0.375
 2 1.10    -0.0588    0.0294     0          0.0588
 3 1.11    -0.167     0.0833     0          0.167
 4 1.12    -0.143     0.0714     0          0.143
 5 1.13    -0.167     0.0833     0          0.167
 6 1.14     0         0          0          0
 7 1.15    -0.0909    0.136      0.0909     0.182
 8 1.16    -0.230     0.171      0.0556     0.286
 9 1.17    -0.0667    0.0333     0          0.0667
10 1.18    -0.0833    0.0833     0.0417     0.125
# ... with 13 more rows
```

Evaluate the design: Compare to a randomized experiment.

The within-set differences look different from those that would be expected from a randomized experiment.

```
xbHS2 <- xBalance(nhTrt~nhAboveHS,
                  strata=list(nostrat=NULL,
                              hsmatch=~fm1),
                  data=meddat,report="all")
xbHS2$results
```

```
, , strata = nostrat

           stat
vars        Control Treatment adj.diff adj.diff.null.sd std.diff      z       p
  nhAboveHS   0.163   0.05829  -0.1047          0.04883   -0.668 -2.144 0.03201

, , strata = hsmatch

           stat
vars        Control Treatment adj.diff adj.diff.null.sd std.diff      z       p
  nhAboveHS   0.113   0.09937 -0.01366         0.008237 -0.08714 -1.658 0.09728

attr(,"originals")
[1] "nhAboveHS"
```

```
xbHS2$overall
```

```
        chisquare df p.value
nostrat     4.598  1 0.03201
hsmatch     2.750  1 0.09728
```

What is xBalance doing?

```
setmeanDiffs <- meddat %>% group_by(fm1) %>%
  summarise(diffAboveHS=mean(nhAboveHS[nhTrt==1])-mean(nhAboveHS[nhTrt==0]),
            nb=n(),
            nTb = sum(nhTrt),
            nCb = sum(1-nhTrt),
            hwt = ( 2*( nCb * nTb ) / (nTb + nCb))
            )
setmeanDiffs
```

```
# A tibble: 14 x 6
   fm1   diffAboveHS    nb   nTb   nCb   hwt
   <fct>       <dbl> <int> <int> <dbl> <dbl>
 1 1.1         0         3     2     1  1.33
 2 1.10        0         2     1     1  1
 3 1.15       -0.0516    6     1     5  1.67
 4 1.16       -0.0102    3     2     1  1.33
 5 1.17        0         8     7     1  1.75
 6 1.19       -0.0667    5     1     4  1.6
 7 1.2         0         2     1     1  1
 8 1.20        0.00476   2     1     1  1
 9 1.3         0         2     1     1  1
10 1.4         0         2     1     1  1
11 1.5         0         2     1     1  1
12 1.6         0.0591    3     1     2  1.33
13 1.7        -0.0857    3     1     2  1.33
14 1.9         0         2     1     1  1
```

What is xBalance doing with multiple sets/blocks?

The test statistic is a weighted average of the set-specific differences (same approach as we would use to test the null in a block-randomized experiment)

```
## The descriptive adj.mean diff from balanceTest
with(setmeanDiffs, sum(diffAboveHS*nTb/sum(nTb)))
```

```
[1] -0.007297
```

```
## The mean diff used as the observed value in the testing
with(setmeanDiffs, sum(diffAboveHS*hwt/sum(hwt)))
```

```
[1] -0.01366
```

```
## Compare to xBalance output
xbHS2$results[,,"hsmatch"]
```

```
        Control        Treatment         adj.diff adj.diff.null.sd         std.diff
       0.113032         0.099373        -0.013659         0.008237        -0.087145        -1.6
```

Notice that balanceTest prints the set-size weighted difference (the updated version differs a little from xBalance):

```
btHS2 <- balanceTest(nhTrt~nhAboveHS+strata(fm1) +strata(pm1),
                     data=meddat,report="all")
btHS2
```

```
          strata():       fm1                                                    pm1
          stat      Treatment Control adj.diff std.diff       z       p Treatment Control adj.di
vars
nhAboveHS              0.0583  0.0656  -0.0073  -0.0541 -1.6582  0.0973    0.0583  0.1431  -0.08
---Overall Test---
    chisquare df     p.value
fm1 2.7496    1.0000 0.0973
pm1 9.8251    1.0000 0.0017
```

Summary of the Day

- We can assess the randomization of a randomized experiment easily using covariates (X): compare the observed treatment-vs-control differences in X with those consistent with no differences that would emerge from repeating the design.
- How to justify an adjustment strategy for an observational study? The linear model adjustment strategy is difficult to justify. A stratification based strategy is easier to justify, inspect, learn from. (We can compare our stratification to a block randomized experiment, to a known design, a known standard.)
- How to choose a stratification? We can do it by hand. Or we can delegate to a computer (i.e. `optmatch`) — we can think of it as an optimization problem and ask the computer to optimize.

Summary of the Day

- We can assess the randomization of a randomized experiment easily using covariates (X): compare the observed treatment-vs-control differences in X with those consistent with no differences that would emerge from repeating the design.
- How to justify an adjustment strategy for an observational study? The linear model adjustment strategy is difficult to justify. A stratification based strategy is easier to justify, inspect, learn from. (We can compare our stratification to a block randomized experiment, to a known design, a known standard.)
- How to choose a stratification? We can do it by hand. Or we can delegate to a computer (i.e. `optmatch`) — we can think of it as an optimization problem and ask the computer to optimize.

Summary of the Day

- We can assess the randomization of a randomized experiment easily using covariates (X): compare the observed treatment-vs-control differences in X with those consistent with no differences that would emerge from repeating the design.
- How to justify an adjustment strategy for an observational study? The linear model adjustment strategy is difficult to justify. A stratification based strategy is easier to justify, inspect, learn from. (We can compare our stratification to a block randomized experiment, to a known design, a known standard.)
- How to choose a stratification? We can do it by hand. Or we can delegate to a computer (i.e. `optmatch`) — we can think of it as an optimization problem and ask the computer to optimize.

References

Arceneaux, Kevin (2005). "Using cluster randomized field experiments to study voting behavior". In: The Annals of the American Academy of Political and Social Science 601.1, pp. 169–179.

Gelman, Andrew and Jennifer Hill (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Hansen, Ben B and Jake Bowers (2008). "Covariate balance in simple, stratified and clustered comparative studies". In: Statistical Science, pp. 219–236.

Ho, Daniel et al. (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". In: Political Analysis 15, pp. 199–236.

Rosenbaum, P.R. (1999). "Choice as an Alternative to Control in Observational Studies (with discussion)". In: Statistical Science 14.3, pp. 259–304.