Difference-in-Difference-in-Differences
CFPLs Case Study

Pdf version of this document

Github repository for this document

Chandler Lutz

Last Updated 2018-03-04

Overview

This case study will use a difference-in-difference-in-differences (DDD) research design to study the impact of the 2008 California Foreclosure Prevention Laws (CFPLs) that aimed to reduce foreclosures at the height of the financial crises. These laws are called the California Foreclosure Prevention Laws. We first start with a classic difference-in-differences analysis and then cover the DDD approach.

Preliminaries

- This tutorial is based on Gabriel et al. (2017)
- The Github Repository for this project is here.
- Download the most recent pdf version here.
- Data can be downloaded in either rds or csv here.

This tutorial will employ the `R` statistical computing environment and the `data.table` and `magrittr` packages for data manipulation as well as the `AER` package for regression analysis.

DDD Notes and resources

- See this CrossValidated Question for an overview of the usual difference-in-differences methodology
- Wooldridge Lecture Notes
- These Lecture Notes (More Advanced)

Policy Background

In 2008, California housing markets were spiraling downwards. California Policymakers implemented a set of new foreclosure laws that increased the pecuniary and time costs of foreclosure in an attempt to mitigate the rise in foreclosures. Our aim is to analyze the effects of these policies on foreclosures (See Gabriel et. al. (2017) for more details).

Data

```r
#Load packages
library(ggplot2); library(magrittr); library(data.table); library(AER)

#Read in the Data
DT <- fread("https://raw.githubusercontent.com/ChandlerLutz/difference-in-difference-in-differences-CFPL
```

The data contain the following variables:

- `CA` – an indicator equal to 1 for California and zero otherwise
- `sand.state` – an indicator equal to 1 for Sand States (AZ, CA, FL, NV), states that experienced the largest boom and bust during the 2000s
- `state.fips` – the two-digit state fips code
- `fips.code` – the five-digit county fips code
- `CFPL` – and indicator that equals `1` for the CFPL treatment period
- `zillow.forc` – the real estate owned (REO) foreclosures per 10,000 homes
- `forc.high` – and indicator equal to 1 for "high" foreclosure counties. Classification of counties as high or low foreclosure counties is from Gabriel et al. (2017).
- `hh2000` – the number of housing units in 2000 from the US Census

```r
head(DT)
```

```
##    CA sand.state state.fips fips.code CFPL zillow.forc forc.high hh2000
## 1:  0          0          1      1097    0    159.7454         0 165101
## 2:  0          0          1      1097    1    735.7570         0 165101
## 3:  0          0          1      1117    0     67.8369         0  59302
## 4:  0          0          1      1117    1    477.7105         0  59302
## 5:  0          1          4      4003    0     54.4248         0  51126
## 6:  0          1          4      4003    1    474.9314         0  51126
```

Research Goal

Our aim is to estimate the effects of the CFPLs on the foreclosures, `zillow.forc`.

Difference-in-differences (DD)

We can first analyze the effects of the program using a simple DD setup. The idea behind a DD research design, is that we can estimate the effects of the policy by comparing the change in mean foreclosures for counties in California relative to the change in mean foreclosures for Arizona and Nevada. Specifically, we will use counties in Arizona and Nevada as the control group and counties in California as the treatment group. Together, California, Arizona, and Nevada make up the "Sand States" (note: Florida is also considered a Sand State, but Florida is not in our dataset).

Let `CFPL = 0` for the pre-CFPL period and `CFPL = 1` for the CFPL treatment period. Also, `CA = 1` for California counties and `CA = 0` for counties in other Sand States. Our DD means estimator is

$$\hat{\delta} = (\bar{Y}_{\text{CA = 1, CFPL = 1}} - \bar{Y}_{\text{CA = 1, CFPL = 0}}) - (\bar{Y}_{\text{CA = 0, CFPL = 1}} - \bar{Y}_{\text{CA = 0, CFPL = 0}})$$

$\hat{\delta}$ is the difference-in-differences estimator (the difference in two differences).

The first difference is the mean change in Y for California counties (the treatment group), $(\bar{Y}_{\text{CA = 1, CFPL = 1}} - \bar{Y}_{\text{CA = 1, CFPL = 0}})$. This first difference is also often called the "before-after" estimator as we compare $\bar{Y}$ in California before and after the policy

The second difference is the mean change in Y for non-California counties (the control group).

Taking the difference of the above differences yields the difference-in-differences estimate

Difference-in-differences (DD) means estimate

Let's first look at the summary statistics in California versus the rest of the sand states, where the other sand states (Arizona and Nevada) are the control group. This will serve as the basis for our DD means estimator

```
dd.means <- DT %>%
    #Get the means by the CA and CFPL dummies for the sand states
    .[sand.state == 1,
      .(mean.zillow.forc = weighted.mean(zillow.forc, w = hh2000)),
      by = .(CA, CFPL)]
print(dd.means)
```

```
##    CA CFPL mean.zillow.forc
## 1:  0    0         206.1253
## 2:  0    1        1539.1376
## 3:  1    0         222.7555
## 4:  1    1         895.6155
```

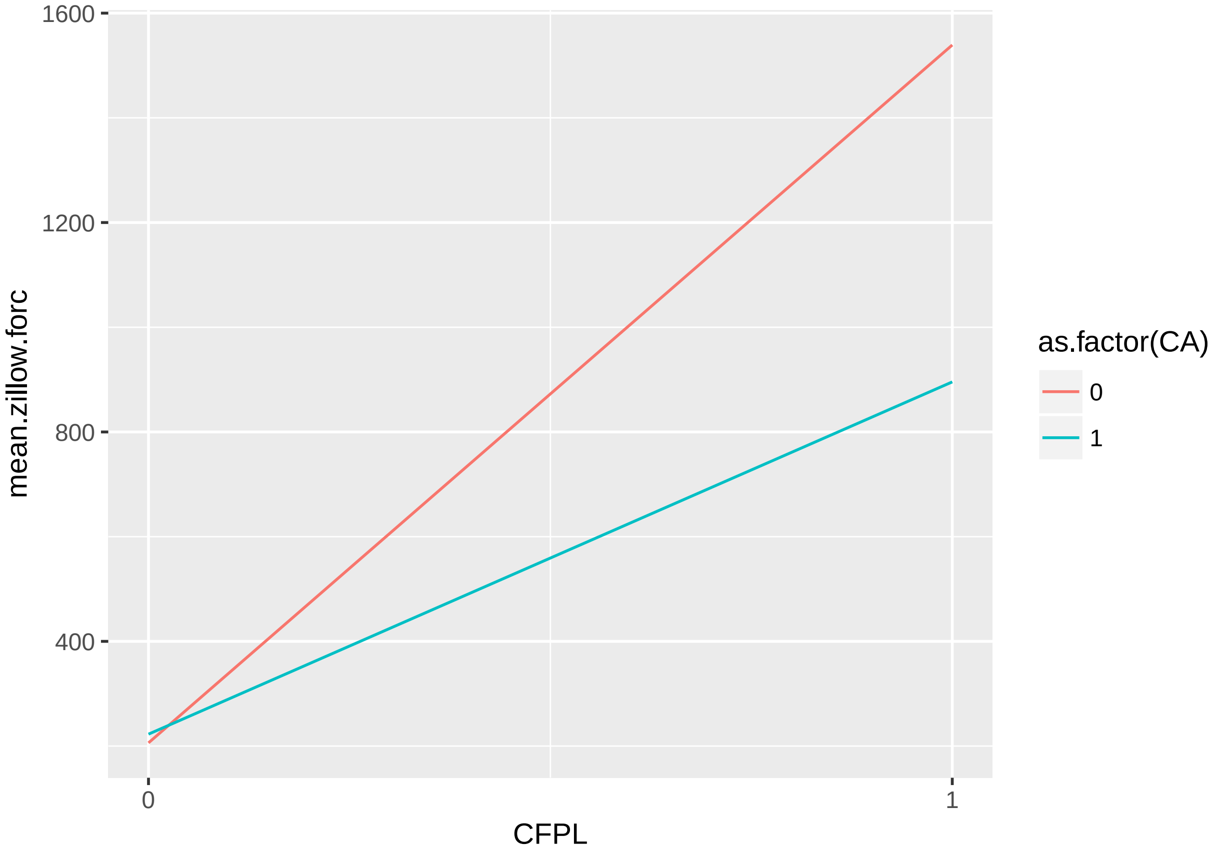
Notice here that we restrict the sample to only include the Sand States as California is a Sand State and the other Sand States (Arizona and Nevada) represent our control group. We also use the weighted mean (weighting by the number of households in 2000) as counties have notably different populations.

For California counties during the pre-treatment period (`CA = 0` and `CFPL = 0`), average foreclosures were `222.7555`. This is similar to the pre-treatment non-CA mean foreclosure estimate: `206.1253`.

During the CFPL treatment period (`CFPL = 1`), however, the mean foreclosure estimate was substantially lower for California counties versus non-California counties (`895.6155` when `CFPL = 0 & CA = 1` vs. `1539.1376` when `CFPL = 1 & CA = 0`). After the implementation of the CFPLs, California mean foreclosures increased to `895.6155` (`CFPL = 1 & CA = 1`), but the non-California mean foreclosures increased to `1539.1376` (`CFPL = 1 & CA = 0`). Even though foreclosures increased in California, they increased by substantially less than compared to the other Sand States. This is the key to the DD estimator: We compare the *change* in the treatment group to the *change* in the control group, hence a "difference-in-differences".

We can plot this data using `ggplot2`. Note that we turn `CA` to a factor so that `ggplot2` classifies `CA` as two different groups:

```
ggplot(data = dd.means, aes(x = CFPL, y = mean.zillow.forc)) +
    geom_line(aes(color = as.factor(CA))) +
    scale_x_continuous(breaks = c(0, 1))
```



From the graph, we see the following points that summarize our DD analysis thus far:

1. In the pre-CFPL period (`CFPL = 0`), average foreclosures in both the treatment group (`CA = 1`) and control group (`CA = 0`) are similar. Thus, the treatment and control groups are similar during the pre-treatment period. In the jargon of DD studies, we say that "the parallel pre-trends assumption is satisfied". The parallel pre-trends assumption, the key assumption for DD studies, says that during the pre-treatment period (in our case here during the pre-CFPL period), that the treatment and control groups only differ by a constant, meaning that their pre-treatment trends are "parallel".
2. Foreclosures increased *both* in California and the other Sand States increase following the implementation of the CFPLs. Thus, if we just looked at California, it would appear as if the CFPLs were ineffective at lowering foreclosures
3. The increase in mean California foreclosures was much smaller than the increase in mean foreclosures for the other states. Thus, if we use the other Sand States as a counterfactual (the path for California in the absence of the policy), the plots shows that the CFPLs noticeably reduced foreclosures in California.

Using the above formula for the means DD estimator, we have:

$$\hat{\delta} = (895.6155 - 222.7555) - (1539.1376 - 206.1253) = -660.1523$$

$\hat{\delta} = -660.1523$ is our DD and means that there were 660.1523 fewer foreclosures in California due to the CFPLs. We discuss statistical significance below

Difference-in-differences (DD) means estimate using regression

We can also obtain our DD estimate and assess our parallel pre-trends assumption through a regression. The advantage of the regression is that it simultaneously outputs the estimates of interest as well as their standard errors (for statistical significance). Note that we'll only use data from the Sand States (which contain both CA, or treatment, and our controls, AZ and NV) and use White standard errors to correct for any potential heteroskedasticity in our data.

```
lm(zillow.forc ~ CA * CFPL, data = DT[sand.state == 1], weights = hh2000) %>%
    coeftest(vcov = sandwich)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   206.125     35.899  5.7418 6.135e-08 ***
## CA             16.630     43.551  0.3819 0.7031809
## CFPL         1333.012    160.728  8.2936 1.110e-13 ***
## CA:CFPL      -660.152    182.939 -3.6086 0.0004357 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The regression output shows us all of the requisite DD output. As `CA`, `CFPL`, and `CA:CFPL` are all indicator variables, we can easily connect the output from the regression to our DD means estimates above:

- $\bar{Y}_{\text{CA = 0, CFPL = 0}} = \texttt{(Intercept)} = \texttt{206.125}$
- $\bar{Y}_{\text{CA = 1, CFPL = 0}} = \texttt{(Intercept)} + \texttt{CA} = \texttt{206.125} + \texttt{16.630} = \texttt{222.755}$
- $\bar{Y}_{\text{CA = 0, CFPL = 1}} = \texttt{(Intercept)} + \texttt{CFPL} = \texttt{206.125} + \texttt{1333.012} = \texttt{1539.137}$
- $\bar{Y}_{\text{CA = 1, CFPL = 1}} = \texttt{(Intercept)} + \texttt{CA} + \texttt{CFPL} + \texttt{CA:CFPL} = \texttt{206.125} + \texttt{16.630} + \texttt{1333.012} + \texttt{-660.152} = \texttt{1539.137} = \texttt{895.615}$

As an exercise, check that this output matches the output from `dd.means`.

Here's what else we learn from the above regression output:

1. The parallel pre-trends assumption is satisfied as there is not a statistically significant difference in pre-treatment foreclosures between California and the other Sand States (the coefficient on `CA` is small and insignificant.
2. The DD coefficient, `CA:CFPL`, the interaction of the `CA` and `CFPL` indicators is large in magnitude and statistically significant. Thus there is a statistically significant difference in foreclosures between California counties and the controls during the CFPL period
3. `-660.152` is the DD estimate and it is statistically significant with a t-statistic of `-3.61`

Other Difference-in-differences estimators

We can leverage the `forc.high` variable and create other DD estimators.

First, we can look at "high" foreclosure counties in California versus "low" foreclosure counties. The advantage of this estimator is that it will account for California macro-level trends (e.g. an economic shock that affects the whole state).

Here is the regression output. Note that we only use California counties.

```
lm(zillow.forc ~ CFPL * forc.high, data = DT[CA == 1], weights = hh2000) %>%
    coeftest(sandwich)
```

```
##
```

```
## t test of coefficients:
##
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      91.469     10.461  8.7438 7.407e-14 ***
## CFPL            446.591     55.797  8.0039 2.788e-12 ***
## forc.high       158.194     34.220  4.6228 1.179e-05 ***
## CFPL:forc.high  272.642    126.622  2.1532   0.03381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient on `forc.high` is significant, indicating that during the pre-CFPL period, that high foreclosure counties experienced more foreclosures than low foreclosure counties (2.73 times more foreclosures, how did I get that number?). This makes sense, but suggests that our parallel pre-trends assumption is violated.

Further, the coefficient on `CFPL:forc.high` is positive and significant, suggesting foreclosures in high foreclosure counties increased more than in low foreclosure counties. It's hard to know what this means in terms of the CFPL efficacy as parallel pre-trends assumption is violated.

We can construct a third DD estimate by comparing high foreclosure counties in California to high foreclosure counties in other states. Note here that we subset the data to `forc.high == 1`. This yields all "high" foreclosure states across the country

```
lm(zillow.forc ~ CA * CFPL, data = DT[forc.high == 1], weights = hh2000) %>%
    coeftest(sandwich)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  245.262     35.145  6.9785 9.639e-10 ***
## CA             4.401     47.925  0.0918   0.92707
## CFPL        1290.082    213.687  6.0373 5.352e-08 ***
## CA:CFPL     -570.848    242.037 -2.3585   0.02092 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output from this regression is rather encouraging: The parallel pre-trends assumption is satisfied (the coefficient on `CA` is small and insignificant), and the DD estimate is negative and significant. The DD estimate tells us that foreclosures were 570.848 lower per 10,000 homes in high foreclosure California counties, relative to high foreclosure counties in other states.

Difference-in-difference-in-differences (DDD)

The above DD estimates are appealing, but none account for within California macro-level trends while satisfying the parallel pre-trends assumption. Here, we're going to explore a comprehensive DDD estimate.

Essentially, the DDD estimator is going to (1) compare the change in means between high and low foreclosure counties within California (first difference); (2) compare the change in means between high and low foreclosure counties outside of California (second difference); and (3) take the difference of the first two differences. In essence, the DDD estimator is the difference in two difference-in-differences estimators; hence the name "Difference-in-difference-in-differences"

To see how the DDD estimator works, let's calculate the three differences separately:

First Difference:

The first difference is the *within* California difference between high a low foreclosure counties:

```r
lm(zillow.forc ~ CFPL * forc.high, data = DT[CA == 1], weights = hh2000) %>%
    coeftest(sandwich)
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      91.469     10.461  8.7438 7.407e-14 ***
## CFPL            446.591     55.797  8.0039 2.788e-12 ***
## forc.high       158.194     34.220  4.6228 1.179e-05 ***
## CFPL:forc.high  272.642    126.622  2.1532   0.03381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that this was one of our DD estimators from above. The within California DD estimate is `272.642`.

Second Difference:

The second difference is difference between high and low foreclosure counties for counties *outside* California

```r
lm(zillow.forc ~ CFPL * forc.high, data = DT[CA == 0], weights = hh2000) %>%
    coeftest(sandwich)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      83.8746     8.3008 10.1045 < 2.2e-16 ***
## CFPL            261.0448    25.4928 10.2399 < 2.2e-16 ***
## forc.high       161.3871    36.1123  4.4690 9.203e-06 ***
## CFPL:forc.high 1029.0368   215.2021  4.7817 2.132e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second difference is thus `1029.0368`

Third Difference (DDD estimate):

Our third difference is the difference in our first two differences: `272.642272.642` − `1029.0368` = `-756.3948`.

This is our DDD estimate and yields the estimate of the CFPLs on high foreclosure counties in California, netting out changes in low foreclosure California counties and change in high foreclosure counties in other states relative to low foreclosure counties in other states.

The DDD Regression:

We can confirm our estimate using a full DDD regression. Note here that we use the full dataset and that `CA` × `forc.high` is the treated group and all other counties (even low counties within California) are controls. The assumption behind identification here is that CFPL foreclosure prevention policies should have an outsized impact on high foreclosure counties.

```r
#We're going to save the model b/c we'll need it later
ddd.mod <- lm(zillow.forc ~ CFPL * forc.high * CA, data = DT, weights = hh2000)
ddd.mod %>% coeftest(sandwich)
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          83.8746     8.3008 10.1045 < 2.2e-16 ***
```

```
## CFPL               261.0448    25.4928 10.2399 < 2.2e-16 ***
## forc.high          161.3871    36.1123  4.4690 9.027e-06 ***
## CA                   7.5941    13.3542  0.5687  0.569747
## CFPL:forc.high    1029.0368   215.2021  4.7817 2.080e-06 ***
## CFPL:CA            185.5466    61.3446  3.0247  0.002571 **
## forc.high:CA        -3.1931    49.7505 -0.0642  0.948842
## CFPL:forc.high:CA -756.3945   249.6901 -3.0293  0.002532 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A couple of things to note about the DDD regression:

1. We can assess the parallel pre-trends assumption: During the pre-treatment period, the mean of foreclosures for non-California, high foreclosure counties is `(Intercept) + forc.high`; while the pre-treatment mean of foreclosures for high foreclosure California counties is `(Intercept) + forc.high + CA + forc.high:CA`. The difference, and what we need to test for the parallel pre-trends, is H_0: `CA + forc.high:CA = 0`. The F-statistics for null that `CA + forc.high:CA = 0` is insignificant:

```
linearHypothesis(ddd.mod, "CA + forc.high:CA = 0", vcov = sandwich)
```

```
## Linear hypothesis test
##
## Hypothesis:
## CA  + forc.high:CA = 0
##
## Model 1: restricted model
## Model 2: zillow.forc ~ CFPL * forc.high * CA
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    777
## 2    776  1 0.0084 0.9269
```

2. Also notice in the above regression that the coefficient on `CA` is insignificant. Thus, low foreclosure California counties are not different from low foreclosure counties in other states during the pre-CFPL period.
3. The DDD estimate is `-756.3945` and statistically significant at the 1 percent level. This estimate means that foreclosures per 10,000 homes in high foreclosure, California counties fell by -756.3945 due to the CFPLs
4. Generally, in both DD and DDD research designs, you can add extra control variables to your regression if "randomization is based on covariates" (e.g. the parallel pre-trends hypothesis is satisfied after controlling for certain variables) or to lower standard errors (recall that if a covariate predicts the dependent variable and is included in the regression, then the variance of the residuals will be lower and the standard errors will fall)
5. See Gabriel et al. (2017) for a panel data and time-varying setup of the DDD approach.