Forest Fires

June 11, 2016

```
Regression Modeling in Practice Course
Wesleyan University

Linear Regression Model
Mario Colosso V.

The sample comes from Cortez and Morais study about predicting forest fires using
metereological data [Cortez and Morais, 2007]. The study includes data from 517
forest fires in the Natural Park Montesinho (Trás-os -Montes, in northeastern Portugal)
January 2000 to December 2003, including meteorological data, the type of vegetation
involved (which determines the six components of the Canadian Forest Fire Weather Index
(FWI) system --see below--) and the total burned area in order to generate a model capable
of predicting the burned area of small fires, which are more frequent.

Measures
The data contains:
* X, Y: location of the fire (x,y axis spatial coordinate within the Montesinho park map:
  from 1 to 9)
* month, day: month and day of the week the fire occurred (january to december and monday
  to sunday)
* FWI system components:
  - FFMC: Fine Fuel Moisture Code (numeric rating of the moisture content of litter and
    other cured fine fuels: 18.7 to 96.2)
  - DMC: Duff Moisture Code (numeric rating of the average moisture content of loosely
    compacted organic layers of moderate depth: 1.1 to 291.3)
  - DC: Drought Code (numeric rating of the average moisture content of deep, compact
    organic layers: 7.9 to 860.6)
  - ISI: Initial Spread Index (numeric rating of the expected rate of fire spread: 0.0
    to 56.1)
* Metereological variables:
  - temp: temperature (2.2 to 33.3 °C)
  - RH: relative humidity (15 to 100%)
  - wind: wind speed (0.4 to 9.4 Km/h)
  - rain: outside rain (0.0 to 6.4 mm/m^2)
* area: the burned area of the forest as response variable (0.0 to 1090.84 Ha).
```

In [1]:
```
%matplotlib inline

import pandas
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
from math import ceil
```

```
pandas.set_option('display.float_format', lambda x:'%.3f'%x)
plt.style.use('ggplot') # Make the graphs a bit prettier
plt.rcParams['figure.figsize'] = (15, 5)
```

0.0.1 Load Forest Fires .csv file

```
In [2]: fires = pandas.read_csv('forestfires.csv')
```

0.1 1. Lets have a brief look of Fires DataFrame

```
In [3]: fires.head()   #Show first rows
```

Out[3]:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.200	26.200	94.300	5.100	8.200	51	6.700	0.000	0.000
1	7	4	oct	tue	90.600	35.400	669.100	6.700	18.000	33	0.900	0.000	0.000
2	7	4	oct	sat	90.600	43.700	686.900	6.700	14.600	33	1.300	0.000	0.000
3	8	6	mar	fri	91.700	33.300	77.500	9.000	8.300	97	4.000	0.200	0.000
4	8	6	mar	sun	89.300	51.300	102.200	9.600	11.400	99	1.800	0.000	0.000

0.1.1 Get some descriptive statistic of the data

```
In [4]: fires_attributes = fires.columns.values.tolist()
        number_of_columns = len(fires_attributes)

In [5]: statistics = pandas.DataFrame(index=range(0, number_of_columns - 2),
                                      columns=('min', 'max', 'mean', 'median', 'std'))

In [6]: idx = 0
        for attr in [0, 1] + list(range(4, number_of_columns)):
            statistics.loc[idx] = {'min':    min(fires[fires_attributes[attr]]),
                                   'max':    max(fires[fires_attributes[attr]]),
                                   'mean':   fires[fires_attributes[attr]].mean(),
                                   'median': fires[fires_attributes[attr]].median(),
                                   'std':    fires[fires_attributes[attr]].std()}
            idx += 1
        statistics.index = [fires_attributes[attr]
                            for attr in [0, 1] + list(range(4, number_of_columns))]

In [7]: statistics.T   #Show min, max, mean, median and standard deviation
```

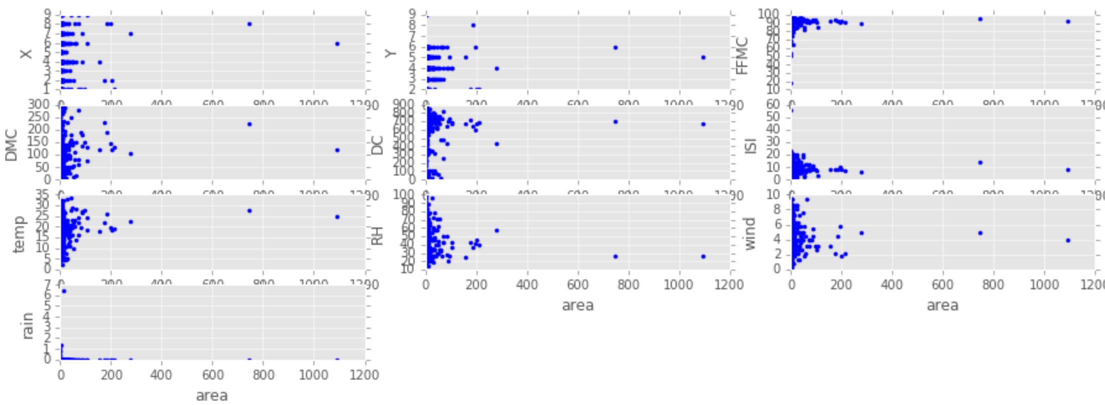
Out[7]:

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	\
min	1.000	2.000	18.700	1.100	7.900	0.000	2.200	15.000	0.400	0.000	
max	9.000	9.000	96.200	291.300	860.600	56.100	33.300	100.000	9.400	6.400	
mean	4.669	4.300	90.645	110.872	547.940	9.022	18.889	44.288	4.018	0.022	
median	4.000	4.000	91.600	108.300	664.200	8.400	19.300	42.000	4.000	0.000	
std	2.314	1.230	5.520	64.046	248.066	4.559	5.807	16.317	1.792	0.296	

	area
min	0.000
max	1090.840
mean	12.847
median	0.520
std	63.656

0.1.2 And display a graph of quantitative variables vs area

```
In [8]: attributes = [0, 1] + list(range(4, number_of_columns - 1))
        n_cols = 3
        n_rows = int(ceil(len(attributes) / n_cols))
        fig = plt.figure()
        idx = 1
        for attr in attributes:
            plt.subplot(n_rows, n_cols, idx)
            plt.plot(fires['area'], fires[fires_attributes[attr]], 'b.')
            plt.xlabel('area')
            plt.ylabel(fires_attributes[attr])
            idx += 1

        plt.show()
```

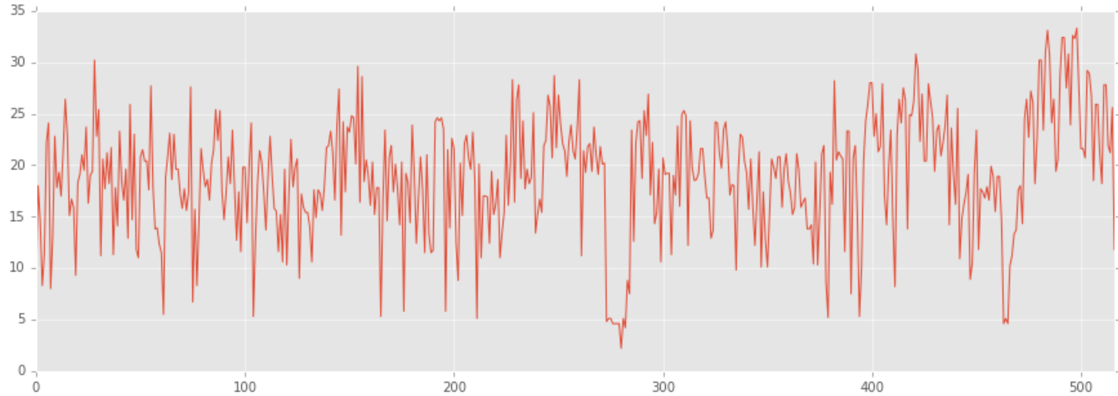


There are some data values where the burned area is away from other values

```
In [9]: fires[fires['area'] > 250]

Out[9]:      X  Y month  day   FFMC     DMC      DC    ISI   temp  RH  wind  rain  \
        238  6  5   sep  sat 92.500 121.100 674.400  8.600 25.100  27 4.000 0.000
        415  8  6   aug  thu 94.800 222.400 698.600 13.900 27.500  27 4.900 0.000
        479  7  4   jul  mon 89.200 103.900 431.600  6.400 22.600  57 4.900 0.000

                area
        238 1090.840
        415  746.280
        479  278.530
```

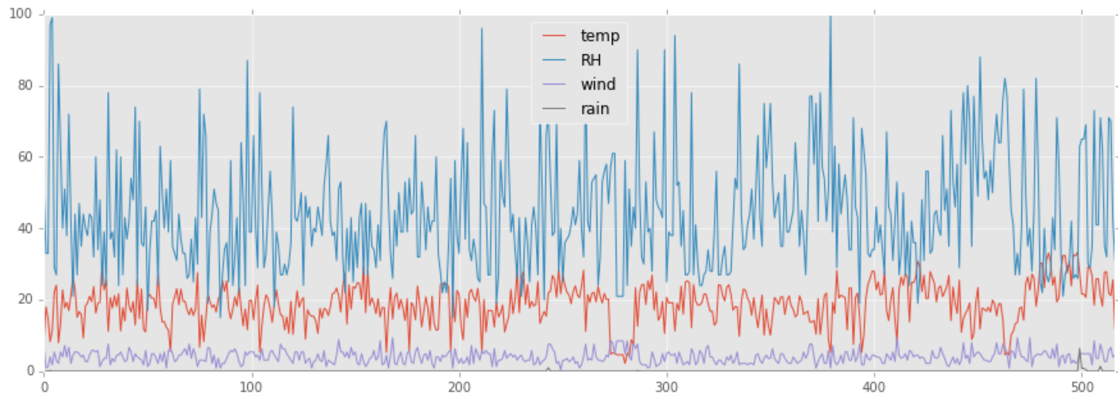
0.1.3 Plot some other variables

```
In [10]: fires['temp'].plot()    #Plot temperature graph

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x2746ce55f98>
```



```
In [11]: fires[['temp', 'RH', 'wind', 'rain']].plot()    #Plot temperature, relative humidity, wind
                                                          #and rain graphs
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x2746cdc97b8>
```




```
In [12]: fires.corr()    #Show correlation between variables
```

Out[12]:

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain \
X	1.000	0.540	-0.021	-0.048	-0.086	0.006	-0.051	0.085	0.019	0.065
Y	0.540	1.000	-0.046	0.008	-0.101	-0.024	-0.024	0.062	-0.020	0.033
FFMC	-0.021	-0.046	1.000	0.383	0.331	0.532	0.432	-0.301	-0.028	0.057
DMC	-0.048	0.008	0.383	1.000	0.682	0.305	0.470	0.074	-0.105	0.075
DC	-0.086	-0.101	0.331	0.682	1.000	0.229	0.496	-0.039	-0.203	0.036
ISI	0.006	-0.024	0.532	0.305	0.229	1.000	0.394	-0.133	0.107	0.068
temp	-0.051	-0.024	0.432	0.470	0.496	0.394	1.000	-0.527	-0.227	0.069
RH	0.085	0.062	-0.301	0.074	-0.039	-0.133	-0.527	1.000	0.069	0.100
wind	0.019	-0.020	-0.028	-0.105	-0.203	0.107	-0.227	0.069	1.000	0.061
rain	0.065	0.033	0.057	0.075	0.036	0.068	0.069	0.100	0.061	1.000
area	0.063	0.045	0.040	0.073	0.049	0.008	0.098	-0.076	0.012	-0.007

	area
X	0.063

```
Y     0.045
FFMC  0.040
DMC   0.073
DC    0.049
ISI   0.008
temp  0.098
RH   -0.076
wind  0.012
rain -0.007
area  1.000
```

0.2 2. Linear regression

0.2.1 Convert categorical variables (months and days) into numerical values

In [13]:
```
months_table = ['jan', 'feb', 'mar', 'apr', 'may', 'jun',
                'jul', 'aug', 'sep', 'oct', 'nov', 'dec']
days_table =   ['sun', 'mon', 'tue', 'wed', 'thu', 'fri', 'sat']

fires['month'] = [months_table.index(month) for month in fires['month'] ]
fires['day'] =   [days_table.index(day)     for day   in fires['day']   ]

fires['X'] -= 1
fires['Y'] -= 2

fires.head()
```

Out[13]:

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	6	3	2	5	86.200	26.200	94.300	5.100	8.200	51	6.700	0.000	0.000
1	6	2	9	2	90.600	35.400	669.100	6.700	18.000	33	0.900	0.000	0.000
2	6	2	9	6	90.600	43.700	686.900	6.700	14.600	33	1.300	0.000	0.000
3	7	4	2	5	91.700	33.300	77.500	9.000	8.300	97	4.000	0.200	0.000
4	7	4	2	0	89.300	51.300	102.200	9.600	11.400	99	1.800	0.000	0.000

0.2.2 Center each explanatory variable

In [14]:
```
for idx in list(range(4, number_of_columns - 1)):   #Exclude categorical variables
    fires[fires_attributes[idx]] = fires[fires_attributes[idx]] - \
                                   fires[fires_attributes[idx]].mean()
```

In [15]:
```
statistics = [fires[fires_attributes[idx]].mean() for idx in range(0, number_of_columns)]
statistics = pandas.DataFrame(statistics,
                              index=fires_attributes,
                              columns=['mean'])
```

In [16]:
```
statistics.T   #Only quantitative explanatory variables (FFMC thru rain) were centered
```

Out[16]:

```
         X     Y  month   day  FFMC    DMC    DC    ISI  temp    RH   wind  \
mean 3.669 2.300  6.476 2.973 0.000 -0.000 0.000 -0.000 0.000 0.000 -0.000

      rain   area
mean 0.000 12.847
```

0.2.3 Generate models to test each variable

In [17]:
```
statistics = list()
for idx in range(0, number_of_columns - 1):
```

```
        model = smf.ols(formula = "area ~ " +
                        fires_attributes[idx], data = fires).fit()

        title = 'Model: area ~ ' + fires_attributes[idx]
        print('+' + "-" * (len(title) + 2) + '+' + '\n' +
              '| ' + title + ' |' + '\n' +
              '+' + "-" * (len(title) + 2) + '+')
        print()
        print(model.summary())
        print()
        statistics.append([model.f_pvalue, model.rsquared])
```

```
+-----------------+
| Model: area ~ X |
+-----------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.077
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.150
Time:                        18:12:03   Log-Likelihood:                -2879.4
No. Observations:                 517   AIC:                             5763.
Df Residuals:                     515   BIC:                             5771.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      6.4487      5.247      1.229      0.220        -3.859    16.756
X              1.7438      1.210      1.441      0.150        -0.633     4.121
==============================================================================
Omnibus:                      981.662   Durbin-Watson:                   1.653
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           802838.467
Skew:                          12.752   Prob(JB):                         0.00
Kurtosis:                     194.360   Cond. No.                         8.45
==============================================================================

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

+-----------------+
| Model: area ~ Y |
+-----------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.039
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.309
Time:                        18:12:03   Log-Likelihood:                -2879.9
No. Observations:                 517   AIC:                             5764.
```

```
Df Residuals:                     515   BIC:                             5772.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      7.5060      5.941      1.263      0.207        -4.165    19.177
Y              2.3225      2.278      1.019      0.309        -2.154     6.799
==============================================================================
Omnibus:                      981.970   Durbin-Watson:                   1.645
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           802937.403
Skew:                          12.761   Prob(JB):                         0.00
Kurtosis:                     194.369   Cond. No.                         6.19
==============================================================================

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

+---------------------+
| Model: area ~ month |
+---------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.649
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.200
Time:                        18:12:03   Log-Likelihood:                -2879.6
No. Observations:                 517   AIC:                             5763.
Df Residuals:                     515   BIC:                             5772.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      2.6149      8.445      0.310      0.757       -13.976    19.206
month          1.5801      1.230      1.284      0.200        -0.837     3.997
==============================================================================
Omnibus:                      983.027   Durbin-Watson:                   1.647
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           807389.375
Skew:                          12.790   Prob(JB):                         0.00
Kurtosis:                     194.901   Cond. No.                         21.1
==============================================================================

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

+-------------------+
| Model: area ~ day |
+-------------------+

                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:                   area   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.207
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.272
Time:                        18:12:03   Log-Likelihood:                -2879.8
No. Observations:                 517   AIC:                             5764.
Df Residuals:                     515   BIC:                             5772.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      8.5785      4.788      1.792      0.074        -0.829    17.986
day            1.4359      1.307      1.099      0.272        -1.132     4.003
==============================================================================
Omnibus:                      980.555   Durbin-Watson:                   1.636
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           794438.352
Skew:                          12.725   Prob(JB):                         0.00
Kurtosis:                     193.346   Cond. No.                         6.58
==============================================================================

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

+---------------------+
| Model: area ~ FFMC |
+---------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.8304
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.363
Time:                        18:12:03   Log-Likelihood:                -2880.0
No. Observations:                 517   AIC:                             5764.
Df Residuals:                     515   BIC:                             5773.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.800      4.588      0.000         7.346    18.348
FFMC           0.4627      0.508      0.911      0.363        -0.535     1.460
==============================================================================
Omnibus:                      983.137   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           808340.065
Skew:                          12.793   Prob(JB):                         0.00
Kurtosis:                     195.015   Cond. No.                         5.51
==============================================================================

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

```
+-------------------+
| Model: area ~ DMC |
+-------------------+

                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   area   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     2.759
Date:                Sat, 11 Jun 2016   Prob (F-statistic):             0.0973
Time:                        18:12:03   Log-Likelihood:                -2879.1
No. Observations:                 517   AIC:                             5762.
Df Residuals:                     515   BIC:                             5771.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.795      4.597      0.000         7.357    18.338
DMC            0.0725      0.044      1.661      0.097        -0.013     0.158
==============================================================================
Omnibus:                      982.803   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           811231.935
Skew:                          12.780   Prob(JB):                         0.00
Kurtosis:                     195.368   Cond. No.                         64.0
==============================================================================

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

+------------------+
| Model: area ~ DC |
+------------------+

                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   area   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.259
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.262
Time:                        18:12:03   Log-Likelihood:                -2879.8
No. Observations:                 517   AIC:                             5764.
Df Residuals:                     515   BIC:                             5772.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.799      4.590      0.000         7.349    18.346
DC             0.0127      0.011      1.122      0.262        -0.010     0.035
==============================================================================
Omnibus:                      982.892   Durbin-Watson:                   1.645
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           807312.305
Skew:                          12.786   Prob(JB):                         0.00
```

```
Kurtosis:                     194.893   Cond. No.                         248.
==============================================================================

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

+--------------------+
| Model: area ~ ISI |
+--------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                   0.03512
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.851
Time:                        18:12:03   Log-Likelihood:                -2880.4
No. Observations:                 517   AIC:                             5765.
Df Residuals:                     515   BIC:                             5773.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.802      4.585      0.000         7.342    18.352
ISI            0.1153      0.615      0.187      0.851        -1.093     1.324
==============================================================================
Omnibus:                      983.625   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           809992.277
Skew:                          12.806   Prob(JB):                         0.00
Kurtosis:                     195.211   Cond. No.                         4.56
==============================================================================

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

+---------------------+
| Model: area ~ temp |
+---------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     4.978
Date:                Sat, 11 Jun 2016   Prob (F-statistic):             0.0261
Time:                        18:12:03   Log-Likelihood:                -2878.0
No. Observations:                 517   AIC:                             5760.
Df Residuals:                     515   BIC:                             5768.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
```

```
Intercept     12.8473      2.789      4.607      0.000         7.368    18.326
temp           1.0726      0.481      2.231      0.026         0.128     2.017
==============================================================================
Omnibus:                      979.270   Durbin-Watson:                   1.650
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           793772.021
Skew:                          12.687   Prob(JB):                         0.00
Kurtosis:                     193.275   Cond. No.                         5.80
==============================================================================

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

+-------------------+
| Model: area ~ RH |
+-------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     2.954
Date:                Sat, 11 Jun 2016   Prob (F-statistic):             0.0863
Time:                        18:12:03   Log-Likelihood:                -2879.0
No. Observations:                 517   AIC:                             5762.
Df Residuals:                     515   BIC:                             5770.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.794      4.598      0.000         7.358    18.337
RH            -0.2946      0.171     -1.719      0.086        -0.631     0.042
==============================================================================
Omnibus:                      980.422   Durbin-Watson:                   1.642
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           795947.965
Skew:                          12.720   Prob(JB):                         0.00
Kurtosis:                     193.531   Cond. No.                         16.3
==============================================================================

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

+---------------------+
| Model: area ~ wind |
+---------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                   0.07815
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.780
Time:                        18:12:03   Log-Likelihood:                -2880.4
No. Observations:                 517   AIC:                             5765.
```

```
Df Residuals:                     515   BIC:                             5773.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.802      4.585      0.000         7.342    18.352
wind           0.4376      1.565      0.280      0.780        -2.638     3.513
==============================================================================
Omnibus:                      983.721   Durbin-Watson:                   1.647
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           810324.708
Skew:                          12.809   Prob(JB):                         0.00
Kurtosis:                     195.251   Cond. No.                         1.79
==============================================================================

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

+---------------------+
| Model: area ~ rain |
+---------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                   0.02794
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.867
Time:                        18:12:03   Log-Likelihood:                -2880.4
No. Observations:                 517   AIC:                             5765.
Df Residuals:                     515   BIC:                             5773.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.802      4.585      0.000         7.342    18.352
rain          -1.5842      9.477     -0.167      0.867       -20.203    17.035
==============================================================================
Omnibus:                      983.726   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           810320.385
Skew:                          12.809   Prob(JB):                         0.00
Kurtosis:                     195.250   Cond. No.                         3.38
==============================================================================

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

0.2.4 Models summary:

```
In [18]: statistics = pandas.DataFrame(statistics,
                                       index=fires_attributes[: number_of_columns - 1],
                                       columns=['p-value', 'R-squared'])
         statistics.T
```

```
Out[18]:                X     Y  month   day  FFMC   DMC    DC   ISI  temp    RH  wind  \
         p-value   0.150 0.309  0.200 0.272 0.363 0.097 0.262 0.851 0.026 0.086 0.780
         R-squared 0.004 0.002  0.003 0.002 0.002 0.005 0.002 0.000 0.010 0.006 0.000

                    rain
         p-value   0.867
         R-squared 0.000
```

```
In [19]: statistics[statistics['p-value'] < 0.05]
```

```
Out[19]:       p-value  R-squared
         temp    0.026      0.010
```

'temp' is the only statistically significant variable (p-value = 0.026) but it only explains the 1% of forest fires. Let's show its linear model summary:

```
In [20]: print((smf.ols(formula = "area ~ temp", data = fires).fit()).summary())
```

```
OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     4.978
Date:                Sat, 11 Jun 2016   Prob (F-statistic):             0.0261
Time:                        18:12:04   Log-Likelihood:                -2878.0
No. Observations:                 517   AIC:                             5760.
Df Residuals:                     515   BIC:                             5768.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.789      4.607      0.000         7.368    18.326
temp           1.0726      0.481      2.231      0.026         0.128     2.017
==============================================================================
Omnibus:                      979.270   Durbin-Watson:                   1.650
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           793772.021
Skew:                          12.687   Prob(JB):                         0.00
Kurtosis:                     193.275   Cond. No.                         5.80
==============================================================================

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

The results of the linear regression models indicated than only temperature (Beta = 1.0726, p = 0.026, R^2 = 0.010) was significantly and positively associated with the total burned area due to forest fires. 'p-value' of other models are greater than treshold value of 0.05 so results are not statistically significant to reject null hypothesis.

0.2.5 Create a Linear Regression Model for a combination of all variables

```
In [21]: explanatory_variables = "X + Y + month + day + FFMC + DMC + DC + ISI + temp + RH + " + \
                                 "wind + rain"
         response_variable =     "area"

         model = smf.ols(formula = response_variable + " ~ " + explanatory_variables,
                         data = fires).fit()
```

In [22]: print(model.summary())

```
OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.025
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.092
Date:                Sat, 11 Jun 2016   Prob (F-statistic):              0.364
Time:                        18:12:04   Log-Likelihood:                -2873.8
No. Observations:                 517   AIC:                             5774.
Df Residuals:                     504   BIC:                             5829.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -17.5974     19.340     -0.910      0.363       -55.595    20.400
X              1.9002      1.450      1.311      0.191        -0.948     4.748
Y              0.3241      2.754      0.118      0.906        -5.086     5.734
month          2.9004      2.791      1.039      0.299        -2.583     8.384
day            1.3269      1.320      1.005      0.315        -1.267     3.921
FFMC          -0.1127      0.663     -0.170      0.865        -1.415     1.190
DMC            0.0966      0.071      1.369      0.172        -0.042     0.235
DC            -0.0315      0.032     -0.981      0.327        -0.095     0.032
ISI           -0.7305      0.772     -0.947      0.344        -2.247     0.786
temp           0.9546      0.797      1.198      0.232        -0.612     2.521
RH            -0.1758      0.241     -0.730      0.466        -0.649     0.297
wind           1.2321      1.702      0.724      0.470        -2.113     4.577
rain          -3.1958      9.683     -0.330      0.742       -22.220    15.829
==============================================================================
Omnibus:                      972.663   Durbin-Watson:                   1.643
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           769640.593
Skew:                          12.508   Prob(JB):                         0.00
Kurtosis:                     190.356   Cond. No.                     1.76e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.76e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

p-value of combination model (p = 0.410) is bigger than treshold value, so the combination of the Canadian Forest Fire Weather Index (FWI) system plus temperature, humidity, wind and rain are not significantly associated with the total burned area due to forest fires. p-value of temperature in combination model (p = 0.282) is not longer statistically significant, a confounder variable?

Also, there is a warning in previous model summary: "The condition number is large, 1.76e+03. This might indicate that there are **strong multicollinearity** or other numerical problems." We will review this issue next week.

In [ ]: