Marijuana Legalization and Violent Crime_02MAR20

March 4, 2020

0.1 Marijuana Legalization and Violent Crime

0.2 Exercise 1

What is a unit of observation (a single row) in this data? What entities are being tracked, and over what time period?

[48]:
```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
```

[49]:
```
url = 'https://media.githubusercontent.com/media/nickeubank/MIDS_Data/master/
↪UDS_arrest_data.csv'
df = pd.read_csv(url)
```

[50]:
```
df.head()
```

[50]:

	YEAR	COUNTY	VIOLENT	F_DRUGOFF	total_population
0	1980	Alameda County	4504	3569	1105379.0
1	1981	Alameda County	4699	3926	1122759.3
2	1982	Alameda County	4389	4436	1140139.6
3	1983	Alameda County	4500	5086	1157519.9
4	1984	Alameda County	3714	5878	1174900.2

[51]:
```
df.describe()
```

[51]:

	YEAR	VIOLENT	F_DRUGOFF	total_population
count	2262.000000	2262.000000	2262.000000	2.262000e+03
mean	1999.000000	2084.767462	2063.761273	5.611930e+05
std	11.257117	5695.691031	5816.478130	1.297315e+06
min	1980.000000	1.000000	0.000000	1.097000e+03
25%	1989.000000	145.000000	109.000000	4.208128e+04
50%	1999.000000	516.000000	399.500000	1.525506e+05
75%	2009.000000	1925.500000	1498.250000	5.144530e+05
max	2018.000000	64790.000000	69667.000000	9.818605e+06

Answer: Each row is a county by year, we are tracking violent arrests, felony drug arrests, and the total population in annual records from 1980 to 2018.

0.3 Exercise 2:

Calculate each county's average drug arrest rate for the period from 2007-2009. Then calculate the median value across counties, and create an indicator called treated for counties with above-median average drug arrest rates during this period. Note that this indicator should be time-invariant – if a county is in the treated group, it should always be identified as being in the treated group.

[52]:
```
#Subset period 2007 - 2009
df_subset_pre = df[df['YEAR'].isin([2007,2008,2009])].reset_index(drop=True)
# https://cmdlinetips.com/2018/02/
 ↪how-to-subset-pandas-dataframe-based-on-values-of-a-column/
```

[53]:
```
#Calculate drug arrest rate btw 2007 - 2009
df_subset_pre['Drug_arrest_rate'] = df_subset_pre['F_DRUGOFF']/
 ↪df_subset_pre['total_population']
```

[54]:
```
#Calculate average drug arrest rate by county
df_subset_pre['Drug_arrest_rate_mean'] = df_subset_pre.groupby(['COUNTY'],␣
 ↪as_index = False)['Drug_arrest_rate'].transform('mean')

#Calculate median in the average drug arrest rate in pre_treatment period␣
 ↪across all counties
df_subset_pre.Drug_arrest_rate_mean.median()

print(f' The median average drug arrest rate in pre_treatment period across all␣
 ↪counties is : {df_subset_pre.Drug_arrest_rate_mean.median():.5f}')
```

```
 The median average drug arrest rate in pre_treatment period across all counties
is : 0.00302
```

[55]:
```
df_subset_pre['Treated'] = 0
list=[]
i=0
for l in df_subset_pre['COUNTY']:
    check_median =␣
 ↪df_subset_pre['Drug_arrest_rate_mean'][df_subset_pre['COUNTY']==l]
    if any (np.array(check_median) > df_subset_pre.Drug_arrest_rate.median()):
        list.append(1)
        i+=1
    else:
        list.append(0)
        i+=1

df_subset_pre['Treated'] = list
```

[56]:
```
df_subset_pre['Treated'].value_counts()
```

[56]:
```
1    87
0    87
Name: Treated, dtype: int64
```

0.4 Exercise 3:

Our outcome in this analysis is the violent arrest rate – if drug liberalization reduces crime overall, we would expect to see this rate fall in counties with high drug arrest rates after liberalization; if not, we would not expect to see any changes. Create a violent_rate variable with is violent arrests per 100,000 people.

[57]:
```
df_subset_pre['Violent_arrest_per100k'] = (df_subset_pre['VIOLENT']/
 ↪df_subset_pre['total_population'])*100000
df_subset_pre
```

[57]:

	YEAR	COUNTY	VIOLENT	F_DRUGOFF	total_population	\
0	2007	Alameda County	4443	6071	1490312.0	
1	2008	Alameda County	4336	5893	1496965.0	
2	2009	Alameda County	4318	5749	1503618.0	
3	2007	Alpine County	8	1	1184.9	
4	2008	Alpine County	4	4	1181.6	
..	…	…	…	…	…	
169	2008	Yolo County	587	632	194411.2	
170	2009	Yolo County	585	614	197630.1	
171	2007	Yuba County	416	309	68574.2	
172	2008	Yuba County	375	214	69767.8	
173	2009	Yuba County	354	211	70961.4	

	Drug_arrest_rate	Drug_arrest_rate_mean	Treated	Violent_arrest_per100k
0	0.004074	0.003945	1	298.125493
1	0.003937	0.003945	1	289.652731
2	0.003823	0.003945	1	287.174003
3	0.000844	0.001976	0	675.162461
4	0.003385	0.001976	0	338.524035
..	…	…	…	…
169	0.003251	0.003338	1	301.937337
170	0.003107	0.003338	1	296.007541
171	0.004506	0.003516	1	606.642148
172	0.003067	0.003516	1	537.497241
173	0.002973	0.003516	1	498.862762

[174 rows x 9 columns]

0.5 Exercise 4:

Calculate (a) the change in violent arrest rates for our treated groups from before legalization to after ($\bar{y}_{T=1,Post} - \bar{y}_{T=1,Pre}$), and (b) our difference in difference estimator $\hat{}$ by calculating these four values. Does doing your difference-in-difference estimate tell you something different from what you'd learn if you had just done a pre-post comparison?

Answer: The outcome from the difference in difference allow us to obtain more precise information regarding the actual effect of the treatment between the treatment and control groups. By only calculating the pre-post average across all groups we would not be able to estimate the impact of the treatment effect, as we will review it in the following questions.

[58]:
```
df_treated_pre = df_subset_pre[df_subset_pre['Treated']==1]
df_control_pre = df_subset_pre[df_subset_pre['Treated']==0]
```

[59]:
```
df_subset_post = df[df['YEAR'].isin([2016,2017,2018])].reset_index(drop=True)
df_subset_post['Treated'] = list
df_subset_post['Violent_arrest_per100k'] = (df_subset_post['VIOLENT']/
 ↪df_subset_post['total_population'])*100000
df_treated_post = df_subset_post[df_subset_post['Treated']==1]
df_control_post = df_subset_post[df_subset_post['Treated']==0]
```

[60]:
```
Diff_treated = df_treated_post.Violent_arrest_per100k.mean() - df_treated_pre.
 ↪Violent_arrest_per100k.mean()
```

[61]:
```
Diff_control = df_control_post.Violent_arrest_per100k.mean() - df_control_pre.
 ↪Violent_arrest_per100k.mean()
```

[62]:
```
print(f' The change in violent arrest rates for our treated group from before␣
 ↪legalization to after is:{Diff_treated:.2f}')
print(f' The change in violent arrest rates for our control group from before␣
 ↪legalization to after is:{Diff_control:.2f}')
print(f' The our difference in difference estimator ^ is:{(Diff_treated -␣
 ↪Diff_control):.2f}')
```

```
 The change in violent arrest rates for our treated group from before
legalization to after is:-26.80
 The change in violent arrest rates for our control group from before
legalization to after is:-19.38
 The our difference in difference estimator ^ is:-7.42
```

0.6 Exercise 5:

Now calculate $\hat{}$ using a regression with an indicator for post-2010, an indicator for treated, and an interaction of the two. Use only the same set of years you used above. How does your estimate compare to the estimate you calculated in Exercise 4?

Answer: From our model regression model, the (difference in difference estimator) haven't changed at all (the coefficient on the interaction term C(Treated)*C(post_2010), is -7.42).

What does this tell you about interpretation of interaction terms with two indicator variables?

Answer: As we are clustering our observations we should be aware that the size of the standard error for the interaction effect is 20.658, which makes this element statistically not significant. The same occurs with the post-2010 indicator. Observe that zero is included in the confidence interval for both elements.

The 'treatment' indicator is statistically significant, meaning that on average, treated counties have -21.07 units in their violent arrest rate for each 100,000 people.

[63]:
```
treated_counties = pd.
 ↪DataFrame(df_subset_pre[df_subset_pre['Treated']==1])['COUNTY'].unique()
treated_counties
```

[63]:
```
array(['Alameda County', 'Calaveras County', 'Del Norte County',
       'Fresno County', 'Glenn County', 'Humboldt County',
       'Imperial County', 'Inyo County', 'Kern County', 'Lake County',
       'Los Angeles County', 'Mendocino County', 'Merced County',
       'Plumas County', 'Sacramento County', 'San Bernardino County',
       'San Francisco County', 'San Joaquin County', 'Santa Cruz County',
       'Siskiyou County', 'Solano County', 'Sonoma County',
       'Stanislaus County', 'Tehama County', 'Trinity County',
       'Tulare County', 'Tuolumne County', 'Yolo County', 'Yuba County'],
      dtype=object)
```

[64]:
```
df['Treated'] = 0
df['Treated'].loc[df['COUNTY'].isin(treated_counties)] = 1
```

[65]:
```
df['post_2010'] = 0
df['post_2010'].loc[df['YEAR']>2010] = 1
df_model_subset = df[df['YEAR'].isin([2007,2008,2009, 2016,2017,2018])].
 ↪reset_index(drop=True)
```

[66]:
```
df_model_subset['Violent_arrest_per100k'] = (df_model_subset['VIOLENT']/
 ↪df_model_subset['total_population'])*100000
```

[67]:
```
# Regression model with an indicator for post-2010, an indicator for treated,␣
 ↪and an interaction of the two
model_q5 = smf.ols('Violent_arrest_per100k ~ C(COUNTY) + YEAR +␣
 ↪C(Treated)*C(post_2010)', df_model_subset).fit()
model_q5.get_robustcov_results(cov_type='cluster', groups =␣
 ↪df_model_subset['COUNTY']).summary()
```

[67]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:     Violent_arrest_per100k   R-squared:
0.795
Model:                               OLS   Adj. R-squared:
0.752
Method:                    Least Squares   F-statistic:
2.454
Date:                   Wed, 04 Mar 2020   Prob (F-statistic):
0.0724
Time:                           08:23:34   Log-Likelihood:
-1861.6
No. Observations:                    348   AIC:
3845.
Df Residuals:                        287   BIC:
4080.
Df Model:                             60
Covariance Type:                  cluster
==============================================================================
=======================
                                    coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
-----------------------
Intercept                         1.284e+04   6600.846      1.946      0.057
-374.138    2.61e+04
C(COUNTY)[T.Alpine County]         585.6198    220.042      2.661      0.010
144.994    1026.246
C(COUNTY)[T.Amador County]         340.6325    220.042      1.548      0.127
-99.993     781.258
C(COUNTY)[T.Butte County]          433.7830    220.042      1.971      0.054
-6.843     874.409
C(COUNTY)[T.Calaveras County]      132.3930      4e-10   3.31e+11      0.000
132.393     132.393
C(COUNTY)[T.Colusa County]         460.6208    220.042      2.093      0.041
19.995     901.247
C(COUNTY)[T.Contra Costa County]   379.4583    220.042      1.724      0.090
-61.167     820.084
C(COUNTY)[T.Del Norte County]      284.4013   4.05e-10   7.01e+11      0.000
284.401     284.401
C(COUNTY)[T.El Dorado County]      409.4839    220.042      1.861      0.068
-31.142     850.110
C(COUNTY)[T.Fresno County]          207.0049   4.07e-10   5.09e+11      0.000
207.005     207.005
C(COUNTY)[T.Glenn County]          115.3571   4.09e-10   2.82e+11      0.000
115.357     115.357
```

```
C(COUNTY)[T.Humboldt County]          86.4667   4.03e-10   2.15e+11      0.000
86.467      86.467
C(COUNTY)[T.Imperial County]          38.2888   4.04e-10   9.47e+10      0.000
38.289      38.289
C(COUNTY)[T.Inyo County]             209.7831   4.03e-10    5.2e+11      0.000
209.783     209.783
C(COUNTY)[T.Kern County]             198.9596   4.05e-10   4.91e+11      0.000
198.960     198.960
C(COUNTY)[T.Kings County]            461.6912    220.042      2.098      0.040
21.065     902.317
C(COUNTY)[T.Lake County]             220.2953   3.97e-10   5.55e+11      0.000
220.295     220.295
C(COUNTY)[T.Lassen County]           430.0203    220.042      1.954      0.056
-10.606     870.646
C(COUNTY)[T.Los Angeles County]       55.5320   4.05e-10   1.37e+11      0.000
55.532      55.532
C(COUNTY)[T.Madera County]           458.9347    220.042      2.086      0.041
18.309     899.560
C(COUNTY)[T.Marin County]            320.0740    220.042      1.455      0.151
-120.552     760.700
C(COUNTY)[T.Mariposa County]         449.5044    220.042      2.043      0.046
8.879     890.130
C(COUNTY)[T.Mendocino County]        189.1406   4.07e-10   4.64e+11      0.000
189.141     189.141
C(COUNTY)[T.Merced County]           175.2687   4.06e-10   4.32e+11      0.000
175.269     175.269
C(COUNTY)[T.Modoc County]            664.1660    220.042      3.018      0.004
223.540    1104.792
C(COUNTY)[T.Mono County]             488.6947    220.042      2.221      0.030
48.069     929.321
C(COUNTY)[T.Monterey County]         444.8322    220.042      2.022      0.048
4.206     885.458
C(COUNTY)[T.Napa County]             386.9274    220.042      1.758      0.084
-53.698     827.553
C(COUNTY)[T.Nevada County]           333.7059    220.042      1.517      0.135
-106.920     774.332
C(COUNTY)[T.Orange County]           322.8686    220.042      1.467      0.148
-117.757     763.494
C(COUNTY)[T.Placer County]           340.8199    220.042      1.549      0.127
-99.806     781.446
C(COUNTY)[T.Plumas County]           197.9096   4.04e-10    4.9e+11      0.000
197.910     197.910
C(COUNTY)[T.Riverside County]        389.2182    220.042      1.769      0.082
-51.408     829.844
C(COUNTY)[T.Sacramento County]       110.8686   4.06e-10   2.73e+11      0.000
110.869     110.869
C(COUNTY)[T.San Benito County]       499.9849    220.042      2.272      0.027
```

```
59.359     940.611
C(COUNTY)[T.San Bernardino County]    168.1644   4.06e-10   4.15e+11      0.000
168.164     168.164
C(COUNTY)[T.San Diego County]         423.7505    220.042      1.926      0.059
-16.875     864.376
C(COUNTY)[T.San Francisco County]      84.2993   4.04e-10   2.08e+11      0.000
84.299      84.299
C(COUNTY)[T.San Joaquin County]       187.7026   4.04e-10   4.65e+11      0.000
187.703     187.703
C(COUNTY)[T.San Luis Obispo County]   354.9974    220.042      1.613      0.112
-85.628     795.623
C(COUNTY)[T.San Mateo County]         311.0308    220.042      1.414      0.163
-129.595     751.657
C(COUNTY)[T.Santa Barbara County]     421.8716    220.042      1.917      0.060
-18.754     862.497
C(COUNTY)[T.Santa Clara County]       362.4240    220.042      1.647      0.105
-78.202     803.050
C(COUNTY)[T.Santa Cruz County]          9.8490   4.02e-10   2.45e+10      0.000
9.849       9.849
C(COUNTY)[T.Shasta County]            406.3349    220.042      1.847      0.070
-34.291     846.961
C(COUNTY)[T.Sierra County]            588.1148    220.042      2.673      0.010
147.489    1028.741
C(COUNTY)[T.Siskiyou County]          122.3128   4.03e-10   3.04e+11      0.000
122.313     122.313
C(COUNTY)[T.Solano County]            128.8165   4.08e-10   3.16e+11      0.000
128.817     128.817
C(COUNTY)[T.Sonoma County]             21.3606   4.04e-10   5.29e+10      0.000
21.361      21.361
C(COUNTY)[T.Stanislaus County]        186.3526   4.07e-10   4.58e+11      0.000
186.353     186.353
C(COUNTY)[T.Sutter County]            585.0921    220.042      2.659      0.010
144.466    1025.718
C(COUNTY)[T.Tehama County]             93.1865   4.08e-10   2.28e+11      0.000
93.187      93.187
C(COUNTY)[T.Trinity County]           214.3527   4.04e-10   5.31e+11      0.000
214.353     214.353
C(COUNTY)[T.Tulare County]            232.5847   4.04e-10   5.76e+11      0.000
232.585     232.585
C(COUNTY)[T.Tuolumne County]           35.0323   4.02e-10   8.71e+10      0.000
35.032      35.032
C(COUNTY)[T.Ventura County]           392.3963    220.042      1.783      0.080
-48.229     833.022
C(COUNTY)[T.Yolo County]               22.1551   4.07e-10   5.44e+10      0.000
22.155      22.155
C(COUNTY)[T.Yuba County]              310.7904   4.11e-10   7.57e+11      0.000
310.790     310.790
```

```
C(Treated)[T.1]                       396.7746    219.873      1.805      0.076
-43.514     837.063
C(post_2010)[T.1]                      38.6758     31.398      1.232      0.223
-24.198     101.549
C(Treated)[T.1]:C(post_2010)[T.1]      -7.4181     20.694     -0.358      0.721
-48.858      34.021
YEAR                                   -6.4508      3.397     -1.899      0.063
-13.252       0.351
==============================================================================
Omnibus:                       36.200   Durbin-Watson:                   1.756
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              169.554
Skew:                           0.224   Prob(JB):                     1.52e-37
Kurtosis:                       6.390   Cond. No.                     7.33e+18
==============================================================================

Warnings:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The smallest eigenvalue is 2.62e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
"""
```

0.7 Exercise 6:

Plot a difference-in-difference model using data from 2000-2009 (inclusive) and from 2016-2018 (inclusive). Note this will have four different geometric components: a time trend for treated counties pre-2010, a time trend for control counties pre-2010, a time trend for treated counties post-2016 (include 2016), and a time trend for control counties post-2016 (include 2016).

Do you see evidence of parallel trends for these two datasets? Does that make you feel more or less confident in your diff-in-diff estimates?

Answer: We observe both groups follow the same parallel trend pre and post 2010. In both cases, the rate was decreasing before 2010 and between 2016 and 2018 both are increasing again. This suggest that the effect of policy on drugs was not relevant as observed by the p_values previously reported.

[68]:
```
#Setting Treated variable as categorical
df['Treated'].dtype
df.Treated = df.Treated.astype('category')
```

[69]:
```
df['Violent_arrest_per100k'] = (df['VIOLENT']/df['total_population'])*100000
df_plot = df.
 ↪loc[(df['YEAR']>=2000)&(df['YEAR']<=2009)|(df['YEAR']>=2016)&(df['YEAR']<=2018)]
df_plot_treated = df_plot[df_plot['Treated']==1]
df_plot_control = df_plot[df_plot['Treated']==0]
```

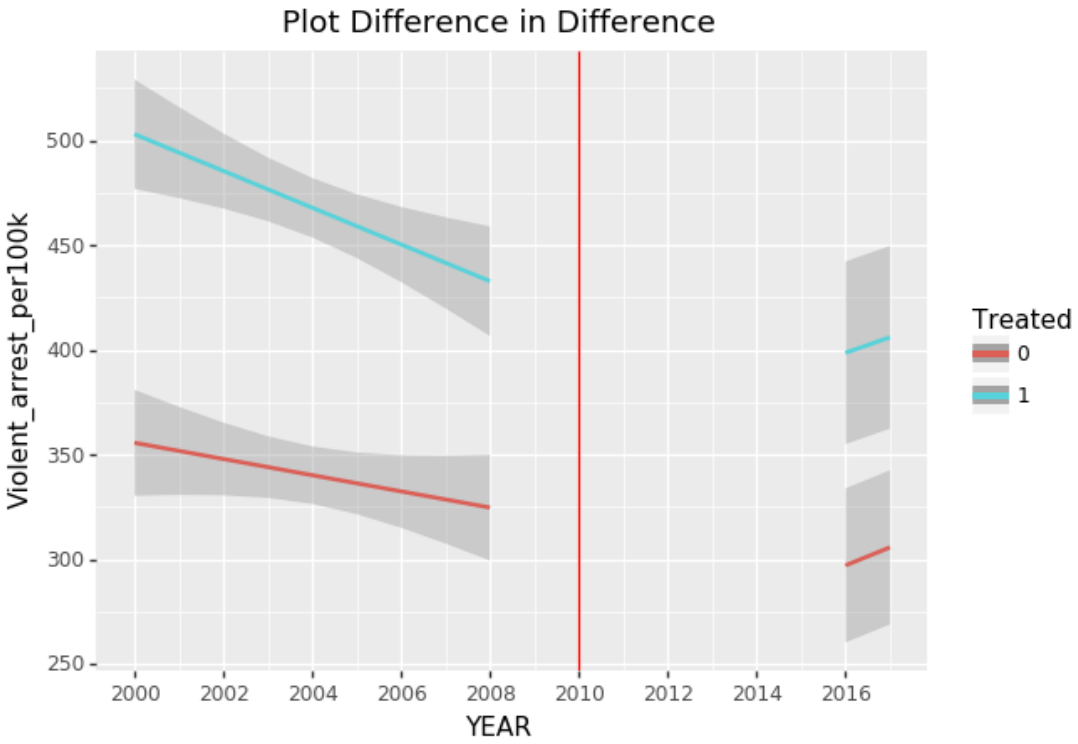
[70]:
```
from plotnine import *

policy_year = 2010

# Define x-axis
start_year = df_plot['YEAR'].min()
end_year = df_plot['YEAR'].max()
bef = range(start_year, policy_year-1)
aft = range(policy_year+6, end_year)
#Treated subset
df_treated_bef = df_plot_treated.copy()
df_treated_bef = df_treated_bef[df_treated_bef['YEAR'].isin(bef)]
df_treated_aft = df_plot_treated.copy()
df_treated_aft = df_treated_aft[df_treated_aft['YEAR'].isin(aft)]
#Control subset
df_control_bef = df_plot_control.copy()
df_control_bef = df_control_bef[df_control_bef['YEAR'].isin(bef)]
df_control_aft = df_plot_control.copy()
df_control_aft = df_control_aft[df_control_aft['YEAR'].isin(aft)]

# Plot data
plot = (ggplot() +
       geom_smooth(df_treated_bef, aes(x = 'YEAR', y =␣
 ↪'Violent_arrest_per100k', color = 'Treated'), method = 'lm', level = 0.95) +
       geom_smooth(df_treated_aft, aes(x = 'YEAR', y =␣
 ↪'Violent_arrest_per100k', color = 'Treated'), method = 'lm', level = 0.95) +
       geom_smooth(df_control_bef, aes(x = 'YEAR', y =␣
 ↪'Violent_arrest_per100k', color = 'Treated'), method = 'lm', level = 0.95) +
       geom_smooth(df_control_aft, aes(x = 'YEAR', y =␣
 ↪'Violent_arrest_per100k', color = 'Treated'), method = 'lm', level = 0.95) +
       scale_x_continuous(breaks = range(start_year, end_year, 2)) +
       ggtitle("Plot Difference in Difference") +
       geom_vline(xintercept = policy_year, color = 'red'))
```

[71]:
```
plot
```



[71]: <ggplot: (-9223371857170095036)>

0.8 Exercise 7

While we can estimate the model described above precisely as a regression, it's actually much easier to estimate a more flexible model by running the regression we ran in Exercise 5 but with both county and year fixed effects. Use PanelOLS (or lfe in R) to estimate this fixed effects regression.

With all these additional fixed effects, do you find evidence that marijuana legalization reduced violent crime?

[72]:
```
from linearmodels import PanelOLS
# Regression model with an indicator for post-2010, an indicator for treated,␣
 ↪and an interaction of the two

# Set multiindex from dataframe used in exercise 5
df_w_multiindex = df_model_subset.set_index(['COUNTY','YEAR'])

# Create model
model_q7 = PanelOLS.
 ↪from_formula('Violent_arrest_per100k~Treated*post_2010+EntityEffects', data␣
 ↪= df_w_multiindex,
                           drop_absorbed = True)
model_q7.fit(cov_type='clustered', cluster_entity=True)
```

[72]:

```
                            PanelOLS Estimation Summary
================================================================================
====
Dep. Variable:     Violent_arrest_per100k   R-squared:
0.0496
Estimator:                       PanelOLS   R-squared (Between):
-0.0617
No. Observations:                     348   R-squared (Within):
0.0496
Date:                    Wed, Mar 04 2020   R-squared (Overall):
-0.0595
Time:                            08:23:39   Log-likelihood
-1863.4
Cov. Estimator:                 Clustered
                                            F-statistic:
7.5117
Entities:                               58   P-value
0.0007
Avg Obs:                            6.0000   Distribution:
F(2,288)
Min Obs:                            6.0000
Max Obs:                            6.0000   F-statistic (robust):
3.3784
                                            P-value
0.0355
Time periods:                            6   Distribution:
F(2,288)
Avg Obs:                           58.000
Min Obs:                           58.000
Max Obs:                           58.000

                              Parameter Estimates
================================================================================
=====
                   Parameter  Std. Err.     T-stat    P-value    Lower CI
Upper CI
--------------------------------------------------------------------------------
-----
post_2010            -19.382     9.7916    -1.9794     0.0487     -38.654
-0.1095
Treated:post_2010    -7.4181     18.679    -0.3971     0.6916     -44.182
29.346
================================================================================
=====

F-test for Poolability: 16.717
P-value: 0.0000
```

```
Distribution: F(57,288)

Included effects: Entity
PanelEffectsResults, id: 0x29d60da9f08
```

Answer: The regression analysis in question 7 results in the same coefficients for the interaction term (Treated*post_2010) in question 5. Also, we confirm that there is no statistical significance of this term as the p_value is 0.69. This is a confirmation of the graph we saw in question 6 showing the parallel behavior in pre and post analysis of the violent crime rate across treated and controlled counties, even when you control for year and county.