Re-analysis of Acemoglu

```r
# Re-analysis of Acemoglu et al.

require(haven)
```

```
## Loading required package: haven
```

```r
require(readxl)
```

```
## Loading required package: readxl
```

```r
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
require(magrittr)
```

```
## Loading required package: magrittr
```

```r
require(tidyr)
```

```
## Loading required package: tidyr

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(modelr)
```

```
## Loading required package: modelr
```

```r
require(broom)
```

```
## Loading required package: broom

##
## Attaching package: 'broom'

## The following object is masked from 'package:modelr':
##
##     bootstrap
```

```r
require(purrr)
```

```
## Loading required package: purrr

##
## Attaching package: 'purrr'

## The following object is masked from 'package:magrittr':
##
##     set_names

## The following objects are masked from 'package:dplyr':
##
##     contains, order_by
```

```r
require(pbapply)
```

```
## Loading required package: pbapply
```

```r
require(parallel)
```

```
## Loading required package: parallel
```

```r
fiveyear <- read_dta("Acemoglu one year panel.dta")
oneyear <- read_excel("Income and Democracy Data AER adjustment.xls",
sheet = "Annual Panel")
oneyear <- oneyear %>% group_by(country) %>% mutate(l1_fhpolrigaug=lag(fhpolrigaug,order_by=year),
                                                     l1_lrgdpch=lag(lrgdpch,order_by=year)) %>%
  filter(!(is.na(fhpolrigaug)) & !(is.na(lrgdpch))) %>% mutate(panel_balance=n())
```

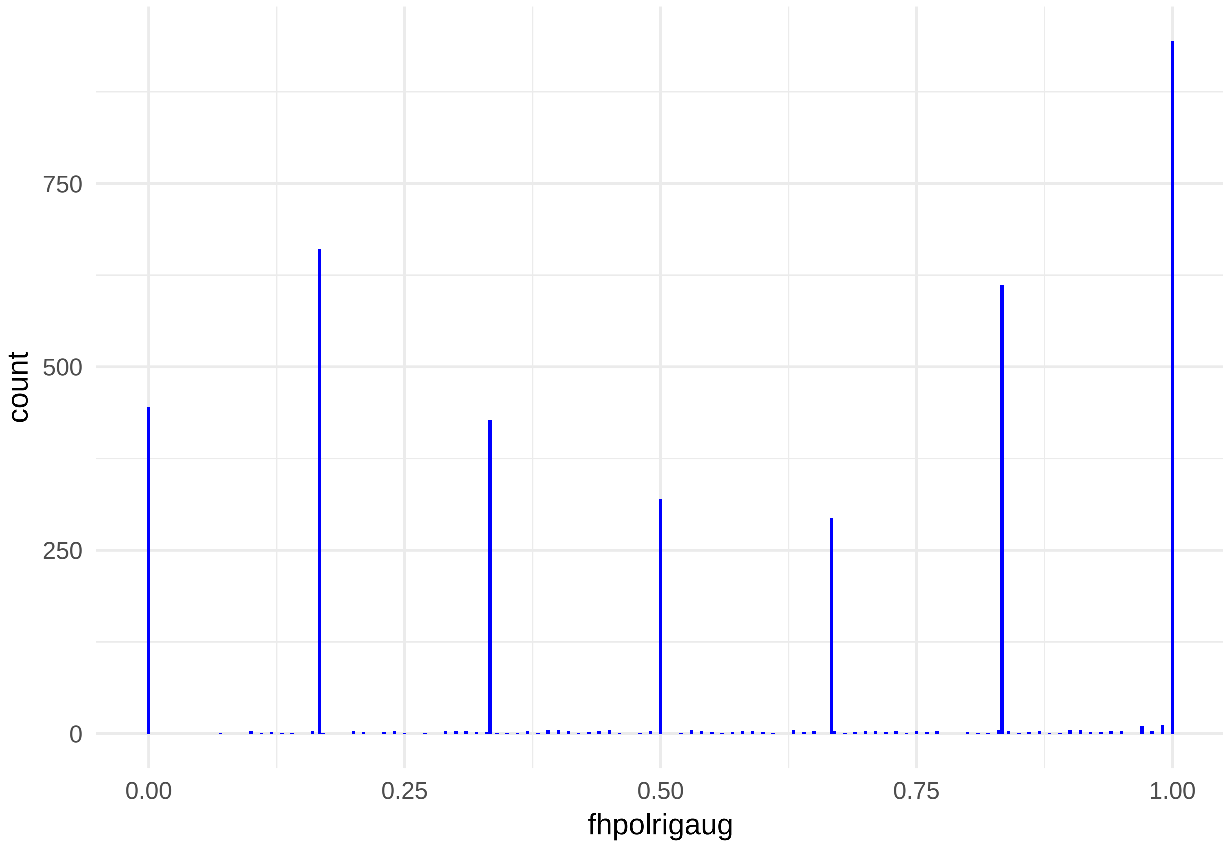
The issue with the DV

One un-noticed issue in the data is that the DV is fixed within countries. We have been focusing on IVs that could be fixed within or between, but not DVs. Of course, the model can't drop the DV, so it seems like it can screw up the 2-way estimate when there isn't enough variance in the DV either in the cross-section or the over-time effect.

I haven't figured out yet how to reproduce this; you might want to run a simulation to test it more accurately.

First we look at the distribution of the DV in the Acemoglu dataset. We see that there are spikes at 0 and 1, and that in general there isn't a whole lot of variation in the DV even though it is being treated as continuous.

```r
ggplot(oneyear,aes(x=fhpolrigaug)) + geom_bar(fill='blue') + theme_minimal()
```



Now to illustrate the problem, I use the dataset only with Canada and the United States.

```
data3 <-  filter(oneyear,country %in% c('United States','Canada')) %>% select(country,year,fhpolrigaug,l
twoway <- round(coef(lm(data=data3,formula=fhpolrigaug~ lrgdpch+ factor(year) + factor(country)))['lrgd
pooled <- round(coef(lm(data=data3,formula=fhpolrigaug~ lrgdpch))['lrgdpch'],3)
time <- round(coef(lm(data=data3,formula=fhpolrigaug~ lrgdpch+ factor(year)))['lrgdpch'],3)
case <- round(coef(lm(data=data3,formula=fhpolrigaug~ lrgdpch+ factor(country)))['lrgdpch'],3)
```

In this situation with these two countries, the 2-way effect is -0.02, the pooled estimate is 0.015, the time estimate is -0.028 and the case estimate is 0.019. Thus 2-way estimate is between the pooled and case estimate, although it is converging to the time estimate.

However, when we look at the distribution of the DV in this dataset, we see a lot of 1s:

```
print(as.data.frame(data3))
```

```
##          country year fhpolrigaug   lrgdpch l1_fhpolrigaug l1_lrgdpch
## 1         Canada 1950        0.99  9.115211             NA         NA
## 2         Canada 1955        1.00  9.205848             NA   9.150495
## 3         Canada 1960        1.00  9.247975             NA   9.251952
## 4         Canada 1965        1.00  9.429002             NA   9.380478
## 5         Canada 1972        1.00  9.625804             NA   9.581972
## 6         Canada 1973        1.00  9.689451              1   9.625804
## 7         Canada 1974        1.00  9.720343              1   9.689451
## 8         Canada 1975        1.00  9.719351              1   9.720343
## 9         Canada 1976        1.00  9.764153              1   9.719351
## 10        Canada 1977        1.00  9.779172              1   9.764153
## 11        Canada 1978        1.00  9.810351              1   9.779172
## 12        Canada 1979        1.00  9.853029              1   9.810351
```

```
## 13        Canada 1980        1.00  9.851376              1   9.853029
## 14        Canada 1981        1.00  9.880587              1   9.851376
## 15        Canada 1982        1.00  9.817758              1   9.880587
## 16        Canada 1983        1.00  9.840611              1   9.817758
## 17        Canada 1984        1.00  9.897246              1   9.840611
## 18        Canada 1985        1.00  9.937892              1   9.897246
## 19        Canada 1986        1.00  9.952429              1   9.937892
## 20        Canada 1987        1.00  9.986536              1   9.952429
## 21        Canada 1988        1.00 10.027080              1   9.986536
## 22        Canada 1989        1.00 10.036050              1  10.027080
## 23        Canada 1990        1.00 10.014570              1  10.036050
## 24        Canada 1991        1.00  9.973947              1  10.014570
## 25        Canada 1992        1.00  9.967081              1   9.973947
## 26        Canada 1993        1.00  9.980831              1   9.967081
## 27        Canada 1994        1.00 10.021720              1   9.980831
## 28        Canada 1995        1.00 10.040930              1  10.021720
## 29        Canada 1996        1.00 10.047190              1  10.040930
## 30        Canada 1997        1.00 10.089140              1  10.047190
## 31        Canada 1998        1.00 10.118300              1  10.089140
## 32        Canada 1999        1.00 10.161900              1  10.118300
## 33        Canada 2000        1.00 10.200050              1  10.161900
## 34 United States 1950        0.99  9.278240             NA         NA
## 35 United States 1955        0.98  9.389680             NA   9.331630
## 36 United States 1960        0.95  9.415136             NA   9.406607
## 37 United States 1965        0.92  9.594625             NA   9.541778
## 38 United States 1972        1.00  9.780488             NA   9.730689
## 39 United States 1973        1.00  9.834690              1   9.780488
## 40 United States 1974        1.00  9.827986              1   9.834690
## 41 United States 1975        1.00  9.800418              1   9.827986
## 42 United States 1976        1.00  9.855587              1   9.800418
## 43 United States 1977        1.00  9.903549              1   9.855587
## 44 United States 1978        1.00  9.952475              1   9.903549
## 45 United States 1979        1.00  9.977451              1   9.952475
## 46 United States 1980        1.00  9.968130              1   9.977451
## 47 United States 1981        1.00  9.983500              1   9.968130
## 48 United States 1982        1.00  9.940727              1   9.983500
## 49 United States 1983        1.00  9.974105              1   9.940727
## 50 United States 1984        1.00 10.045710              1   9.974105
## 51 United States 1985        1.00 10.070010              1  10.045710
## 52 United States 1986        1.00 10.092960              1  10.070010
## 53 United States 1987        1.00 10.119860              1  10.092960
## 54 United States 1988        1.00 10.150640              1  10.119860
## 55 United States 1989        1.00 10.176520              1  10.150640
## 56 United States 1990        1.00 10.183310              1  10.176520
## 57 United States 1991        1.00 10.161920              1  10.183310
## 58 United States 1992        1.00 10.184470              1  10.161920
## 59 United States 1993        1.00 10.201960              1  10.184470
## 60 United States 1994        1.00 10.235590              1  10.201960
## 61 United States 1995        1.00 10.254460              1  10.235590
## 62 United States 1996        1.00 10.281720              1  10.254460
## 63 United States 1997        1.00 10.315280              1  10.281720
## 64 United States 1998        1.00 10.344660              1  10.315280
## 65 United States 1999        1.00 10.377480              1  10.344660
## 66 United States 2000        1.00 10.413100              1  10.377480
```

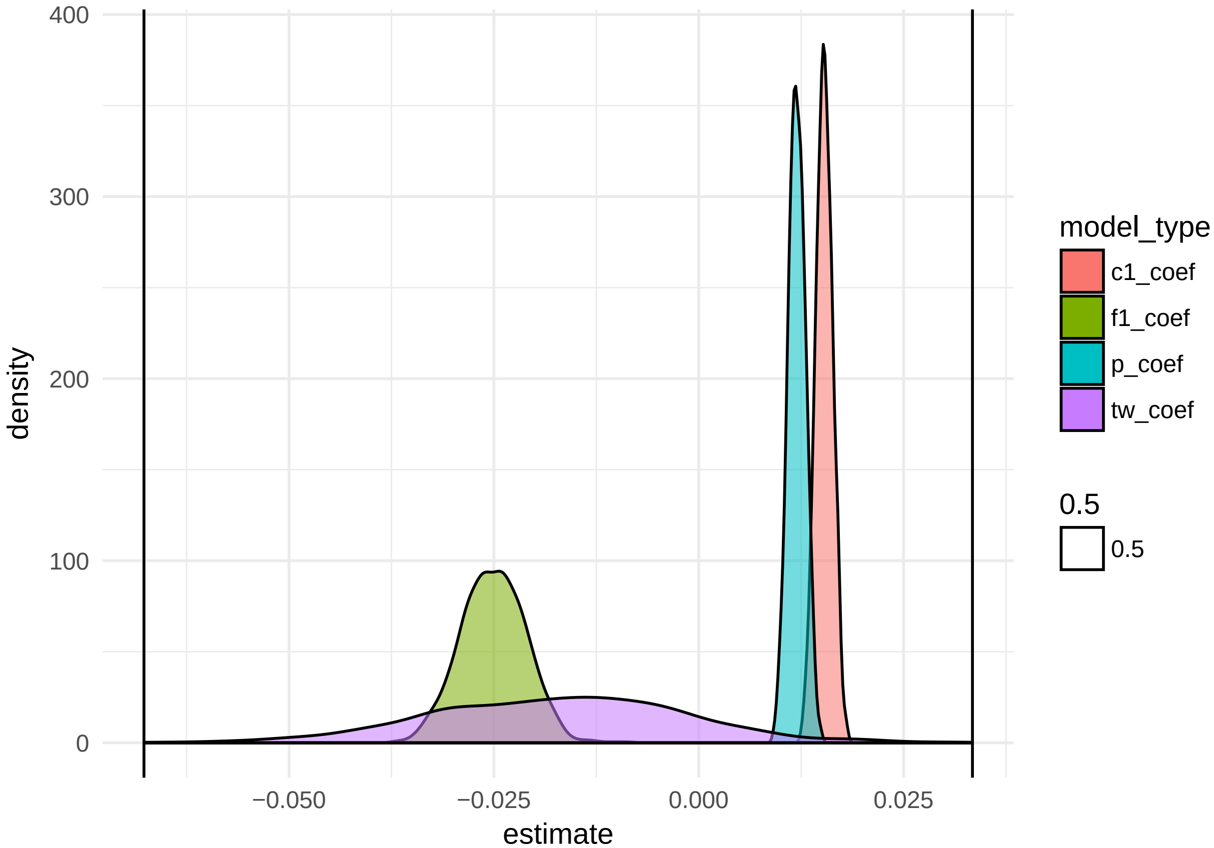
Canada apparently went from .99 to 1 between 1950 to 1955. They must have given the Mounties the right to vote.

Given our framework, this dataset is particularly problematic for two-way estimation because there isn't much variation in the DV, and the 2-way estimate has to combine variation along several dimensions. To test this, I run a simulation in which I replace the 1s with a random number between 0.99 and .9999. It is a very small difference in the variation, but it can lead to startling results.

```
tw_coef <- 1:1000
f1_coef <- 1:1000
c1_coef <- 1:1000
p_coef <- 1:1000
m_replace <- 1:1000
m_diff <- 1:1000
for(i in 1:1000) {
data3 <-  filter(oneyear,country %in% c('United States','Canada')) %>% select(country,year,fhpolrigaug,l
replacement <- runif(nrow(data3),0.99,.9999)
data3 <- ungroup(data3) %>% mutate(fhpolrigaug2 = ifelse(fhpolrigaug==1,replacement,fhpolrigaug))
tw_coef[i] <- coef(lm(data=data3,formula=fhpolrigaug2~ lrgdpch+ factor(year) + factor(country)))['lrgdp
f1_coef[i] <- coef(lm(data=data3,formula=fhpolrigaug2~ lrgdpch+ factor(year)))['lrgdpch']
c1_coef[i] <- coef(lm(data=data3,formula=fhpolrigaug2~ lrgdpch+ factor(country)))['lrgdpch']
p_coef[i] <- coef(lm(data=data3,formula=fhpolrigaug2~ lrgdpch))['lrgdpch']
m_replace[i] <- mean(ifelse(data3$fhpolrigaug==1,replacement,data3$fhpolrigaug),na.rm=TRUE)
out_diff <- data3 %>% group_by(country) %>% summarize(mean_countries=mean(fhpolrigaug2))
m_diff[i] <- (out_diff$mean_countries[1] - out_diff$mean_countries[2])^2
}
```

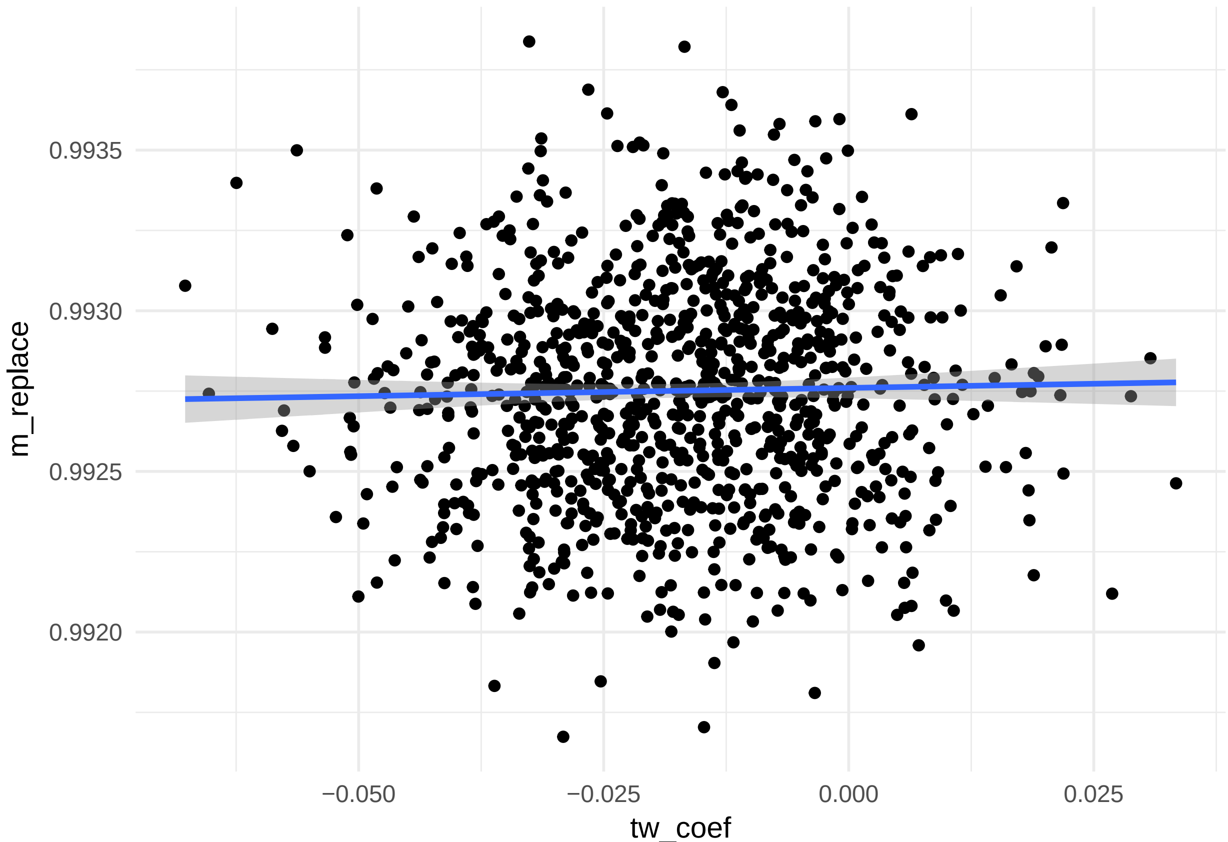
Then we can look at the distribution of coefficients across the different models. The black bars on either side of the graph indicate the maximum and minimum observed coefficient from the two-way FE model. As can be seen, the 1-way and pooled estimates (c1,f1 and p) are largely stable, but the two_way estimate (tw_coef) has a greater range and much, much wider variance than any of the other models. The 1-way estimates are largely immune to the small amount of random noise injected into the model, but the 2-way estimates are highly influenced by it.

```
data_frame(tw_coef,f1_coef,c1_coef,p_coef) %>% gather(model_type,estimate) %>%
  ggplot(aes(x=estimate,fill=model_type,alpha=0.5)) + geom_density() + theme_minimal() +
  geom_vline(xintercept=c(min(tw_coef),max(tw_coef)))
```



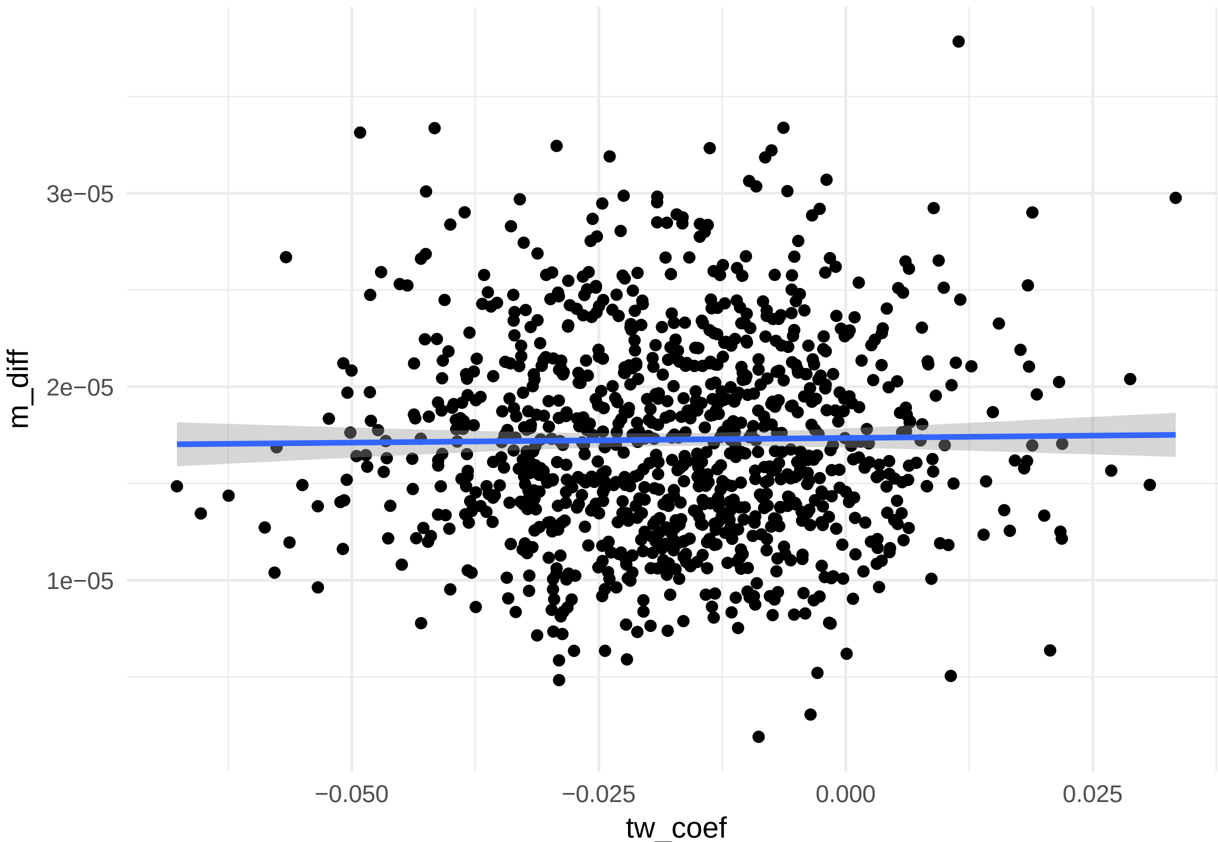
We can see if small differences in the uniform random values may have driven the change in the 2-way coefficient by comparing mean values of the random numbers to the 2-way coefficients:

```
data_frame(m_replace,tw_coef) %>% ggplot(aes(y=m_replace,x=tw_coef)) + geom_point() + stat_smooth(metho
```



It is not related to the small differences in the sampling error of the random uniform numbers. We can also look at mean differences in the DV between the US and Canada:

```
data_frame(m_diff,tw_coef) %>% ggplot(aes(y=m_diff,x=tw_coef)) + geom_point() + stat_smooth(method='lm')
```



As can be seen, it is difficult to predict when and on what side the two-way coefficient will fall even given minute levels of random noise. It seems we should warn the reader of this problem, and also mention it in the simulations.