Week 9 Assignment

Importing the data into Python and Graphing the 4 plots with regression lines.

First we need to connect to the database. Make sure to change the path to the location of the file on your machine.

In [11]:
```
import sqlite3
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from scipy import stats
%matplotlib inline
conn = sqlite3.connect('D:/Erik/cunyweek9.sqlite')
cur = conn.cursor()
```

Next we create the helper functions that handle the analysis and generate the plots for our analysis. * read_table() reads in the data from a SQLite database and sorts the data. * plot_reg() generates the scaterplots and preforms the regression that is graphed with the scatterplots. * des_stats() generates descriptive statistics for the given data set. * least_squares() generates the data from an ordinarly least squares regression including residuals and Durbin-Watson results.

In [49]:
```
#Takes the data from the SQLite database and generates a set of points
def read_table(tablename):
    table = str(tablename)
    cur.execute('SELECT * FROM %s' %table)
    points = []
    for row in cur:
        points.append(row)
        points.sort()
    return points

#Takes a list of points(as a tuple or list), a title for the plot, and degree
 of the
#model and generates a scatterplot and a regression line for the data.
def plot_reg(ax, points, title, degree=1):
    x = []
    y = []
    for point in points:
        x.append(point[0])
        y.append(point[1])
    fit = np.polyfit(x, y, degree)
    yp = np.polyval(fit, x)
    ax.scatter(x, y) #Scatter plot of the data
```

```
    ax.plot(x,yp) #Plots the regression line
    ax.set_xlabel('x')
    ax.set_ylabel('y')
    ax.set_title(title)

#Generates and output of the descriptive statistics for the data set
def des_stats(points,title):
    n, min_max, mean, var, skew, kurt = stats.describe(points) #scipy functio
n for descriptive stats
    print title
    print 'Number of elements: %d' % n
    min_val = min_max[0][1]
    max_val = min_max[1][1]
    print 'Minimum: %f Maximum: %f' % (min_val, max_val)
    print 'Mean: %f' % mean[1]
    print 'Variance: %f' % var[1]
    print 'Skew : %f' % skew[1]
    print 'Kurtosis: %f' % kurt[1]
    print ''

#Generates the Ordininary Least Squares regression and prints the summary sta
ts.
def least_squares(points,title):
    x = []
    y = []
    for point in points:
        x.append(point[0])
        y.append(point[1])
    ols_model = sm.OLS(y, x) #OLS must be capatialized here.
    ols_fit = ols_model.fit()
    print title
    print ols_fit.summary()
    print ' '
```

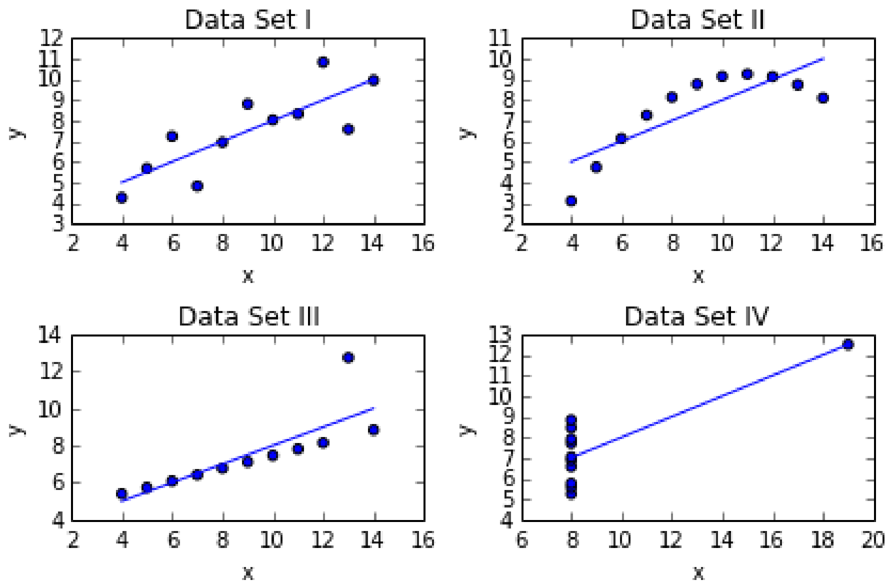
Finally we call the function for each table in the database and assign each to it's own list.

In [50]:
```
points1 = read_table('I')
points2 = read_table('II')
points3 = read_table('III')
points4 = read_table('IV')
```

Plotting the 4 data sets with regression lines.

In [51]:

```
#fig allows you to put multiple plots in one resizable figure
fig, ((ax1, ax2),(ax3,ax4)) = plt.subplots(nrows=2, ncols=2)
plot_reg(ax1, points1, 'Data Set I')
plot_reg(ax2, points2, 'Data Set II')
plot_reg(ax3, points3, 'Data Set III')
plot_reg(ax4, points4, 'Data Set IV')
plt.tight_layout()
```



The following shows the data from running the linear regression on the four data sets. If you notice the data sets are very similar for each of teh measures. This points out the weakness of looking at just the descriptive values for the data sets without actually looking at the graphs. From the descriptive stats we would asume that all of the data sets are modeled by the same function but it is obvious from the graphs above that this is not the case.

In [52]:

```
least_squares(points1,'Data Set I')
least_squares(points2,'Data Set II')
least_squares(points3,'Data Set III')
least_squares(points4,'Data Set IV')
```

```
Data Set I
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                      y   R-squared:                       0.96
3
Model:                            OLS   Adj. R-squared:                  0.95
9
Method:                 Least Squares   F-statistic:                     257.
9
Date:                Fri, 01 Aug 2014   Prob (F-statistic):           1.81e-0
8
Time:                        16:13:36   Log-Likelihood:                -20.04
4
No. Observations:                  11   AIC:                             42.0
9
```

```
Df Residuals:                      10   BIC:                             42.4
9
Df Model:                           1

==============================================================================
=
                 coef    std err          t      P>|t|      [95.0% Conf. Int.
]
------------------------------------------------------------------------------
-
x1             0.7968      0.050     16.059      0.000         0.686     0.90
7
==============================================================================
=
Omnibus:                        1.171   Durbin-Watson:                   1.57
6
Prob(Omnibus):                  0.557   Jarque-Bera (JB):                0.68
4
Skew:                          -0.572   Prob(JB):                        0.71
0
Kurtosis:                       2.573   Cond. No.                         1.0
0
==============================================================================
=

Data Set II
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                      y   R-squared:                       0.96
3
Model:                            OLS   Adj. R-squared:                  0.95
9
Method:                 Least Squares   F-statistic:                     257.
7
Date:                Fri, 01 Aug 2014   Prob (F-statistic):          1.82e-0
8
Time:                        16:13:36   Log-Likelihood:                -20.04
9
No. Observations:                  11   AIC:                             42.1
0
Df Residuals:                      10   BIC:                             42.5
0
Df Model:                           1

==============================================================================
=
                 coef    std err          t      P>|t|      [95.0% Conf. Int.
]
------------------------------------------------------------------------------
-
x1             0.7968      0.050     16.053      0.000         0.686     0.90
7
==============================================================================
=
```

```
Omnibus:                        4.616   Durbin-Watson:                   0.25
1
Prob(Omnibus):                  0.099   Jarque-Bera (JB):                2.20
2
Skew:                          -1.093   Prob(JB):                        0.33
3
Kurtosis:                       3.153   Cond. No.                         1.0
0
==============================================================================
=

Data Set III
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                      y   R-squared:                       0.96
3
Model:                            OLS   Adj. R-squared:                  0.95
9
Method:                 Least Squares   F-statistic:                     257.
7
Date:                Fri, 01 Aug 2014   Prob (F-statistic):          1.82e-0
8
Time:                        16:13:36   Log-Likelihood:                -20.04
7
No. Observations:                  11   AIC:                             42.0
9
Df Residuals:                      10   BIC:                             42.4
9
Df Model:                           1

==============================================================================
=
                 coef    std err          t      P>|t|      [95.0% Conf. Int.
]
------------------------------------------------------------------------------
-
x1             0.7967      0.050     16.053      0.000         0.686     0.90
7
==============================================================================
=
Omnibus:                        0.727   Durbin-Watson:                   1.54
5
Prob(Omnibus):                  0.695   Jarque-Bera (JB):                0.61
4
Skew:                          -0.215   Prob(JB):                        0.73
5
Kurtosis:                       1.925   Cond. No.                         1.0
0
==============================================================================
=

Data Set IV
                            OLS Regression Results
```

```
==============================================================================
=
Dep. Variable:                      y   R-squared:                       0.96
3
Model:                            OLS   Adj. R-squared:                  0.95
9
Method:                 Least Squares   F-statistic:                     258.
0
Date:                Fri, 01 Aug 2014   Prob (F-statistic):           1.81e-0
8
Time:                        16:13:36   Log-Likelihood:                -20.04
3
No. Observations:                  11   AIC:                             42.0
9
Df Residuals:                      10   BIC:                             42.4
8
Df Model:                           1

==============================================================================
=
                 coef    std err          t      P>|t|      [95.0% Conf. Int.
]
------------------------------------------------------------------------------
-
x1             0.7968      0.050     16.062      0.000         0.686     0.90
7
==============================================================================
=
Omnibus:                        0.522   Durbin-Watson:                   1.13
4
Prob(Omnibus):                  0.770   Jarque-Bera (JB):                0.46
8
Skew:                          -0.395   Prob(JB):                        0.79
1
Kurtosis:                       2.370   Cond. No.                         1.0
0
==============================================================================
=
```

The following shows the descriptive statistics for the four data sets. Notice that the 4 sets are very similar for the mean and variance but have very different skew and kutosis values.

In [56]:
```
#using scipy to do descriptive stats on the data set
des_stats(points1, 'Data Set I')
des_stats(points2, 'Data Set II')
des_stats(points3, 'Data Set III')
des_stats(points4, 'Data Set IV')
```

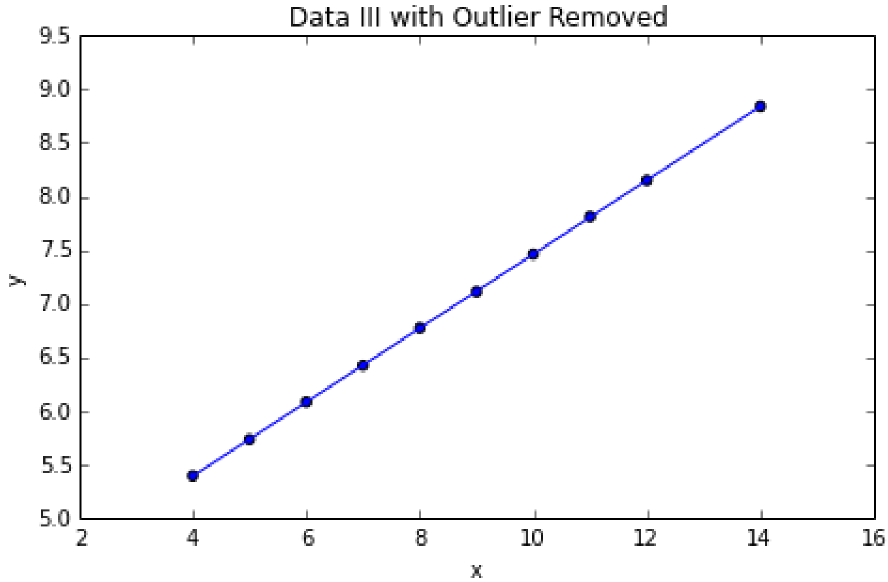
```
Data Set I
Number of elements: 11
Minimum: 4.260000 Maximum: 10.840000
Mean: 7.500909
Variance: 4.127269
Skew : -0.055808
Kurtosis: -0.820939

Data Set II
Number of elements: 11
Minimum: 3.100000 Maximum: 9.260000
Mean: 7.500909
Variance: 4.127629
Skew : -1.129108
Kurtosis: 0.007674

Data Set III
Number of elements: 11
Minimum: 5.390000 Maximum: 12.740000
Mean: 7.500000
Variance: 4.122620
Skew : 1.592231
Kurtosis: 2.130453

Data Set IV
Number of elements: 11
Minimum: 5.250000 Maximum: 12.500000
Mean: 7.500909
Variance: 4.123249
Skew : 1.293025
Kurtosis: 1.390789
```

In the following we evaluate the Data Set III again with the outlier removed.

In [54]:

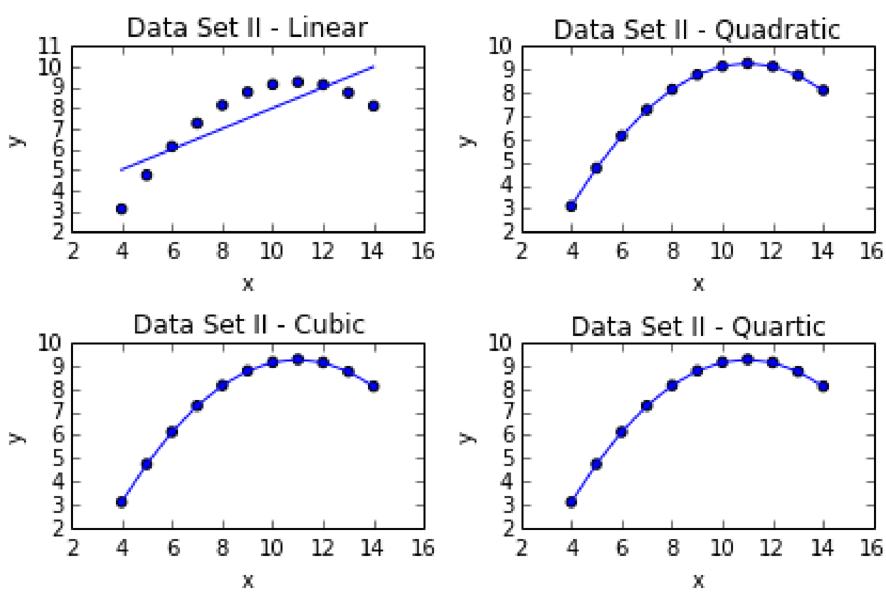
```
#Removing the outlier and sorting the data.
points5 = [(10,7.46), (8,6.77), (9, 7.11), (11, 7.81), (14, 8.84), (6, 6.08),
 (4, 5.39),
           (12, 8.15), (7, 6.42), (5, 5.73)]
points5.sort()

#Generating the plot.
fig, ax = plt.subplots()
plot_reg(ax, points5, 'Data III with Outlier Removed')
plt.tight_layout()

#Generating the regression statistics.
least_squares(points5, 'Data III with Outlier Removed')

#Generating the descriptive statistics.
des_stats(points5, 'Data III with Outlier Removed')
```

```
Data III with Outlier Removed
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                      y   R-squared:                       0.96
4
Model:                            OLS   Adj. R-squared:                  0.96
0
Method:                 Least Squares   F-statistic:                     242.
4
Date:                Fri, 01 Aug 2014   Prob (F-statistic):           8.17e-0
8
Time:                        16:14:06   Log-Likelihood:               -17.07
8
No. Observations:                  10   AIC:                             36.1
6
Df Residuals:                       9   BIC:                             36.4
6
Df Model:                           1

==============================================================================
=
                 coef    std err          t      P>|t|      [95.0% Conf. Int.
]
------------------------------------------------------------------------------
-
x1             0.7594      0.049     15.568      0.000         0.649     0.87
0
==============================================================================
=
Omnibus:                        0.416   Durbin-Watson:                   0.11
6
Prob(Omnibus):                  0.812   Jarque-Bera (JB):                0.47
9
Skew:                          -0.169   Prob(JB):                        0.78
7
Kurtosis:                       1.982   Cond. No.                         1.0
0
==============================================================================
=

Data III with Outlier Removed
Number of elements: 10
Minimum: 5.390000 Maximum: 8.840000
Mean: 6.976000
Variance: 1.224760
Skew : 0.169936
Kurtosis: -1.025550
```



In the following section we try fitting various polynomial functions to Data Set II. Notice that we use the same function as we used when doing the linear regression but this time we pass a third argument which sets the degree of the polynomial we are using.

In [55]:
```
#Trying four different polynomial regressions.
fig, ((ax1, ax2),(ax3,ax4)) = plt.subplots(nrows=2, ncols=2)
plot_reg(ax1, points2, 'Data Set II - Linear', 1)
plot_reg(ax2, points2, 'Data Set II - Quadratic', 2)
plot_reg(ax3, points2, 'Data Set II - Cubic', 3)
plot_reg(ax4, points2, 'Data Set II - Quartic', 4)
plt.tight_layout()
```



Generating descriptive statistics for these models is an area that we could expand on.