# Forest Fires - week 3

June 14, 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).

# 1 Forest Fires

## 1.1 Import required libraries and set global options

```
In [1]: %matplotlib inline

        import pandas
        import matplotlib.pyplot as plt
        import seaborn
        import statsmodels.api as sm
        import statsmodels.formula.api as smf
        from pandas.tools.plotting import scatter_matrix
        from math import ceil

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

### 1.2 Load Forest Fires .csv file

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

## 2 Data Exploration

```
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
```

### 2.1 Get some descriptive statistic of the data

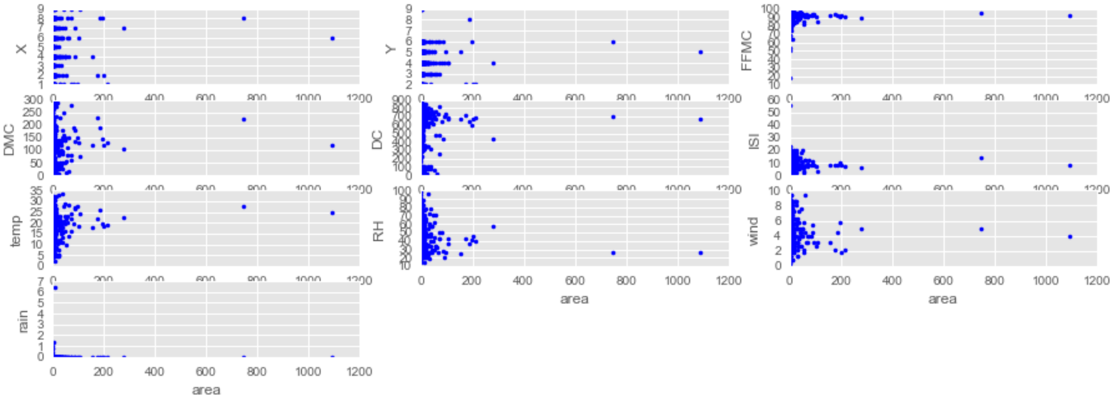
```
In [4]: fires_attributes = fires.columns.values.tolist()
        number_of_columns = len(fires_attributes)
```

```
In [5]: fires.describe()    #Original data
```

```
Out[5]:               X       Y    FFMC     DMC      DC     ISI    temp      RH    wind  \
        count 517.000 517.000 517.000 517.000 517.000 517.000 517.000 517.000 517.000
        mean    4.669   4.300  90.645 110.872 547.940   9.022  18.889  44.288   4.018
        std     2.314   1.230   5.520  64.046 248.066   4.559   5.807  16.317   1.792
        min     1.000   2.000  18.700   1.100   7.900   0.000   2.200  15.000   0.400
        25%     3.000   4.000  90.200  68.600 437.700   6.500  15.500  33.000   2.700
        50%     4.000   4.000  91.600 108.300 664.200   8.400  19.300  42.000   4.000
        75%     7.000   5.000  92.900 142.400 713.900  10.800  22.800  53.000   4.900
        max     9.000   9.000  96.200 291.300 860.600  56.100  33.300 100.000   9.400

                 rain     area
        count 517.000  517.000
        mean    0.022   12.847
        std     0.296   63.656
        min     0.000    0.000
        25%     0.000    0.000
        50%     0.000    0.520
        75%     0.000    6.570
        max     6.400 1090.840
```

## 2.2 Display a graph of quantitative variables vs area

```
In [6]: 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.')
        #    seaborn.regplot(x = fires['area'], y = fires[fires_attributes[attr]],
        #                     scatter = True, color = 'b', data = fires)
            plt.xlabel('area')
            plt.ylabel(fires_attributes[attr])
            idx += 1

        plt.show()
```



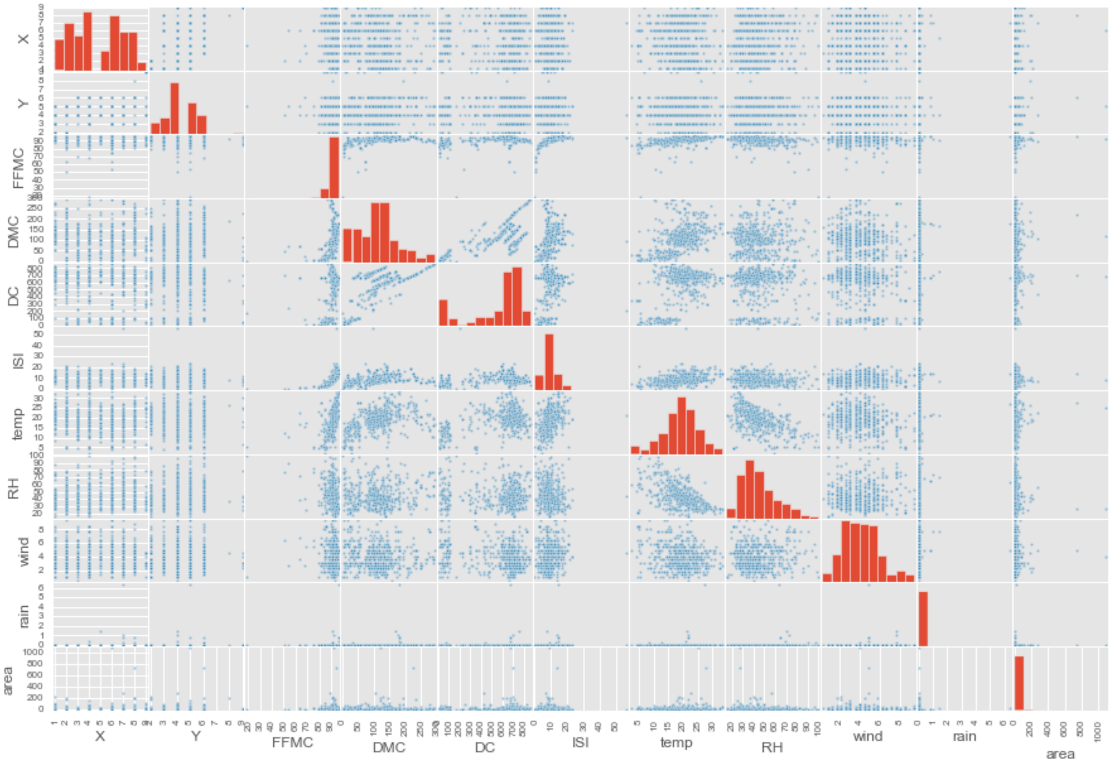
**There are some data values where the burned area is away from other values:**

```
In [7]: fires[fires['area'] > 250]
```

```
Out[7]:      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
```

## 2.3 Plot some other variables

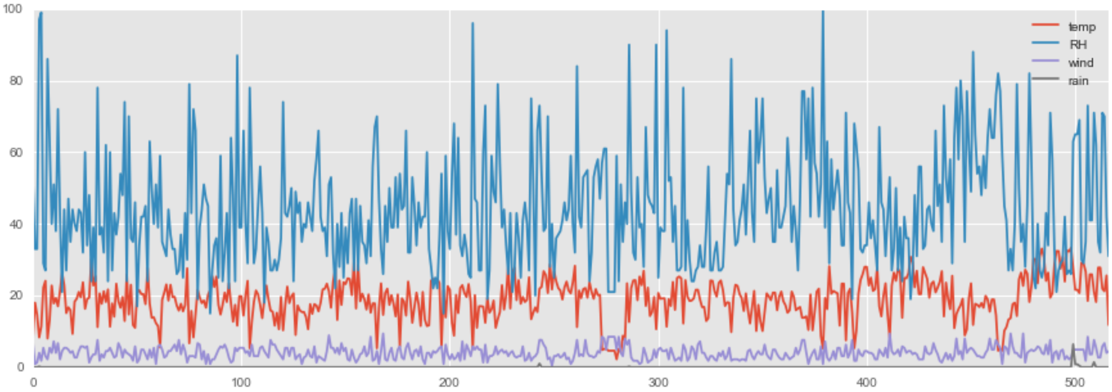
```
In [8]: scatter_matrix(fires, figsize = (15,10))
        plt.show()
```



High bias are appreciated in **FFMC**, **DC**, **ISI**, **wind** and **area** variables

```
In [9]: fires[['temp', 'RH', 'wind', 'rain']].plot()    #Plot temperature, relative humidity, wind
                                                         #and rain graphs
```

```
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x1bae4eb7a20>
```

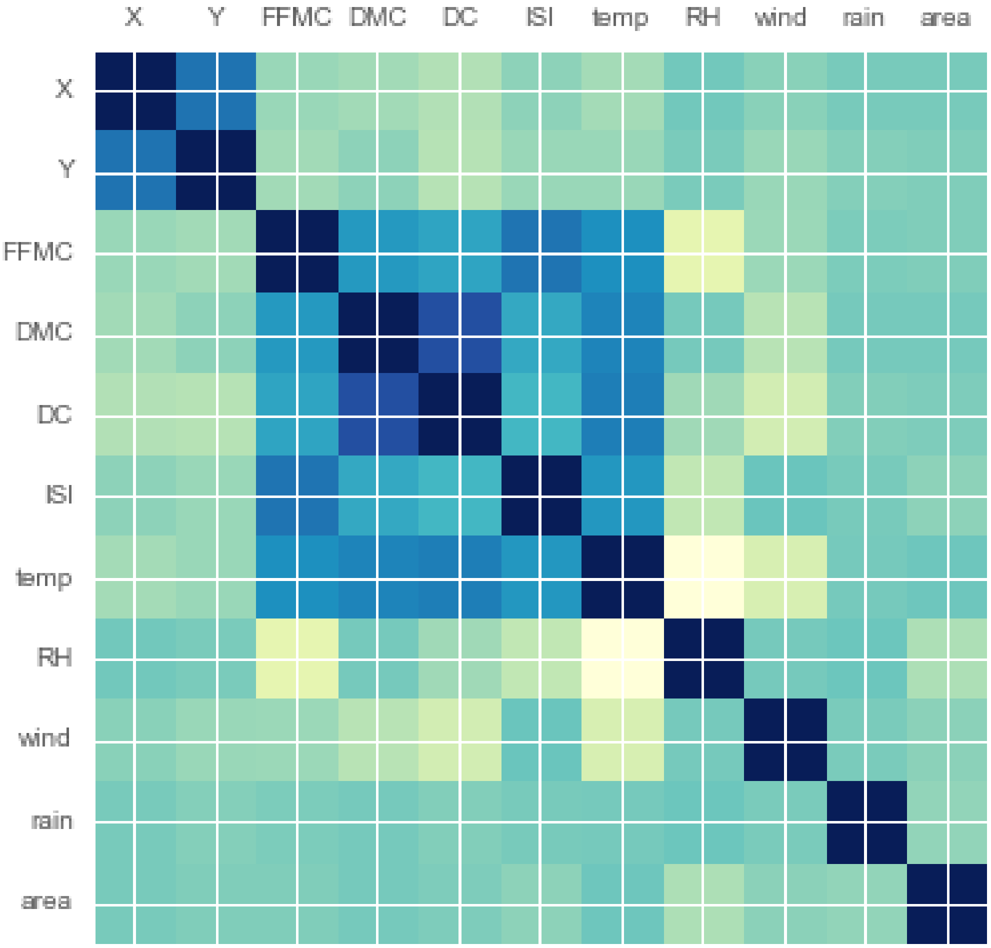


```
In [10]: fires.corr()    #Show correlation between variables
```

```
Out[10]:            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
```

```
In [11]: def plot_corr(df, size=10):
             '''Function plots a graphical correlation matrix for each pair of columns
                in the dataframe, including the names of the attributes
             Input:
                 df: pandas DataFrame
                 size: vertical and horizontal size of the plot

             Code taken from:
             http://stackoverflow.com/questions/29432629/correlation-matrix-using-pandas
             '''

             corr = df.corr()
             fig, ax = plt.subplots(figsize=(size, size))
             ax.matshow(corr, cmap = 'YlGnBu')
             plt.xticks(range(len(corr.columns)), corr.columns);
             plt.yticks(range(len(corr.columns)), corr.columns);

         #plt.matshow(fires.corr())
         plot_corr(fires, size = 6)
```



There is a medium-high correlation (**0.682**) between **DC** (Drought Code: numeric rating of the average moisture content of deep, compact organic layers) and **DMC** (Duff Moisture Code: numeric rating of the average moisture content of loosely compacted organic layers of moderate depth) and medium correlation (**0.532**) between **ISI** (Initial Spread Index: numeric rating of the expected rate of fire spread) and **FFMC** (Fine Fuel Moisture Code: numeric rating of the moisture content of litter and other cured fine fuels). Also, there is a inverse medium correlation (**-0.527**) between temperature (**temp**) and relative humidity (**RH**). Other relationships are noted between temperature (**temp**) and FWI system components (**FFMC**, **DCM**, **DC** and **ISI**)

# 3 Linear regression

## 3.1 Convert categorical variables (months and days) into numerical values

```
In [12]: 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[12]:    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
```

### 3.2 Center each explanatory variable

```
In [13]: 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 [14]: fires.describe()   #Only quantitative explanatory variables (FFMC thru rain) were centered

Out[14]:             X       Y   month     day    FFMC      DMC       DC     ISI  \
         count 517.000 517.000 517.000 517.000 517.000  517.000  517.000 517.000
         mean    3.669   2.300   6.476   2.973   0.000   -0.000    0.000  -0.000
         std     2.314   1.230   2.276   2.144   5.520   64.046  248.066   4.559
         min     0.000   0.000   0.000   0.000 -71.945 -109.772 -540.040  -9.022
         25%     2.000   2.000   6.000   1.000  -0.445  -42.272 -110.240  -2.522
         50%     3.000   2.000   7.000   3.000   0.955   -2.572  116.260  -0.622
         75%     6.000   3.000   8.000   5.000   2.255   31.528  165.960   1.778
         max     8.000   7.000  11.000   6.000   5.555  180.428  312.660  47.078

                  temp      RH    wind    rain     area
         count 517.000 517.000 517.000 517.000  517.000
         mean    0.000   0.000  -0.000   0.000   12.847
         std     5.807  16.317   1.792   0.296   63.656
         min   -16.689 -29.288  -3.618  -0.022    0.000
         25%    -3.389 -11.288  -1.318  -0.022    0.000
         50%     0.411  -2.288  -0.018  -0.022    0.520
         75%     3.911   8.712   0.882  -0.022    6.570
         max    14.411  55.712   5.382   6.378 1090.840
```

### 3.3 Generate models to test each variable

```
In [15]: def print_title(title):
             print(’+’ + "-" * (len(title) + 2) + ’+’ + ’\n’ +
                   ’| ’ + title + ’ |’ + ’\n’ +
                   ’+’ + "-" * (len(title) + 2) + ’+’)

In [16]: statistics = list()
         for idx in range(0, number_of_columns - 1):
             model = smf.ols(formula = "area ~ " +
                             fires_attributes[idx], data = fires).fit()

             print_title(’Model: area ~ ’ + fires_attributes[idx])
             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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.150
Time:                        22:34:25   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.309
Time:                        22:34:25   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.200
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.272
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.363
Time:                        22:34:26   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:               Tue, 14 Jun 2016   Prob (F-statistic):             0.0973
Time:                       22:34:26   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:               Tue, 14 Jun 2016   Prob (F-statistic):              0.262
Time:                       22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.851
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0261
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0863
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.780
Time:                        22:34:26   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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.867
Time:                        22:34:26   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.
```

### 3.3.1 Summary:

```
In [17]: statistics = pandas.DataFrame(statistics,
                                       index=fires_attributes[: number_of_columns - 1],
                                       columns=['p-value', 'R-squared'])
         statistics.T
```

```
Out[17]:                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 [18]: statistics[statistics['p-value'] < 0.05]
```

```
Out[18]:       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 [19]: 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:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0261
Time:                        22:34:26   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.**

## 3.4 Create a Linear Regression Model for a combination of all variables

```
In [20]: 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 [21]: 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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.364
```

```
Time:                        22:34:27   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?**

# 4 Test a Multiple Regression Model

## 4.1 Sort explanatory variables by p-value

```
In [22]: statistics = statistics.sort_values(by='p-value')
```

## 4.2 Define an useful function to plot QQ and Residual plots

```
In [23]: def print_qqplot_and_residuals_plot(model):
             # qq-plot
             ax1 = plt.subplot(1, 3, 1)
             qq_plot = sm.qqplot(model.resid, line = 'r', ax = ax1)

             # Residuals plot
             ax2 = plt.subplot(1, 3, 2)
```

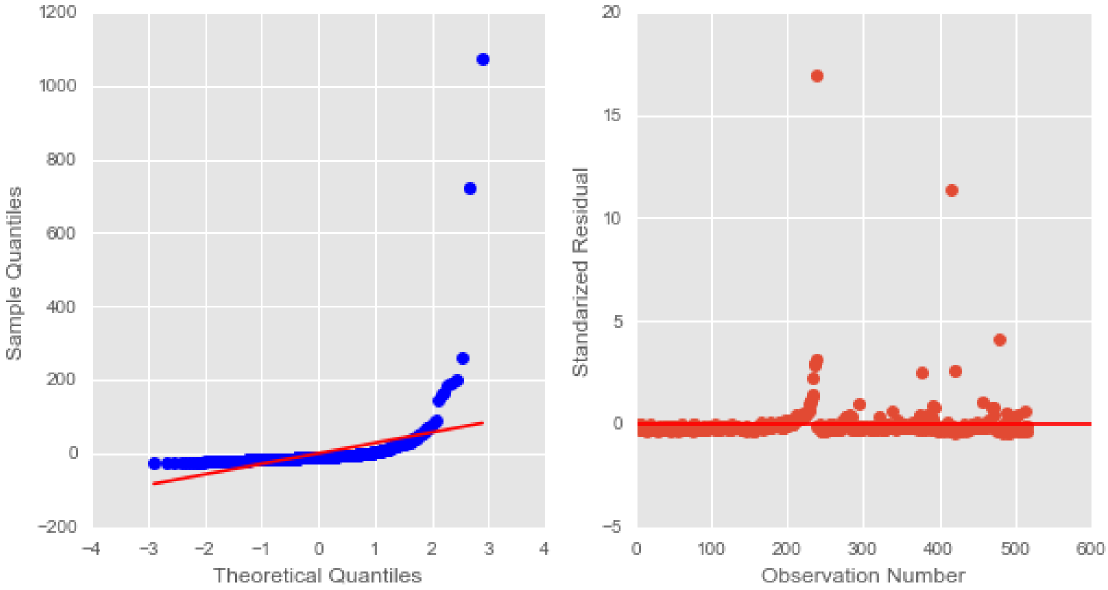
```
        stdres = pandas.DataFrame(model.resid_pearson)
        residuals_plot = plt.plot(stdres, 'o', ls = 'None')
        plt.axhline(y = 0, color = 'r')
        plt.ylabel('Standarized Residual')
        plt.xlabel('Observation Number')

        plt.show()
```

### 4.3 Generate linear models adding one explanatory variable a time

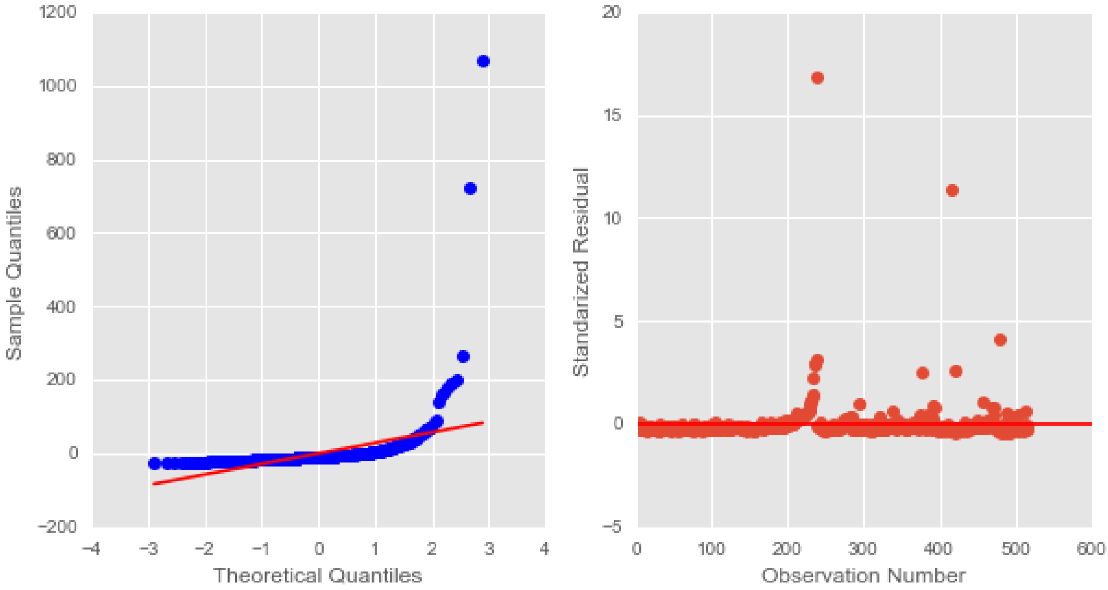
```
In [24]: explanatory_variables = None
         response_variable =     "area"

         saved_models = list()

         for variable in list(statistics[: number_of_columns - 1].index.values):
             if explanatory_variables == None:
                 explanatory_variables = variable
             else:
                 explanatory_variables += " + " + variable
             model = smf.ols(formula = response_variable + " ~ " + explanatory_variables,
                             data = fires).fit()
             saved_models.append(model)

             print_title('Model: ' + response_variable + " ~ " + explanatory_variables)
             print()
             print(model.summary())
             print_qqplot_and_residuals_plot(model)
             print()
```

```
+---------------------+
| 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:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0261
Time:                        22:34:27   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
```

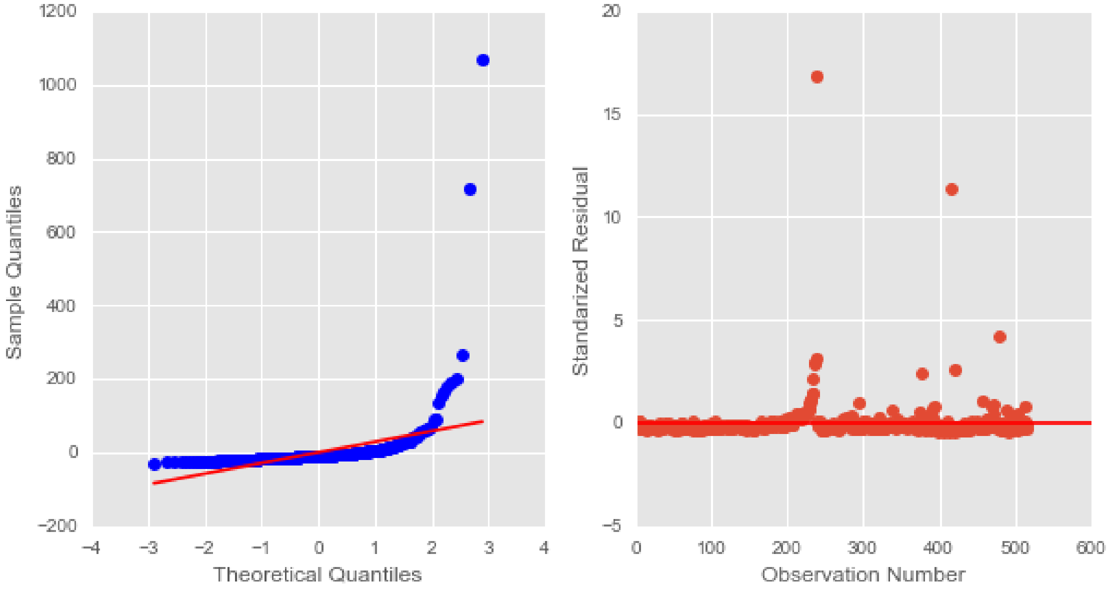
```
==============================================================================

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



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

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     2.692
Date:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0687
Time:                        22:34:28   Log-Likelihood:                -2877.8
No. Observations:                 517   AIC:                             5762.
Df Residuals:                     514   BIC:                             5774.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.790      4.604      0.000         7.365    18.329
temp           0.8811      0.566      1.556      0.120        -0.231     1.993
RH            -0.1293      0.201     -0.642      0.521        -0.525     0.267
==============================================================================
Omnibus:                      978.601   Durbin-Watson:                   1.648
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           790442.645
Skew:                          12.669   Prob(JB):                         0.00
Kurtosis:                     192.873   Cond. No.                         16.6
==============================================================================
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

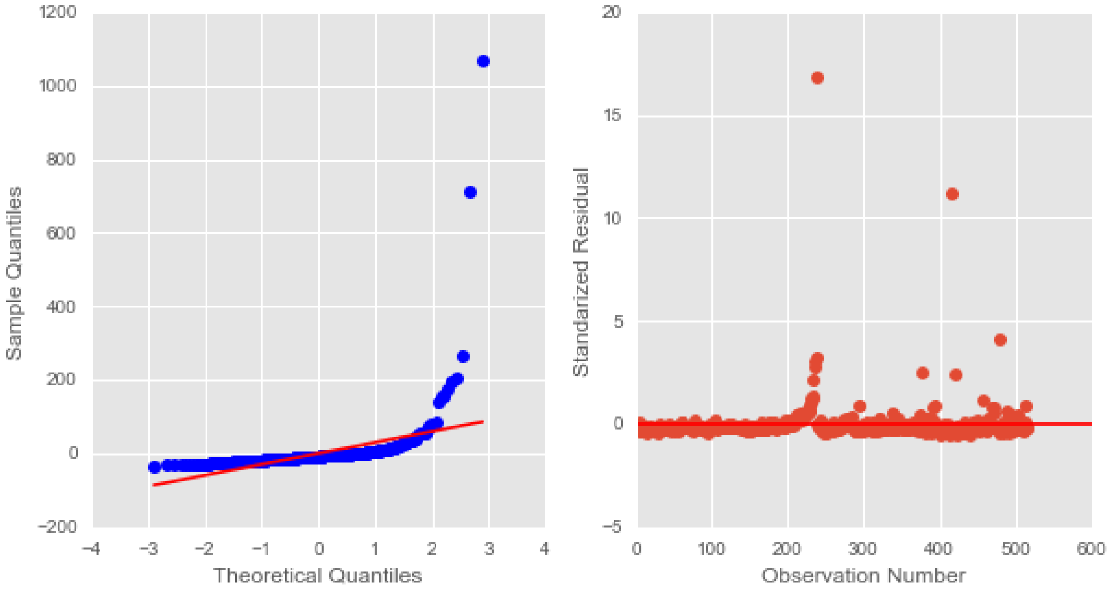


```
+-------------------------------+
| Model: area ~ temp + RH + DMC |
+-------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.013
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     2.182
Date:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0892
Time:                        22:34:29   Log-Likelihood:                -2877.2
No. Observations:                 517   AIC:                             5762.
Df Residuals:                     513   BIC:                             5779.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     12.8473      2.790      4.605      0.000         7.366    18.329
temp           0.4236      0.708      0.599      0.550        -0.967     1.814
RH            -0.2322      0.223     -1.041      0.298        -0.670     0.206
DMC            0.0589      0.055      1.077      0.282        -0.049     0.166
==============================================================================
Omnibus:                      978.623   Durbin-Watson:                   1.644
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           793249.169
Skew:                          12.668   Prob(JB):                         0.00
Kurtosis:                     193.216   Cond. No.                         64.1
==============================================================================
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



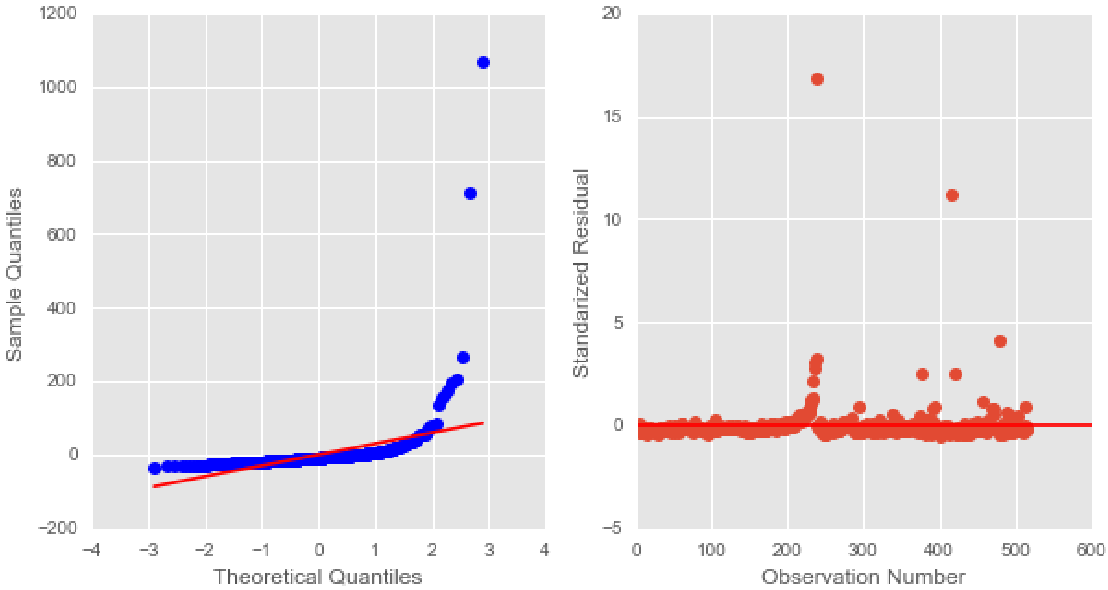
```
+-----------------------------------+
| Model: area ~ temp + RH + DMC + X |
+-----------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     2.351
Date:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0532
Time:                        22:34:30   Log-Likelihood:                -2875.7
No. Observations:                 517   AIC:                             5761.
Df Residuals:                     512   BIC:                             5783.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      5.3630      5.246      1.022      0.307        -4.944    15.670
temp           0.3855      0.707      0.545      0.586        -1.003     1.774
RH            -0.2656      0.223     -1.189      0.235        -0.705     0.173
DMC            0.0647      0.055      1.183      0.237        -0.043     0.172
X              2.0397      1.212      1.683      0.093        -0.341     4.420
==============================================================================
Omnibus:                      975.911   Durbin-Watson:                   1.648
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           783930.607
Skew:                          12.593   Prob(JB):                         0.00
Kurtosis:                     192.095   Cond. No.                         123.
```

```
==============================================================================

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



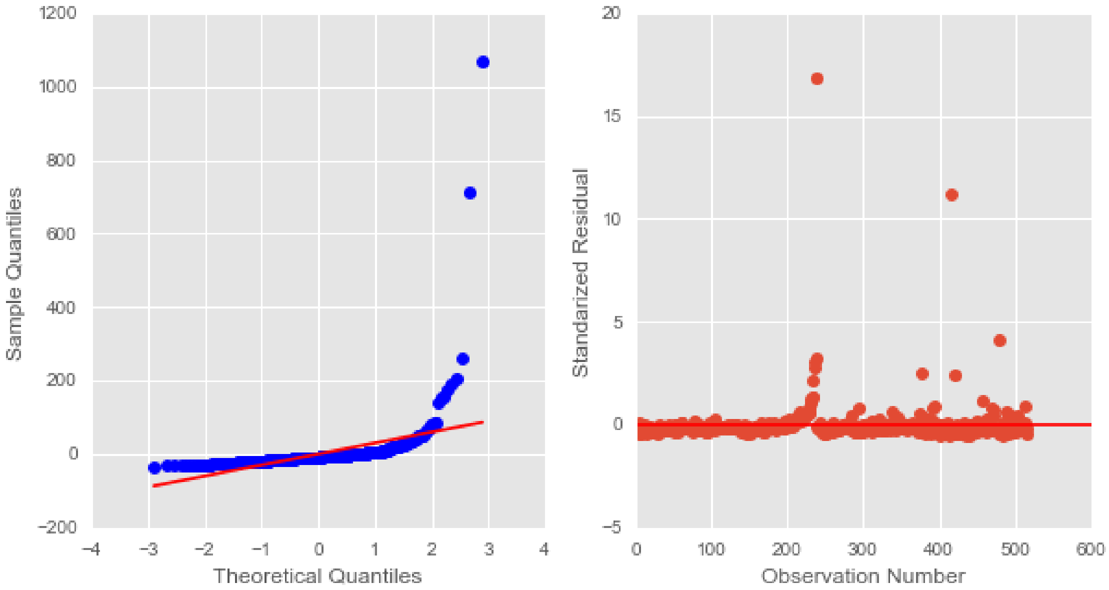
```
+-------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month |
+-------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     1.896
Date:                Tue, 14 Jun 2016   Prob (F-statistic):             0.0935
Time:                        22:34:31   Log-Likelihood:                -2875.7
No. Observations:                 517   AIC:                             5763.
Df Residuals:                     511   BIC:                             5789.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      2.5422     10.711      0.237      0.812       -18.501    23.585
temp           0.3575      0.713      0.501      0.617        -1.044     1.759
RH            -0.2637      0.224     -1.179      0.239        -0.703     0.176
DMC            0.0588      0.058      1.012      0.312        -0.055     0.173
X              2.0544      1.214      1.693      0.091        -0.330     4.439
month          0.4273      1.414      0.302      0.763        -2.351     3.206
==============================================================================
Omnibus:                      975.906   Durbin-Watson:                   1.647
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           783473.752
```

```
Skew:                          12.594   Prob(JB):                         0.00
Kurtosis:                     192.039   Cond. No.                         248.
==============================================================================

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



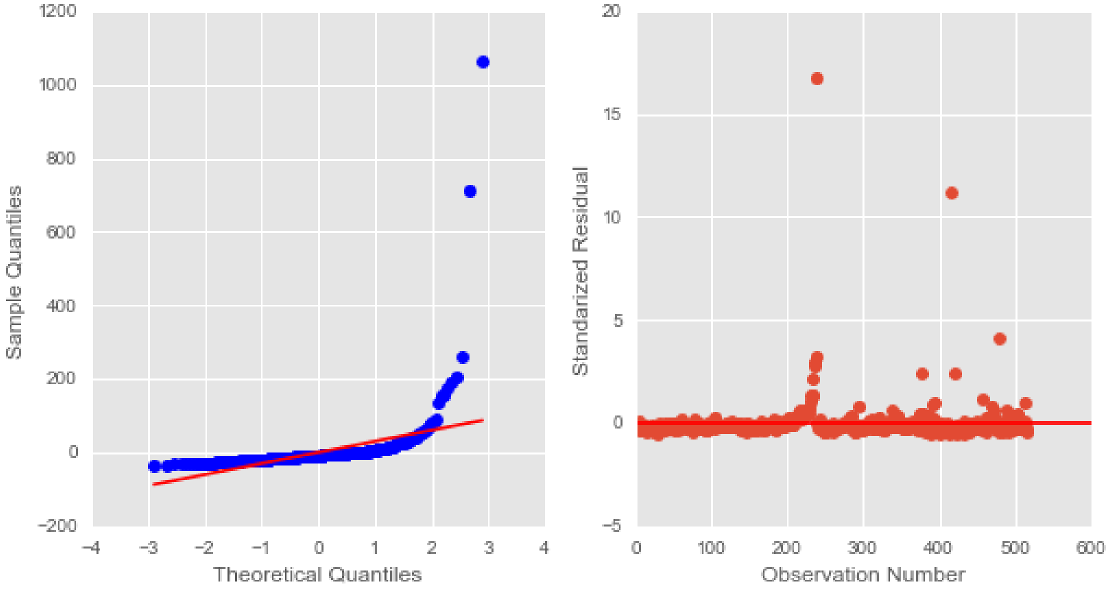
```
+-------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC |
+-------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.020
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     1.758
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.106
Time:                        22:34:32   Log-Likelihood:                -2875.2
No. Observations:                 517   AIC:                             5764.
Df Residuals:                     510   BIC:                             5794.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -12.4210     18.017     -0.689      0.491       -47.818    22.976
temp           0.5717      0.743      0.770      0.442        -0.888     2.031
RH            -0.2196      0.228     -0.964      0.335        -0.667     0.228
DMC            0.0929      0.067      1.390      0.165        -0.038     0.224
X              1.9615      1.217      1.612      0.108        -0.429     4.352
month          2.7906      2.690      1.037      0.300        -2.494     8.075
DC            -0.0316      0.031     -1.033      0.302        -0.092     0.028
```

```
==============================================================================
Omnibus:                      976.248   Durbin-Watson:                   1.655
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           786247.138
Skew:                          12.602   Prob(JB):                         0.00
Kurtosis:                     192.377   Cond. No.                     1.64e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.64e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



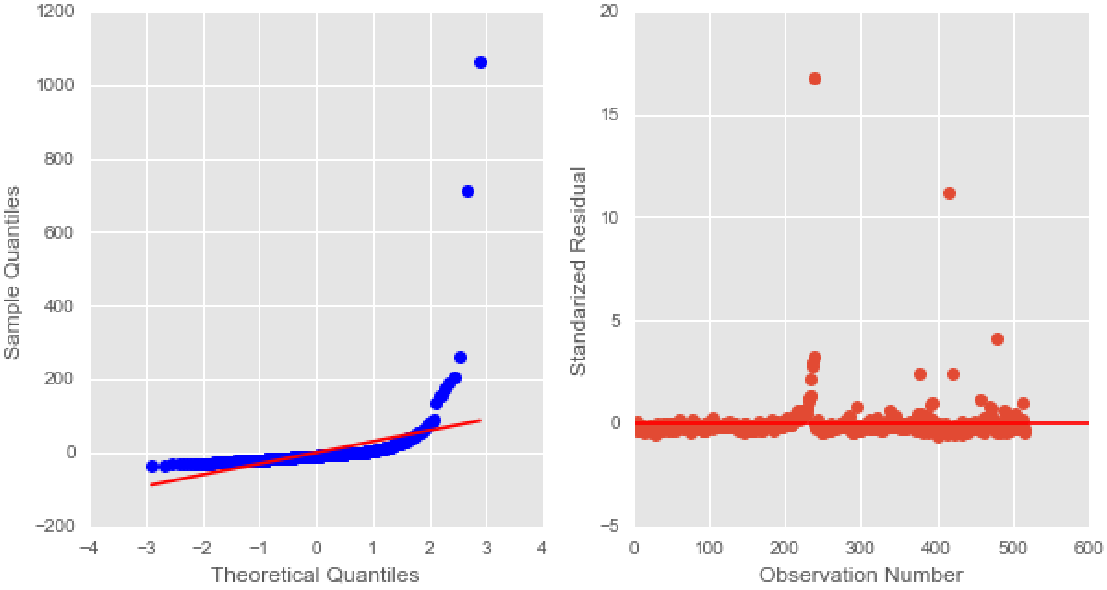
```
+----------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day |
+----------------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     1.642
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.121
Time:                        22:34:33   Log-Likelihood:                -2874.7
No. Observations:                 517   AIC:                             5765.
Df Residuals:                     509   BIC:                             5799.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -17.5970     18.784     -0.937      0.349       -54.500    19.306
temp           0.6091      0.744      0.819      0.413        -0.853     2.071
```

```
RH            -0.1956      0.229     -0.854      0.393       -0.646     0.254
DMC            0.0907      0.067      1.356      0.176       -0.041     0.222
X              1.9310      1.217      1.586      0.113       -0.461     4.323
month          3.0189      2.700      1.118      0.264       -2.286     8.324
DC            -0.0334      0.031     -1.091      0.276       -0.094     0.027
day            1.2812      1.314      0.975      0.330       -1.300     3.862
==============================================================================
Omnibus:                      973.549   Durbin-Watson:                   1.643
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           772827.176
Skew:                          12.531   Prob(JB):                         0.00
Kurtosis:                     190.744   Cond. No.                     1.71e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



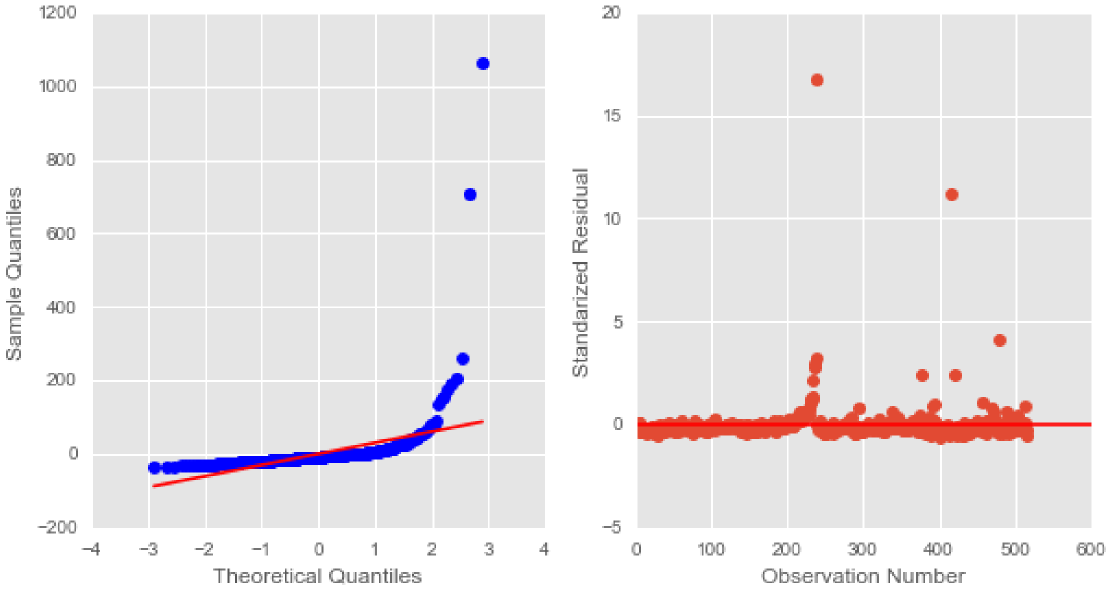
```
+------------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day + Y |
+------------------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     1.437
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.178
Time:                        22:34:34   Log-Likelihood:                -2874.7
No. Observations:                 517   AIC:                             5767.
Df Residuals:                     508   BIC:                             5806.
Df Model:                           8
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -17.8818     18.903     -0.946      0.345       -55.020    19.256
temp           0.6048      0.745      0.811      0.417        -0.859     2.069
RH            -0.1967      0.229     -0.857      0.392        -0.647     0.254
DMC            0.0895      0.067      1.327      0.185        -0.043     0.222
X              1.8191      1.441      1.262      0.207        -1.012     4.651
month          2.9834      2.714      1.099      0.272        -2.348     8.315
DC            -0.0328      0.031     -1.058      0.291        -0.094     0.028
day            1.2850      1.315      0.977      0.329        -1.299     3.869
Y              0.3977      2.734      0.145      0.884        -4.974     5.770
==============================================================================
Omnibus:                      973.369   Durbin-Watson:                   1.642
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           771951.439
Skew:                          12.527   Prob(JB):                         0.00
Kurtosis:                     190.637   Cond. No.                     1.72e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

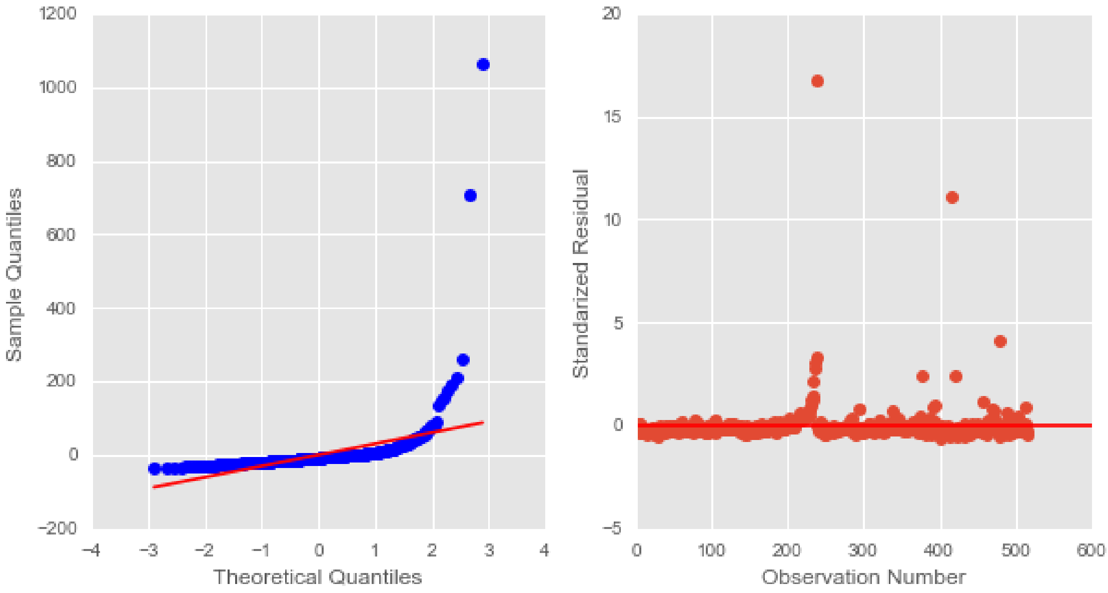


```
+-----------------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day + Y + FFMC |
+-----------------------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.023
```

```
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                     1.320
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.223
Time:                        22:34:34   Log-Likelihood:                -2874.5
No. Observations:                 517   AIC:                             5769.
Df Residuals:                     507   BIC:                             5811.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -19.0738     19.008     -1.003      0.316       -56.418    18.270
temp           0.6607      0.751      0.880      0.379        -0.815     2.136
RH            -0.2260      0.234     -0.965      0.335        -0.686     0.234
DMC            0.1008      0.070      1.444      0.149        -0.036     0.238
X              1.8695      1.444      1.294      0.196        -0.968     4.707
month          3.1573      2.729      1.157      0.248        -2.205     8.519
DC            -0.0341      0.031     -1.098      0.273        -0.095     0.027
day            1.3289      1.318      1.008      0.314        -1.260     3.918
Y              0.2888      2.741      0.105      0.916        -5.097     5.674
FFMC          -0.3772      0.596     -0.633      0.527        -1.547     0.793
==============================================================================
Omnibus:                      973.080   Durbin-Watson:                   1.643
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           771099.036
Skew:                          12.519   Prob(JB):                         0.00
Kurtosis:                     190.534   Cond. No.                     1.73e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

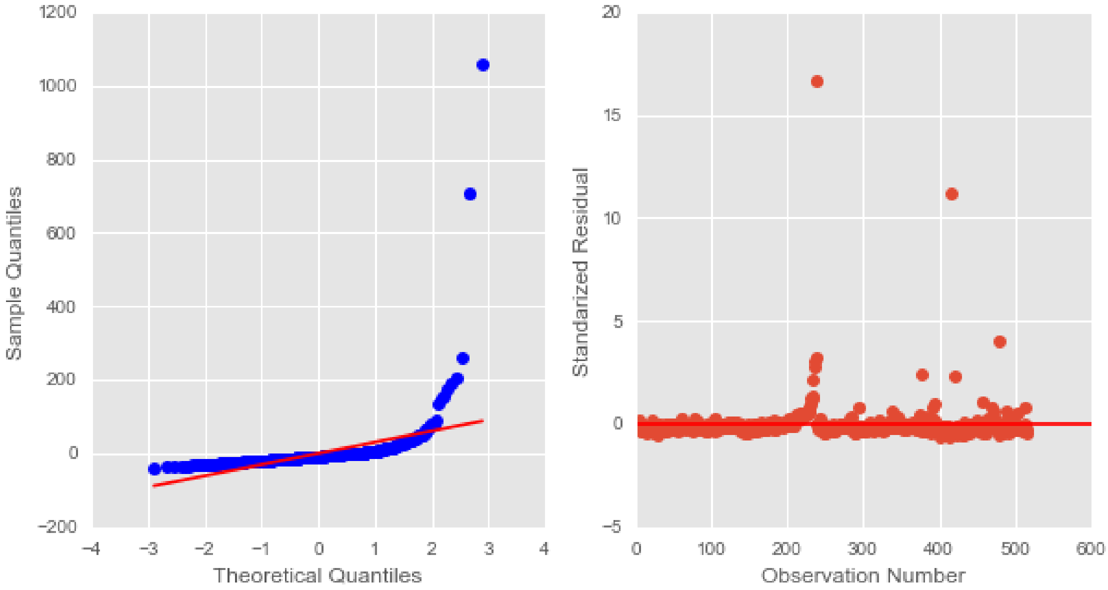


```
+------------------------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day + Y + FFMC + wind |
+------------------------------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     1.214
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.279
Time:                        22:34:35   Log-Likelihood:                -2874.3
No. Observations:                 517   AIC:                             5771.
Df Residuals:                     506   BIC:                             5817.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -17.3156     19.319     -0.896      0.371       -55.271    20.640
temp           0.7126      0.758      0.940      0.348        -0.777     2.202
RH            -0.2247      0.234     -0.959      0.338        -0.685     0.236
DMC            0.0958      0.071      1.359      0.175        -0.043     0.234
X              1.8451      1.446      1.276      0.203        -0.996     4.686
month          2.8663      2.788      1.028      0.304        -2.611     8.344
DC            -0.0301      0.032     -0.939      0.348        -0.093     0.033
day            1.3241      1.319      1.004      0.316        -1.267     3.915
Y              0.3890      2.750      0.141      0.888        -5.014     5.792
FFMC          -0.3932      0.597     -0.659      0.510        -1.566     0.779
wind           0.8642      1.660      0.521      0.603        -2.398     4.126
==============================================================================
Omnibus:                      972.761   Durbin-Watson:                   1.637
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           769820.664
Skew:                          12.510   Prob(JB):                         0.00
Kurtosis:                     190.377   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.
```



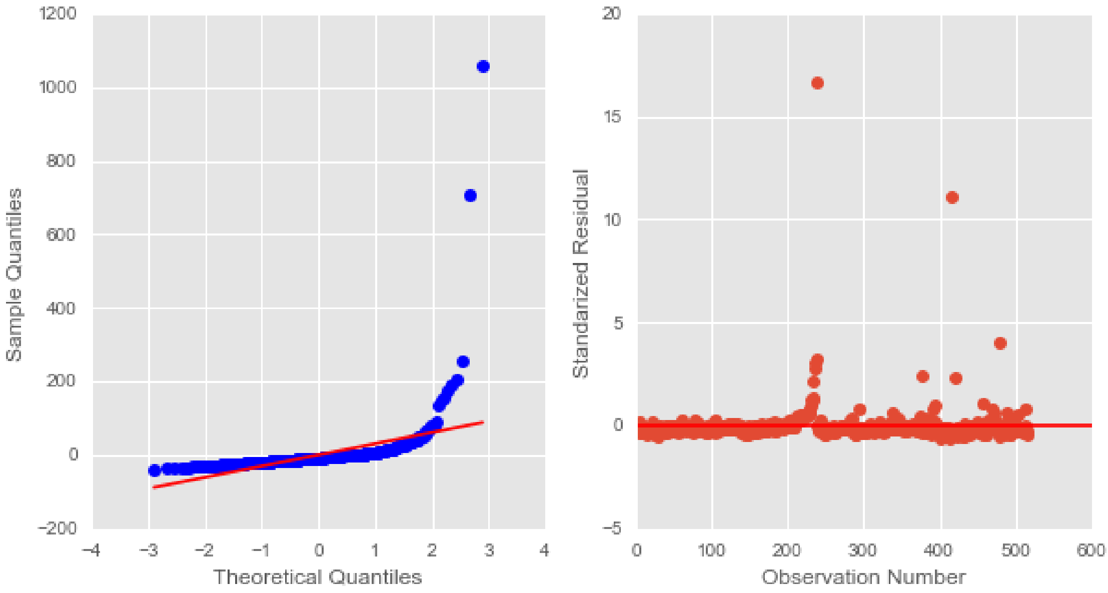
```
+------------------------------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day + Y + FFMC + wind + ISI |
+------------------------------------------------------------------------------+

                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.025
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     1.184
Date:                Tue, 14 Jun 2016   Prob (F-statistic):              0.295
Time:                        22:34:36   Log-Likelihood:                -2873.9
No. Observations:                 517   AIC:                             5772.
Df Residuals:                     505   BIC:                             5823.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -17.5792     19.323     -0.910      0.363       -55.543    20.385
temp           0.9189      0.789      1.165      0.245        -0.631     2.469
RH            -0.1885      0.237     -0.794      0.428        -0.655     0.278
DMC            0.0970      0.071      1.375      0.170        -0.042     0.236
X              1.8773      1.447      1.298      0.195        -0.965     4.720
month          2.9056      2.789      1.042      0.298        -2.573     8.384
DC            -0.0314      0.032     -0.978      0.329        -0.094     0.032
day            1.3346      1.319      1.012      0.312        -1.257     3.926
Y              0.3280      2.751      0.119      0.905        -5.077     5.733
FFMC          -0.1239      0.662     -0.187      0.851        -1.424     1.176
wind           1.1858      1.695      0.700      0.485        -2.145     4.516
ISI           -0.7271      0.771     -0.943      0.346        -2.242     0.787
==============================================================================
Omnibus:                      972.712   Durbin-Watson:                   1.642
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           769667.142
Skew:                          12.509   Prob(JB):                         0.00
Kurtosis:                     190.359   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.
```



```
+-------------------------------------------------------------------------------------+
| Model: area ~ temp + RH + DMC + X + month + DC + day + Y + FFMC + wind + ISI + rain |
+-------------------------------------------------------------------------------------+

                            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:                Tue, 14 Jun 2016   Prob (F-statistic):              0.364
Time:                        22:34:37   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
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
DMC            0.0966      0.071      1.369      0.172        -0.042     0.235
```

```
X              1.9002      1.450      1.311      0.191      -0.948     4.748
month          2.9004      2.791      1.039      0.299      -2.583     8.384
DC            -0.0315      0.032     -0.981      0.327      -0.095     0.032
day            1.3269      1.320      1.005      0.315      -1.267     3.921
Y              0.3241      2.754      0.118      0.906      -5.086     5.734
FFMC          -0.1127      0.663     -0.170      0.865      -1.415     1.190
wind           1.2321      1.702      0.724      0.470      -2.113     4.577
ISI           -0.7305      0.772     -0.947      0.344      -2.247     0.786
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.
```



From above results can be seen that after adding a second variable to the model (relative humidity –RH–), temperature –temp– is not longer statistically significant, so its a confounder variable: "temperature increases due to forest fire, or forest fire is helped by high temperatures?"

In all generated models we can see the summary of the multiple regression model, the quantile-quantile plot (qq-plot), left side, which "plots the quantiles of the residuals that we would theoretically see if the residuals followed a normal distribution, against the quantiles for the residuals estimated from our regression model", and, right side, the Standarized Residuals plot, which are, “simply, the residual values transformed to have a mean of zero and a standard deviation of one”

The qq-plots of our regression models shows a straight line with high deviations at the lower and higher quantiles, indicating the residuals do not follow a normal distribution.

The Standarized Residuals plot shows there are a group of observations which standarized residuals are greather than 2 standard deviations, and more than 3 standard deviations implying the presence of extreme outliers

```
In [26]: stdres = pandas.DataFrame(saved_models[11].resid_pearson)
         stdres[(stdres[0] < -2.5) | (stdres[0] > 2.5)].T
```

```
Out[26]:      235    236    237     238     415    479
         0 2.760 3.019 3.262 16.681 11.154 4.055
```

In all generated models, more than 1% of our observations has standardized residuals with an absolute value greater than 2.5, then there is evidence that the level of error within our model is unacceptable. That is the model is a fairly poor fit to the observed data.

```
In [27]: pct_outliers = list()
         for idx in range(len(saved_models)):
             stdres = pandas.DataFrame(saved_models[idx].resid_pearson)
             pct_outliers.append(round(len(stdres[(stdres[0] < -2.5) | (stdres[0] > 2.5)]) / len(fires)

         print(pct_outliers)
```

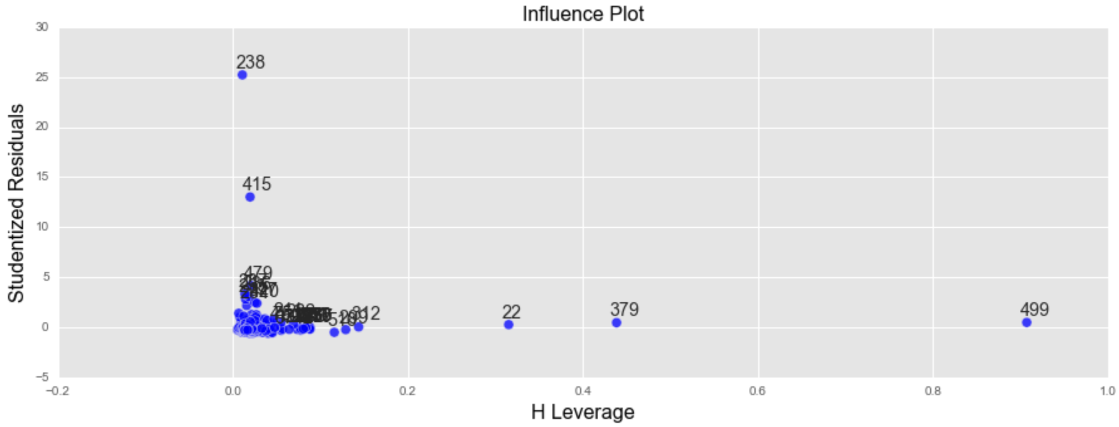
```
[0.015, 0.015, 0.014, 0.012, 0.012, 0.012, 0.012, 0.012, 0.012, 0.012, 0.012, 0.012]
```

The second warning of full model indicates “The condition number is large, 1.76e+03. This might indicate that there are strong multicollinearity or other numerical problems”, as it could be expected.

During data exploration we found that there is a medium to medium-high correlation between the average moisture content of deep, compact organic layers and the average moisture content of loosely compacted organic layers of moderate depth (DC-DMC: 0.682) and between the expected rate of fire spread and the moisture content of litter and other cured fine fuels (ISI- FFMC: 0.532). Also, there is a inverse medium correlation (-0.527) between temperature (temp) and relative humidity (RH). Other relationships are noted between temperature (temp) and FWI system components (FFMC, DCM, DC and ISI)

## 4.4 Leverage plot

```
In [29]: sm.graphics.influence_plot(model, size=8) #Leverage plot for full model
         plt.show()
```



**Leverage plot permits identify observations that have an unusually large influence on the estimation of the predicted value of the response variable, burned area, or that are outliers, or both. The graph of full model shows one observation with a very high influence (observation 499, with near 90%), one with medium influence (observation 379, with near 45%) and one with medium-low influence (observation 22, with near 32%). The rest of the observations have influence under 20%**

**The graph of full model also show us a group of ouliers. Note this extreme outliers are the same observations we found during data exploration: 238, 415, 479, plus a cloud of minor outliers (residuals outside range -2 to 2 standard deviations), but with low influence (< 5%) on the estimation of the regression model**

**No observations in this data are both high leverage and outliers.**

```
In [30]: fires.iloc[[22, 379, 499]]
```

```
Out[30]:      X  Y  month  day    FFMC      DMC       DC    ISI    temp     RH   wind  \
         22   6  2      5    0   3.655  -14.572 -347.940 47.078   2.111 -0.288  0.482
         379  3  3      0    0 -71.945 -109.772 -376.540 -9.022 -13.689 55.712 -3.118
         499  6  3      7    2   5.455   70.228  123.260  5.278   8.411 18.712  0.882

               rain   area
         22  -0.022  0.000
         379 -0.022  0.000
         499  6.378 10.820
```

```
In [ ]:
```