5-16-2021-Panel and DiD in Python

- Name: Jikhan Jeong
- Ch 13. Using Python for Introductory Econometrics
- Ref: http://www.upfie.net/downloads.html
- Ref code: http://www.upfie.net/downloads13.html
- statsmodels package: https://www.statsmodels.org/stable/index.html (for DiD)
- linearmodels package: https://bashtage.github.io/linearmodels/ (for Panel dataset)
- Using **statsmodels**

```
# (if not installed) pip install wooldridge
import wooldridge as woo
```

```
import pandas as pd
import statsmodels.formula.api as smf
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.u
  import pandas.util.testing as tm
```

Before running code, please running user-defiend functions in the bottom of this code.

13.1 Pooled OLS

```
data = woo.dataWoo('cps78_85')
```

```
data_summary(data)
```

```
(1084, 15)
Index(['educ', 'south', 'nonwhite', 'female', 'married', 'exper', 'expersq',
```

```
reg = smf.ols('lwage ~y85*(educ+female) + exper + I((exper**2)/100) + union', data )
results = reg.fit()
results.summary()
```

OLS Regression Results

Dep. Variable:	lwage	**R-squared:**	0.426
Model:	OLS	**Adj. R-squared:**	0.422
Method:	Least Squares	**F-statistic:**	99.80
Date:	Mon, 17 May 2021	**Prob (F-statistic):**	4.46e-124
Time:	01:20:24	**Log-Likelihood:**	-574.24
No. Observations:	1084	**AIC:**	1166.
Df Residuals:	1075	**BIC:**	1211.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.4589	0.093	4.911	0.000	0.276	0.642
y85	0.1178	0.124	0.952	0.341	-0.125	0.361
educ	0.0747	0.007	11.192	0.000	0.062	0.088
female	-0.3167	0.037	-8.648	0.000	-0.389	-0.245
y85:educ	0.0185	0.009	1.974	0.049	0.000	0.037
y85:female	0.0851	0.051	1.658	0.098	-0.016	0.186
exper	0.0296	0.004	8.293	0.000	0.023	0.037
I((exper ** 2) / 100)	-0.0399	0.008	-5.151	0.000	-0.055	-0.025
union	0.2021	0.030	6.672	0.000	0.143	0.262

Omnibus:	83.747	**Durbin-Watson:**	1.918
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	317.985
Skew:	-0.271	**Prob(JB):**	8.92e-70
Kurtosis:	5.597	**Cond. No.**	296.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
stat_ols('lwage ~y85*(educ+female) + exper + I((exper**2)/100) + union', data)
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.426
Model:                            OLS   Adj. R-squared:                  0.422
Method:                 Least Squares   F-statistic:                     99.80
Date:                Mon, 17 May 2021   Prob (F-statistic):          4.46e-124
Time:                        01:35:56   Log-Likelihood:                -574.24
No. Observations:                1084   AIC:                             1166.
Df Residuals:                    1075   BIC:                             1211.
Df Model:                           8                                         
Covariance Type:            nonrobust                                         
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               0.4589      0.093      4.911      0.000       0.276       0.642
y85                     0.1178      0.124      0.952      0.341      -0.125       0.361
educ                    0.0747      0.007     11.192      0.000       0.062       0.088
female                 -0.3167      0.037     -8.648      0.000      -0.389      -0.245
y85:educ                0.0185      0.009      1.974      0.049       0.000       0.037
y85:female              0.0851      0.051      1.658      0.098      -0.016       0.186
exper                   0.0296      0.004      8.293      0.000       0.023       0.037
I((exper ** 2) / 100)  -0.0399      0.008     -5.151      0.000      -0.055      -0.025
union                   0.2021      0.030      6.672      0.000       0.143       0.262
==============================================================================
Omnibus:                       83.747   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              317.985
Skew:                          -0.271   Prob(JB):                     8.92e-70
Kurtosis:                       5.597   Cond. No.                         296.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7f3dcb8dc750>
```

```
dir(results)
```

```
 'centered_tss',
 'compare_f_test',
 'compare_lm_test',
 'compare_lr_test',
 'condition_number',
 'conf_int',
 'conf_int_el',
 'cov_HC0',
 'cov_HC1',
 'cov_HC2',
 'cov_HC3',
 'cov_kwds',
 'cov_params',
 'cov_type',
 'df_model',
 'df_resid',
 'diagn',
 'eigenvals',
 'el_test',
 'ess',
 'f_pvalue',
 'f_test',
 'fittedvalues',
 'fvalue',
 'get_influence',
 'get_prediction',
 'get_robustcov_results',
 'het_scale',
 'initialize',
 'k_constant',
 'llf',
 'load',
 'model',
 'mse_model',
 'mse_resid',
 'mse_total',
 'nobs',
 'normalized_cov_params',
 'outlier_test',
 'params',
 'predict',
 'pvalues',
 'remove_data',
 'resid',
 'resid_pearson',
 'rsquared',
 'rsquared_adj',
 'save',
 'scale',
 'ssr',
 'summary',
 'summary2',
 't_test',
 't_test_pairwise',
 'tvalues',
 'uncentered_tss',
 'use_t',
 'wald_test',
 'wald_test_terms',
 'wresid']
```

```
table = pd.DataFrame({'coefficient': round(results.params, 4),
                      'se': round(results.bse, 4),
                      't': round(results.tvalues, 4),
                      'pval': round(results.pvalues, 4)})
print(f'table: ₩n{table}₩n')
```

```
table:
                       coefficient      se        t    pval
Intercept                   0.4589  0.0934   4.9111  0.0000
y85                         0.1178  0.1238   0.9517  0.3415
educ                        0.0747  0.0067  11.1917  0.0000
female                     -0.3167  0.0366  -8.6482  0.0000
y85:educ                    0.0185  0.0094   1.9735  0.0487
y85:female                  0.0851  0.0513   1.6576  0.0977
exper                       0.0296  0.0036   8.2932  0.0000
I((exper ** 2) / 100)      -0.0399  0.0078  -5.1513  0.0000
union                       0.2021  0.0303   6.6722  0.0000
```

difference in difference (DiD)

```
data2= woo.dataWoo('kielmc')
data_summary(data2)
```

```
(321, 25)
Index(['year', 'age', 'agesq', 'nbh', 'cbd', 'intst', 'lintst', 'price',
       'rooms', 'area', 'land', 'baths', 'dist', 'ldist', 'wind', 'lprice',
       'y81', 'larea', 'lland', 'y81ldist', 'lintstsq', 'nearinc', 'y81nrinc',
       'rprice', 'lrprice'],
      dtype='object')
```

	year	age	agesq	nbh	cbd	intst	lintst	price	rooms	area	land	baths	
0	1978	48	2304.0	4	3000.0	1000.0	6.9078	60000.0	7	1660	4578.0	1	10
1	1978	83	6889.0	4	4000.0	1000.0	6.9078	40000.0	6	2612	8370.0	2	11
2	1978	58	3364.0	4	4000.0	1000.0	6.9078	34000.0	6	1144	5000.0	1	11

```
y78 = (data2['year'] == 1978)
print(type(y78))
y78
```

```
<class 'pandas.core.series.Series'>
0       True
1       True
2       True
3       True
4       True
       ...
316    False
317    False
318    False
319    False
320    False
Name: year, Length: 321, dtype: bool
```

separate regressions for 1978 and 1981:

```
print('year 1978')
y78 = (data2['year'] == 1978) # subset logic array
results78 = stat_ols('rprice ~ nearinc', data2, subset=y78)
```

```
year 1978
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rprice   R-squared:                       0.115
Model:                            OLS   Adj. R-squared:                  0.112
Method:                 Least Squares   F-statistic:                     41.32
Date:                Mon, 17 May 2021   Prob (F-statistic):           4.72e-10
Time:                        01:39:43   Log-Likelihood:                -3776.4
No. Observations:                 321   AIC:                             7557.
Df Residuals:                     319   BIC:                             7564.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
Intercept   9.104e+04   2080.730     43.752      0.000    8.69e+04    9.51e+04
nearinc    -2.446e+04   3804.807     -6.428      0.000   -3.19e+04    -1.7e+04
==============================================================================
Omnibus:                      179.474   Durbin-Watson:                   1.481
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1761.079
Skew:                           2.116   Prob(JB):                         0.00
Kurtosis:                      13.666   Cond. No.                         2.42
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
print('year 1981')
y81 = (data2['year'] == 1981) # subset logic array
results81 = stat_ols('rprice ~ nearinc', data2, subset=y81)
```

```
year 1981
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rprice   R-squared:                       0.115
Model:                            OLS   Adj. R-squared:                  0.112
Method:                 Least Squares   F-statistic:                     41.32
Date:                Mon, 17 May 2021   Prob (F-statistic):           4.72e-10
Time:                        01:39:09   Log-Likelihood:                -3776.4
No. Observations:                 321   AIC:                             7557.
Df Residuals:                     319   BIC:                             7564.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9.104e+04   2080.730     43.752      0.000    8.69e+04    9.51e+04
nearinc    -2.446e+04   3804.807     -6.428      0.000   -3.19e+04    -1.7e+04
==============================================================================
Omnibus:                      179.474   Durbin-Watson:                   1.481
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1761.079
Skew:                           2.116   Prob(JB):                         0.00
Kurtosis:                      13.666   Cond. No.                         2.42
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

joint regression including an interaction term:

- "C" denote category variables
- "*" will include the individual columns that were multiplied together (Interaction)
- Ref: https://www.statsmodels.org/stable/examples/notebooks/generated/formuas.html

```
result_joint = stat_ols('rprice ~ nearinc * C(year)', data2)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rprice   R-squared:                       0.174
Model:                            OLS   Adj. R-squared:                  0.166
Method:                 Least Squares   F-statistic:                     22.25
Date:                Mon, 17 May 2021   Prob (F-statistic):           4.22e-13
Time:                        01:40:38   Log-Likelihood:                -3765.2
No. Observations:                 321   AIC:                             7538.
Df Residuals:                     317   BIC:                             7554.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept              8.252e+04   2726.910     30.260      0.000    7.72e+04    8.79e+04
C(year)[T.1981]        1.879e+04   4050.065      4.640      0.000    1.08e+04    2.68e+04
nearinc               -1.882e+04   4875.322     -3.861      0.000   -2.84e+04   -9232.293
nearinc:C(year)[T.1981] -1.186e+04   7456.646     -1.591      0.113   -2.65e+04    2806.867
==============================================================================
Omnibus:                      192.562   Durbin-Watson:                   1.557
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2462.071
Skew:                           2.217   Prob(JB):                         0.00
Kurtosis:                      15.822   Cond. No.                         6.05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

difference in difference (DiD) with log price

```
# difference in difference (DiD):
result_did_with_log_price = stat_ols('np.log(rprice) ~ nearinc*C(year)', data2)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         np.log(rprice)   R-squared:                       0.246
Model:                            OLS   Adj. R-squared:                  0.239
Method:                 Least Squares   F-statistic:                     34.47
Date:                Mon, 17 May 2021   Prob (F-statistic):           2.62e-19
Time:                        01:47:41   Log-Likelihood:                -105.68
No. Observations:                 321   AIC:                             219.4
Df Residuals:                     317   BIC:                             234.4
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept                11.2854      0.031    369.839      0.000      11.225      11.345
C(year)[T.1981]           0.1931      0.045      4.261      0.000       0.104       0.282
nearinc                  -0.3399      0.055     -6.231      0.000      -0.447      -0.233
nearinc:C(year)[T.1981]  -0.0626      0.083     -0.751      0.453      -0.227       0.102
==============================================================================
Omnibus:                       29.462   Durbin-Watson:                   1.568
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               84.380
Skew:                           0.370   Prob(JB):                     4.75e-19
Kurtosis:                       5.400   Cond. No.                         6.05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

DiD with control variables

```
# DiD with control variables:

result_did_with_log_price_controls = stat_ols('np.log(rprice) ~ nearinc*C(year) + age +'
                           'I(age**2) + np.log(intst) + np.log(land) +'
                           'np.log(area) + rooms + baths', data2)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         np.log(rprice)   R-squared:                       0.733
Model:                            OLS   Adj. R-squared:                  0.724
Method:                 Least Squares   F-statistic:                     84.91
Date:                Mon, 17 May 2021   Prob (F-statistic):           1.24e-82
Time:                        01:51:30   Log-Likelihood:                 60.690
No. Observations:                 321   AIC:                            -99.38
Df Residuals:                     310   BIC:                            -57.89
Df Model:                          10
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                   7.6517      0.416     18.399      0.000       6.833       8.470
C(year)[T.1981]             0.1621      0.028      5.687      0.000       0.106       0.218
nearinc                     0.0322      0.047      0.679      0.498      -0.061       0.126
nearinc:C(year)[T.1981]    -0.1315      0.052     -2.531      0.012      -0.234      -0.029
age                        -0.0084      0.001     -5.924      0.000      -0.011      -0.006
I(age ** 2)              3.763e-05   8.67e-06      4.342      0.000    2.06e-05    5.47e-05
np.log(intst)              -0.0614      0.032     -1.950      0.052      -0.123       0.001
np.log(land)                0.0998      0.024      4.077      0.000       0.052       0.148
np.log(area)                0.3508      0.051      6.813      0.000       0.249       0.452
rooms                       0.0473      0.017      2.732      0.007       0.013       0.081
baths                       0.0943      0.028      3.400      0.001       0.040       0.149
==============================================================================
Omnibus:                       67.366   Durbin-Watson:                   1.710
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              356.621
Skew:                          -0.734   Prob(JB):                     3.64e-78
Kurtosis:                       7.951   Cond. No.                     1.83e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
[2] The condition number is large, 1.83e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

13.4 Panel: First Differenced Estimator

```
# (if not installed) pip install linearmodels
import linearmodels as plm
```

```
data3 = woo.dataWoo('crime4')
data_summary(data3)
```

```
(630, 59)
Index(['county', 'year', 'crmrte', 'prbarr', 'prbconv', 'prbpris', 'avgsen',
       'polpc', 'density', 'taxpc', 'west', 'central', 'urban', 'pctmin80',
       'wcon', 'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc',
       'mix', 'pctymle', 'd82', 'd83', 'd84', 'd85', 'd86', 'd87', 'lcrmrte',
       'lprbarr', 'lprbconv', 'lprbpris', 'lavgsen', 'lpolpc', 'ldensity',
       'ltaxpc', 'lwcon', 'lwtuc', 'lwtrd', 'lwfir', 'lwser', 'lwmfg', 'lwfed',
       'lwsta', 'lwloc', 'lmix', 'lpctymle', 'lpctmin', 'clcrmrte', 'clprbarr',
       'clprbcon', 'clprbpri', 'clavgsen', 'clpolpc', 'cltaxpc', 'clmix'],
      dtype='object')
```

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	ta
0	1	81	0.039885	0.289696	0.402062	0.472222	5.61	0.001787	2.307159	25.69
1	1	82	0.038345	0.338111	0.433005	0.506993	5.59	0.001767	2.330254	24.87
2	1	83	0.030305	0.330449	0.525703	0.479705	5.80	0.001836	2.341801	26.45

```
data3 = data3.set_index(['county', 'year'], drop=False)
data3.head()
```

		county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	den
county	**year**									
1	**81**	1	81	0.039885	0.289696	0.402062	0.472222	5.61	0.001787	2.30
	82	1	82	0.038345	0.338111	0.433005	0.506993	5.59	0.001767	2.33
	83	1	83	0.030305	0.330449	0.525703	0.479705	5.80	0.001836	2.34
	84	1	84	0.034726	0.362525	0.604706	0.520104	6.89	0.001886	2.34
	85	1	85	0.036573	0.325395	0.578723	0.497059	6.55	0.001924	2.36

```
reg = plm.FirstDifferenceOLS.from_formula('np.log(crmrte) ~ year + d83 + d84 + d85 + d86 + d87 +lprbar
first_difference_results = reg.fit()
print(first_difference_results)
print(first_difference_results)
```

```
                          FirstDifferenceOLS Estimation Summary
================================================================================
Dep. Variable:          np.log(crmrte)   R-squared:                        0.4326
Estimator:          FirstDifferenceOLS   R-squared (Between):              0.6003
No. Observations:                  540   R-squared (Within):               0.4281
Date:                 Mon, May 17 2021   R-squared (Overall):              0.6000
Time:                         02:05:38   Log-likelihood                    248.48
Cov. Estimator:             Unadjusted
                                         F-statistic:                      36.661
Entities:                           90   P-value                           0.0000
Avg Obs:                        7.0000   Distribution:                   F(11,529)
Min Obs:                        7.0000
Max Obs:                        7.0000   F-statistic (robust):             36.661
                                         P-value                           0.0000
Time periods:                        7   Distribution:                   F(11,529)
Avg Obs:                        90.000
Min Obs:                        90.000
Max Obs:                        90.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
year           0.0077     0.0171     0.4522     0.6513     -0.0258      0.0412
d83           -0.0999     0.0239    -4.1793     0.0000     -0.1468     -0.0529
d84           -0.1478     0.0413    -3.5806     0.0004     -0.2289     -0.0667
d85           -0.1524     0.0584    -2.6098     0.0093     -0.2671     -0.0377
d86           -0.1249     0.0760    -1.6433     0.1009     -0.2742      0.0244
d87           -0.0841     0.0940    -0.8944     0.3715     -0.2687      0.1006
lprbarr       -0.3275     0.0300    -10.924     0.0000     -0.3864     -0.2686
lprbconv      -0.2381     0.0182    -13.058     0.0000     -0.2739     -0.2023
lprbpris      -0.1650     0.0260    -6.3555     0.0000     -0.2161     -0.1140
lavgsen       -0.0218     0.0221    -0.9850     0.3251     -0.0652      0.0216
lpolpc         0.3984     0.0269     14.821     0.0000      0.3456      0.4512
==============================================================================
```

User-Define Functions for Statmodels

- Name: Jikhan Jeong
- Date: 5-16-2021 (Updated)

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf

def data_summary(df):
    print(df.shape)
    print(df.columns)
    return df.head(3)

def stat_ols(formulas,df,subset=None, drop_cols=None):
     '''
     Name: Jikhan Jeong
     * Simpler ols
     * Requires: import statsmodels.formula.api as smf
     '''
     reg = smf.ols(formulas, df,subset=None, drop_cols=None)
     results = reg.fit()
     print(results.summary())
     return results

# (ex) stat_ols('lwage ~y85*(educ+female) + exper + I((exper**2)/100) + union', data)
```

✓ 0초 오후 7:05에 완료됨 ● ×