Section 9 : Matching IV

Andrew Bertoli

30 October 2013

Roadmap

1. Matching

2. Homework

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Good Practices

1. Custom Loss Functions
2. Dealing with NAs
3. Increasing speed
4. Saving your weights
5. Saving your output

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

Custom Loss Function

Your custom loss function should take a vector containing all t-tests p-values, followed by all the ks-test p-values, for the covariates in the BalanceMatrix.

Therefore, the length of the vector it will receive is twice the length of the number of covariates in the BalanceMatrix.

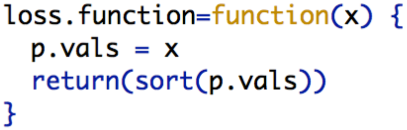
Your loss function should return either a number or a vector.

If it returns a number, then GenMatch will try to maximize that number.

If it returns a vector, then GenMatch will try to maximize the first number in that vector. In cases where multiple matching schemes result in the same first number, then GenMatch will look at the second number, and then the third number, and so on.

Matching

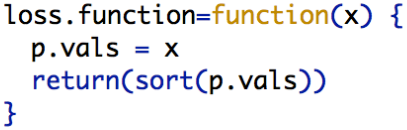
What does this loss function do?



```
loss.function=function(x) {
  p.vals = x
  return(sort(p.vals))
}
```

Matching

What does this loss function do?

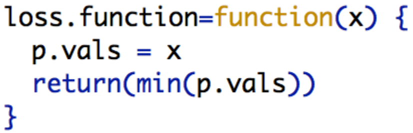


```
loss.function=function(x) {
  p.vals = x
  return(sort(p.vals))
}
```

Answer: Maximize the minimum p-value, with ties are broken by looking at the second lowest p-value. This is the default loss function.

Matching

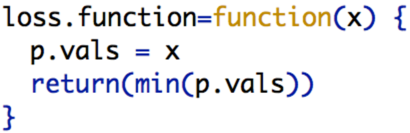
What does this loss function do?



```
loss.function=function(x) {
  p.vals = x
  return(min(p.vals))
}
```

Matching

What does this loss function do?

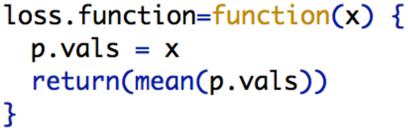


```
loss.function=function(x) {
  p.vals = x
  return(min(p.vals))
}
```

Answer: Maximize the minimum p-value. This is the loss function if you set loss=2.

Matching

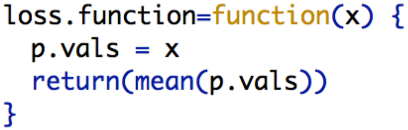
What does this loss function do?



```
loss.function=function(x) {
  p.vals = x
  return(mean(p.vals))
}
```

Matching

What does this loss function do?

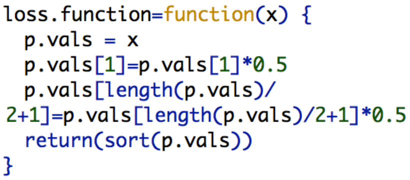


```
loss.function=function(x) {
  p.vals = x
  return(mean(p.vals))
}
```

Answer: Maximize the mean of the p-values.

Matching

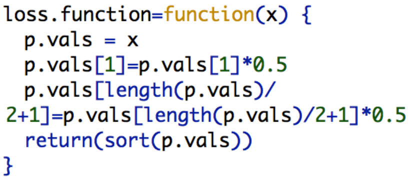
What does this loss function do?



```
loss.function=function(x) {
  p.vals = x
  p.vals[1]=p.vals[1]*0.5
  p.vals[length(p.vals)/
2+1]=p.vals[length(p.vals)/2+1]*0.5
  return(sort(p.vals))
}
```

Matching

What does this loss function do?

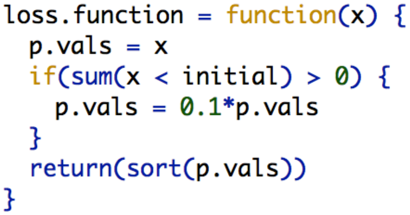


```
loss.function=function(x) {
  p.vals = x
  p.vals[1]=p.vals[1]*0.5
  p.vals[length(p.vals)/
2+1]=p.vals[length(p.vals)/2+1]*0.5
  return(sort(p.vals))
}
```

Answer: Maximize the minimum p-value, giving more weight to the first covariate (we will treat the p-values for the first covariate as half their actual value)

Matching

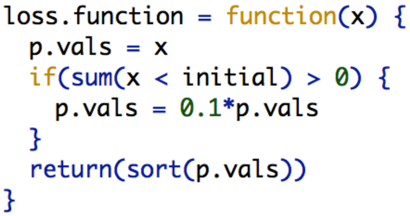
What does this loss function do (if ‘initial’ is the vector of p-values before matching)?



```
loss.function = function(x) {
  p.vals = x
  if(sum(x < initial) > 0) {
    p.vals = 0.1*p.vals
  }
  return(sort(p.vals))
}
```

Matching

What does this loss function do (if ‘initial’ is the vector of p-values before matching)?



```
loss.function = function(x) {
  p.vals = x
  if(sum(x < initial) > 0) {
    p.vals = 0.1*p.vals
  }
  return(sort(p.vals))
}
```

Answer: Restricts the search to only the weighting schemes where balance on every covariate is better than before matching. (Note: If GenMatch cannot find any such balancing scheme, the loss function defaults to maximize the minimum balance.)

Matching

If you want to work with something besides the vector of p-values from t-tests and KS-tests, you can change the **fit.func** argument.

```
GenMatch(Tr, X, BalanceMatrix=X, estimand=“ATT”, M=1,
weights=NULL, pop.size = 100, max.generations=100,
wait.generations=4, hard.generation.limit=FALSE,
starting.values=rep(1,ncol(X)), fit.func=”pvals”,
MemoryMatrix=TRUE, exact=NULL, caliper=NULL,
replace=TRUE, ties=TRUE, CommonSupport=FALSE, nboots =
0, ks=TRUE, verbose=FALSE, distance.tolerance=1e-05,
tolerance=sqrt(.Machine$double.eps), min.weight=0,
max.weight=1000, Domains=NULL, print.level=2,
project.path=NULL, paired=TRUE, loss=1,
data.type.integer=FALSE, restrict=NULL, cluster=FALSE,
balance=TRUE, ...)
```

See the GenMatch() help library for details.

Matching

Question: How do deal with NAs when you are matching?

Matching

Question: How do deal with NAs when you are matching?

Answer: Match on missingness

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Example

Say we on a farm, and we are doing a study on chicks.

We decide to match on weight and age, but we are missing some values for age.

Step 1: Create a new vector that keeps track of which units have NAs for age. Code the units with missing values as 0 and the units with known values as 1.

Step 2: Set all the NAs for age at the mean of that variable.

Step 3: Match on weight, age, and missing age.

This method can be used for anytime you are missing data. It can be used for any matching method, including Genetic Matching.

Matching

Increasing Speed

You run GenMatch using multiple chips in your computer.

Details are available here:

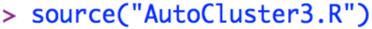
`http://sekhon.berkeley.edu/rgenoud/`

Matching

Instructions for Mac

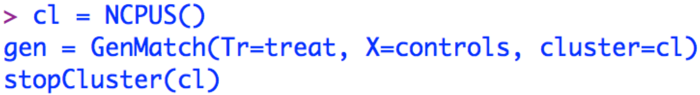
Put AutoCluster3.R into a text file in your working directory.

Enter this code:



```
> source("AutoCluster3.R")
```

Then type in this code



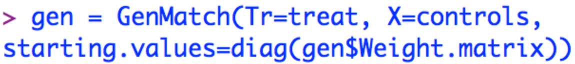
```
> cl = NCPUS()
gen = GenMatch(Tr=treat, X=controls, cluster=cl)
stopCluster(cl)
```

Matching

Many choices for the GenMatch function depend on the data you have.

Thus, you will typically run GenMatch multiple times to try to find what choices result in the best balance.

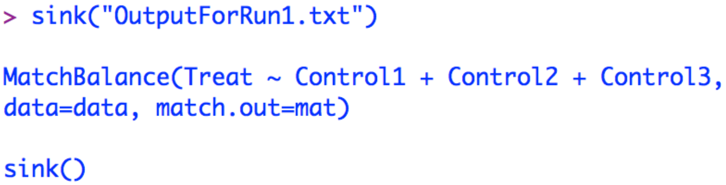
After every run, you should use the weights that you ended on last time.



```
> gen = GenMatch(Tr=treat, X=controls,
starting.values=diag(gen$Weight.matrix))
```

Matching

You can save the balance summary after each run.



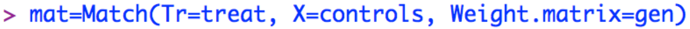
```
> sink("OutputForRun1.txt")

MatchBalance(Treat ~ Control1 + Control2 + Control3,
data=data, match.out=mat)

sink()
```

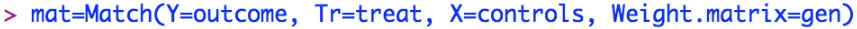
Matching

While playing around with the GenMatch options to find the best balance, you should not even use the outcome variable.



```
> mat=Match(Tr=treat, X=controls, Weight.matrix=gen)
```

Once you have settled on a matching scheme, then you can look at the outcome variable.



```
> mat=Match(Y=outcome, Tr=treat, X=controls, Weight.matrix=gen)
```

Homework

Problem 1

Ten thousand high school students take a test, and everyone who scores in the top 10% receive a college scholarship. A researcher decides to test the effect of receiving the scholarship on the likelihood of going to college.

Homework

a) The researcher compares students who scored between the 89th and 90th percentile to students who scored between the 90th and 91st percentile using a standard t-test. The results are barely significant at the 5% level, and the estimated treatment effect is small. However, the researcher claims that the significant results provide strong evidence that the scholarship had an effect. Name at least two objections that could be raised against this researcher's conclusions.

Homework

a) The researcher compares students who scored between the 89th and 90th percentile to students who scored between the 90th and 91st percentile using a standard t-test. The results are barely significant at the 5% level, and the estimated treatment effect is small. However, the researcher claims that the significant results provide strong evidence that the scholarship had an effect. Name at least two objections that could be raised against this researcher's conclusions.

1. The results may be sensitive to the window the researcher picked.

Homework

a) The researcher compares students who scored between the 89th and 90th percentile to students who scored between the 90th and 91st percentile using a standard t-test. The results are barely significant at the 5% level, and the estimated treatment effect is small. However, the researcher claims that the significant results provide strong evidence that the scholarship had an effect. Name at least two objections that could be raised against this researcher's conclusions.

1. The results may be sensitive to the window the researcher picked.

2. People between the 90th and 91st percentile will tend to be a little different than people between the 89th and 90th percentile, in terms of intelligence and many other factors. Thus, the significant but small estimated treatment effect could result from these small baseline differences rather than a true effect of the scholarship.

Homework

b. The researcher than uses local linear regression to estimate the treatment effect at the cut-point. How does this approach help resolve the problems mentioned in Part (a). Name one new problem that this approach creates.

Homework

b. The researcher than uses local linear regression to estimate the treatment effect at the cut-point. How does this approach help resolve the problems mentioned in Part (a). Name one new problem that this approach creates.

The results will depend on the model being correct. Data not close to the cut-point may influence the results, depending on the bandwidth that is chosen. The results will also likely be sensitive to bandwidth selection.

Homework

c) The researcher decides to run linear regression for the data on both sides of the cut-point to estimate the treatment effect at the cut-point. Show formally that, when estimating a regression within a bounded region, the estimates at the boundaries are more variable then anywhere else. What does this mean for the researcher's estimates?

Homework

c) The researcher decides to run linear regression for the data on both sides of the cut-point to estimate the treatment effect at the cut-point. Show formally that, when estimating a regression within a bounded region, the estimates at the boundaries are more variable then anywhere else. What does this mean for the researcher's estimates?

It can be shown (for example, in Mathematical Statistics and Data Analysis by John Rice) that the variance for the regression estimate at x_i is

$$\sigma^2 \left(\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{N} (x_j - \bar{x})^2} \right)$$

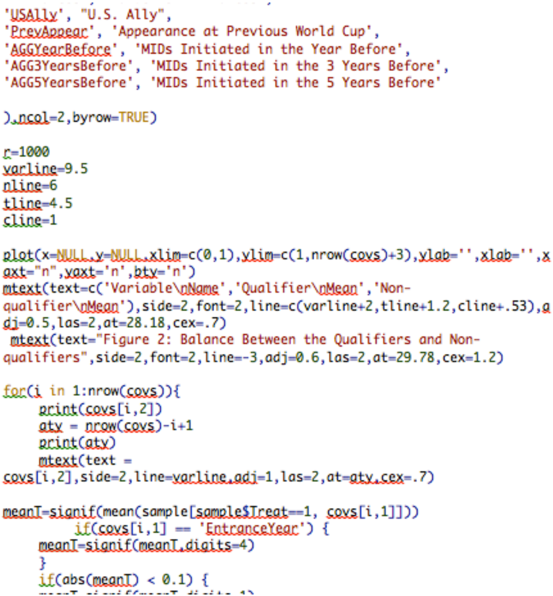
This variance increases as the distance between x_i and $\bar{x}$ grows. So estimates are most variable at the boundary.

Homework

Problem 3

a) Recreate the balance plot using permutation tests instead of paired t-tests. Plot only the ten covariates that you think are most important. Make another plot using equivalence tests. Remember to preserve the paired structure of the data

Homework



```
'USAlly', "U.S. Ally",
'PrevAppear', 'Appearance at Previous World Cup',
'AGGYearBefore', 'MIDs Initiated in the Year Before',
'AGG3YearsBefore', 'MIDs Initiated in the 3 Years Before',
'AGG5YearsBefore', 'MIDs Initiated in the 5 Years Before'

),ncol=2,byrow=TRUE)

r=1000
varline=9.5
nline=6
tline=4.5
cline=1

plot(x=NULL,y=NULL,xlim=c(0,1),ylim=c(1,nrow(covs)+3),ylab='',xlab='',x
axt="n",yaxt='n',bty='n')
mtext(text=c('Variable\nName','Qualifier\nMean','Non-
qualifier\nMean'),side=2,font=2,line=c(varline+2,tline+1.2,cline+.53),a
dj=0.5,las=2,at=28.18,cex=.7)
 mtext(text="Figure 2: Balance Between the Qualifiers and Non-
qualifiers",side=2,font=2,line=-3,adj=0.6,las=2,at=29.78,cex=1.2)

for(i in 1:nrow(covs)){
     print(covs[i,2])
     aty = nrow(covs)-i+1
     print(aty)
     mtext(text =
covs[i,2],side=2,line=varline,adj=1,las=2,at=aty,cex=.7)

meanT=signif(mean(sample[sample$Treat==1, covs[i,1]]))
          if(covs[i,1] == 'EntranceYear') {
     meanT=signif(meanT,digits=4)
     }
     if(abs(meanT) < 0.1) {
     meanT=signif(meanT,digits=1)
```

Homework

b) Run three tests to assess whether the World Cup has an effect on state aggression. Which one do you think is most appropriate given the nature of the data?

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$E[\hat{\tau}_{Dem}] = E[\frac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \frac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$$E[\hat{\tau}_{Dem}] = E[\frac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \frac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k}\sum_{i=1}^{k} E[\{Agg_i : T_i = 1, D_i = 1\}] - \frac{1}{j}\sum_{i=1}^{j} E[\{Agg_i : T_i = 0, D_i = 1\}]$$

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$$E[\hat{\tau}_{Dem}] = E[\tfrac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \tfrac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \tfrac{1}{k}\sum_{i=1}^{k} E[\{Agg_i : T_i = 1, D_i = 1\}] - \tfrac{1}{j}\sum_{i=1}^{j} E[\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \tfrac{1}{k} \cdot k \cdot mean(\{Agg : T_i = 1, D_i = 1\}) - \tfrac{1}{j} \cdot j \cdot mean(\{Agg : T_i = 0, D_i = 1\})$$

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$$E[\hat{\tau}_{Dem}] = E[\frac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \frac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k}\sum_{i=1}^{k} E[\{Agg_i : T_i = 1, D_i = 1\}] - \frac{1}{j}\sum_{i=1}^{j} E[\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k} \cdot k \cdot mean(\{Agg : T_i = 1, D_i = 1\}) - \frac{1}{j} \cdot j \cdot mean(\{Agg : T_i = 0, D_i = 1\})$$

$$E[\hat{\tau}_{Dem}] = mean(\{Agg : T_i = 1, D_i = 1\}) - mean(\{Agg : T_i = 0, D_i = 1\})$$

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$$E[\hat{\tau}_{Dem}] = E[\frac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \frac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k}\sum_{i=1}^{k} E[\{Agg_i : T_i = 1, D_i = 1\}] - \frac{1}{j}\sum_{i=1}^{j} E[\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k} \cdot k \cdot mean(\{Agg : T_i = 1, D_i = 1\}) - \frac{1}{j} \cdot j \cdot mean(\{Agg : T_i = 0, D_i = 1\})$$

$$E[\hat{\tau}_{Dem}] = mean(\{Agg : T_i = 1, D_i = 1\}) - mean(\{Agg : T_i = 0, D_i = 1\})$$

$$E[\hat{\tau}_{Dem}] = \tau_{Dem}$$

Homework

c) Assuming randomness in the RD window, are the estimated treatment effects for democracies and non-democracies guaranteed to be unbiased? Prove your answer.

The estimator is unbiased. Let k be the number of treated democracies and j be the number of control democracies.

$$E[\hat{\tau}_{Dem}] = E[\frac{1}{k}\sum_{i=1}^{k}\{Agg_i : T_i = 1, D_i = 1\} - \frac{1}{j}\sum_{i=1}^{j}\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k}\sum_{i=1}^{k} E[\{Agg_i : T_i = 1, D_i = 1\}] - \frac{1}{j}\sum_{i=1}^{j} E[\{Agg_i : T_i = 0, D_i = 1\}]$$

$$E[\hat{\tau}_{Dem}] = \frac{1}{k}\cdot k \cdot mean(\{Agg : T_i = 1, D_i = 1\}) - \frac{1}{j}\cdot j \cdot mean(\{Agg : T_i = 0, D_i = 1\})$$

$$E[\hat{\tau}_{Dem}] = mean(\{Agg : T_i = 1, D_i = 1\}) - mean(\{Agg : T_i = 0, D_i = 1\})$$

$$E[\hat{\tau}_{Dem}] = \tau_{Dem}$$

The same logic shows that τ_{Nondem} is unbiased.

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

But we have

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

But we have

$$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{0+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{2+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{0+2+0}{3}]$$

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

But we have

$$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{0+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{2+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{0+2+0}{3}]$$

$$E[\hat{\tau}] = \frac{2}{3} - \frac{2}{9} - \frac{2}{9} = \frac{6}{3} - \frac{2}{9} - \frac{2}{9}$$

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

But we have

$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{0+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{2+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{0+2+0}{3}]$

$E[\hat{\tau}] = \frac{2}{3} - \frac{2}{9} - \frac{2}{9} = \frac{6}{3} - \frac{2}{9} - \frac{2}{9} = \frac{2}{9}$

Homework

d) Imagine that rather than using pairs of countries, I used all states close to the cut-point to increase my sample size. For instance, if Italy scored 7 (did not qualify), Austria scored 8 (qualified), and Germany scored 9 (qualified), I would use all three countries in the sample. Would the difference in means estimator be guaranteed to be unbiased? Prove your answer.

The estimator is biased. Imagine that the three countries above are the only ones in our sample. Say that Austria and Germany will start 0 MIDS whether they go or not, and Italy will start 2 if they go and 4 if they do not. We also have France and Britain in our sample, both of which started will start no MIDs regardless of whether they go.

First, note that $\tau = \frac{0+0+(4-2)+0+0}{5} = \frac{2}{5}$

But we have

$$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{0+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{2+0+0}{3}] + \frac{1}{3}[\frac{0+0}{2} - \frac{0+2+0}{3}]$$

$$E[\hat{\tau}] = \frac{2}{3} - \frac{2}{9} - \frac{2}{9} = \frac{6}{3} - \frac{2}{9} - \frac{2}{9} = \frac{2}{9} \neq \tau$$

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

In the previous example we would have

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

In the previous example we would have

$$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{(0+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(2+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(0+2)/2+0}{2}]$$

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

In the previous example we would have

$$E[\hat{\tau}] = \tfrac{1}{3}[\tfrac{4+0}{2} - \tfrac{(0+0)/2+0}{2}] + \tfrac{1}{3}[\tfrac{0+0}{2} - \tfrac{(2+0)/2+0}{2}] + \tfrac{1}{3}[\tfrac{0+0}{2} - \tfrac{(0+2)/2+0}{2}]$$

$$E[\hat{\tau}] = \tfrac{2}{3} - \tfrac{1}{6} - \tfrac{1}{6}$$

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

In the previous example we would have

$$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{(0+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(2+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(0+2)/2+0}{2}]$$

$$E[\hat{\tau}] = \frac{2}{3} - \frac{1}{6} - \frac{1}{6}$$

$$E[\hat{\tau}] = \frac{1}{3} \neq \tau$$

Homework

e) Say I used the design in Part (d), but I took the mean of Austria and Germany and counted it as a single data point. Would the difference in means estimator be unbiased? Prove it.

In the previous example we would have

$E[\hat{\tau}] = \frac{1}{3}[\frac{4+0}{2} - \frac{(0+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(2+0)/2+0}{2}] + \frac{1}{3}[\frac{0+0}{2} - \frac{(0+2)/2+0}{2}]$

$E[\hat{\tau}] = \frac{2}{3} - \frac{1}{6} - \frac{1}{6}$

$E[\hat{\tau}] = \frac{1}{3} \neq \tau$

So the estimator is biased.

Homework

Problem 3

a) If close elections were random, are the units on either side of the cut-point exchangeable? Why or why not?

Homework

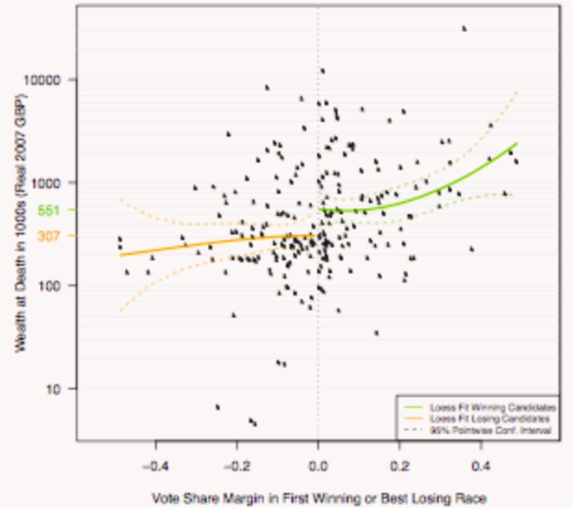
Problem 3

a) If close elections were random, are the units on either side of the cut-point exchangeable? Why or why not?

No, the near winner and near loser in each election are exchangeable. If each candidate only ran once, then this exchangeability would hold. But since the near losers are plotted once (for their best race), the paired structure of the data is violated.

Homework

Problem 3



Homework

b) Along similar lines, Eggers and Hainmueller show that the units to the right and left of the cut-point are similar across a number of covariates. Does that provide the same type of evidence that the RD worked as the balance plot from the World Cup example? Explain your answer.

Homework

b) Along similar lines, Eggers and Hainmueller show that the units to the right and left of the cut-point are similar across a number of covariates. Does that provide the same type of evidence that the RD worked as the balance plot from the World Cup example? Explain your answer.

No. In the World Cup example, the units are paired. Eggers and Hainmueller provide evidence that units are exchangeable that are not (in theory) exchangeable based on their design.

Homework

Bonus

Run a simulation in R where candidates run against each other in pairs across 5 elections. Start with 1000 candidates, each with a fixed amount of wealth drawn from Exp(1/100,000). For each election, the vote difference between Candidate i and Candidate j is drawn from Unif(-2,2). If a candidate wins, record their vote share and wealth, and remove them from the sample. If a candidate loses, they run again with probability $\sqrt{W_i/W_{max}}$, where W_i is their wealth and W_{max} is the maximum wealth of any candidate. After each election, create new candidates to replace removed candidates (either winners or losers that do not run again). Create a continuity plot like the one above. What do your results suggest about the findings from the paper?