IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Regression Analysis Model Building - I

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Introduction

- Model building is the process of developing an estimated regression equation that describes the relationship between a dependent variable and one or more independent variables.
- The major issues in model building are finding the proper functional form of the relationship and selecting the independent variables to be included in the model.

General Linear Regression Model

- Suppose we collected data for one dependent variable *y* and *k* independent variables $x_1, x_2, \ldots, x_k$.
- Objective is to use these data to develop an estimated regression equation that provides the best relationship between the dependent and independent variables.

GENERAL LINEAR MODEL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon$$

- z_j (where $j = 1, 2, \ldots, p$) is a function of $x_1, x_2, \ldots, x_k$ (the variables for which data are collected).
- In some cases, each z_j may be a function of only one *x* variable.

Simple first-order model with one predictor variable

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Modelling Curvilinear Relationships

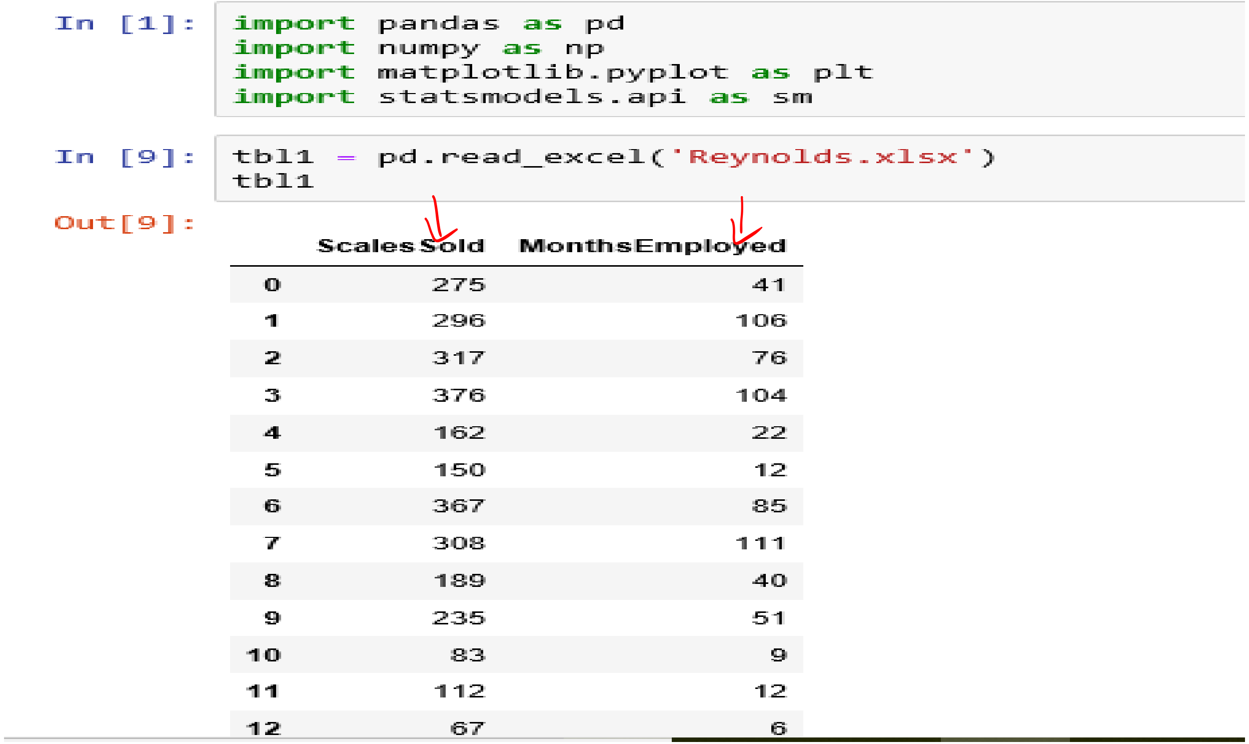
- To illustrate, let us consider the problem facing Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment.
- Managers at Reynolds want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold.
- Table in the next slide gives the number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm.

Sources: Statistics for Business and Economics,11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

Data

y	x
Scales Sold	Months Employed
275	41
296	106
317	76
376	104
162	22
150	12
367	85
308	111
189	40
235	51
83	9
112	12
67	6
325	56
189	19

Importing libraries and table



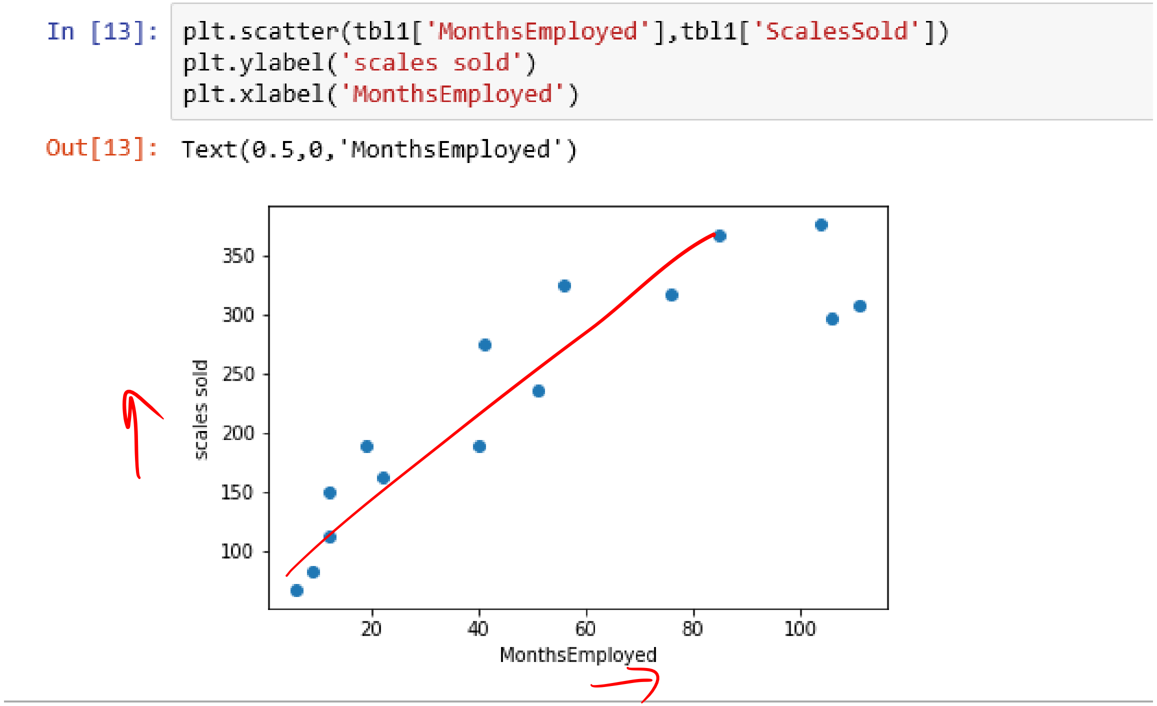
```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import statsmodels.api as sm
```

```
In [9]: tbl1 = pd.read_excel('Reynolds.xlsx')
        tbl1
```

Out[9]:

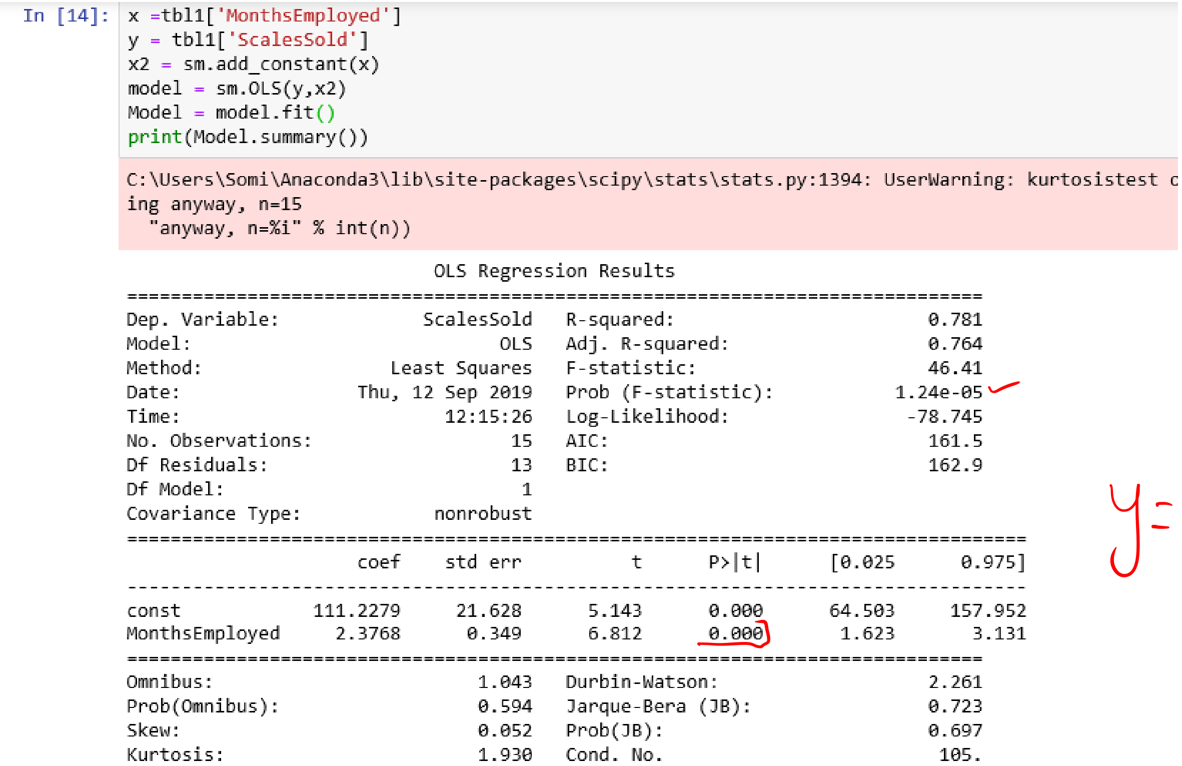
	Scales Sold	Months Employed
0	275	41
1	296	106
2	317	76
3	376	104
4	162	22
5	150	12
6	367	85
7	308	111
8	189	40
9	235	51
10	83	9
11	112	12
12	67	6

SCATTER DIAGRAM FOR THE REYNOLDS EXAMPLE



```
In [13]: plt.scatter(tbl1['MonthsEmployed'],tbl1['ScalesSold'])
         plt.ylabel('scales sold')
         plt.xlabel('MonthsEmployed')

Out[13]: Text(0.5,0,'MonthsEmployed')
```

Python code for the Reynolds example: first-order model



```
In [14]: x =tbl1['MonthsEmployed']
         y = tbl1['ScalesSold']
         x2 = sm.add_constant(x)
         model = sm.OLS(y,x2)
         Model = model.fit()
         print(Model.summary())
```

```
C:\Users\Somi\Anaconda3\lib\site-packages\scipy\stats\stats.py:1394: UserWarning: kurtosistest c
ing anyway, n=15
  "anyway, n=%i" % int(n))
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              ScalesSold   R-squared:                       0.781
Model:                             OLS   Adj. R-squared:                  0.764
Method:                  Least Squares   F-statistic:                     46.41
Date:                 Thu, 12 Sep 2019   Prob (F-statistic):           1.24e-05
Time:                         12:15:26   Log-Likelihood:                -78.745
No. Observations:                   15   AIC:                             161.5
Df Residuals:                       13   BIC:                             162.9
Df Model:                            1                                         
Covariance Type:             nonrobust                                         
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            111.2279     21.628      5.143      0.000      64.503     157.952
MonthsEmployed     2.3768      0.349      6.812      0.000       1.623       3.131
==================================================================================
Omnibus:                        1.043   Durbin-Watson:                   2.261
Prob(Omnibus):                  0.594   Jarque-Bera (JB):                0.723
Skew:                           0.052   Prob(JB):                        0.697
Kurtosis:                       1.930   Cond. No.                         105.
```

$y = 111.22 + 2.37$ Months Employed

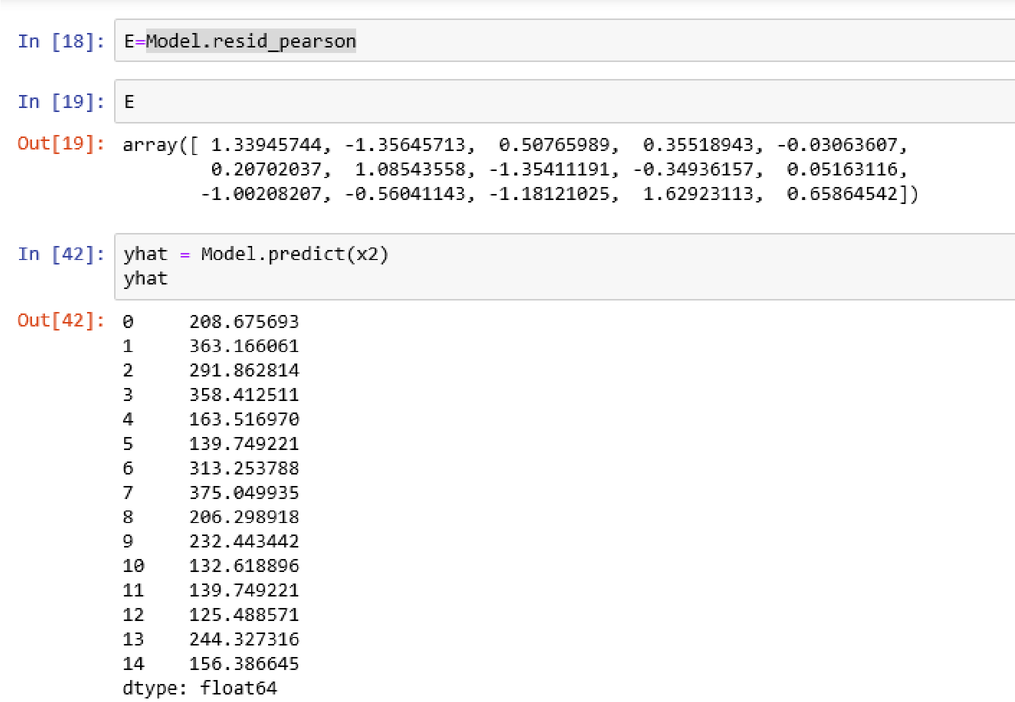
First-order regression equation

$$\text{Sales} = 111 + 2.38\,\text{Months}$$

where

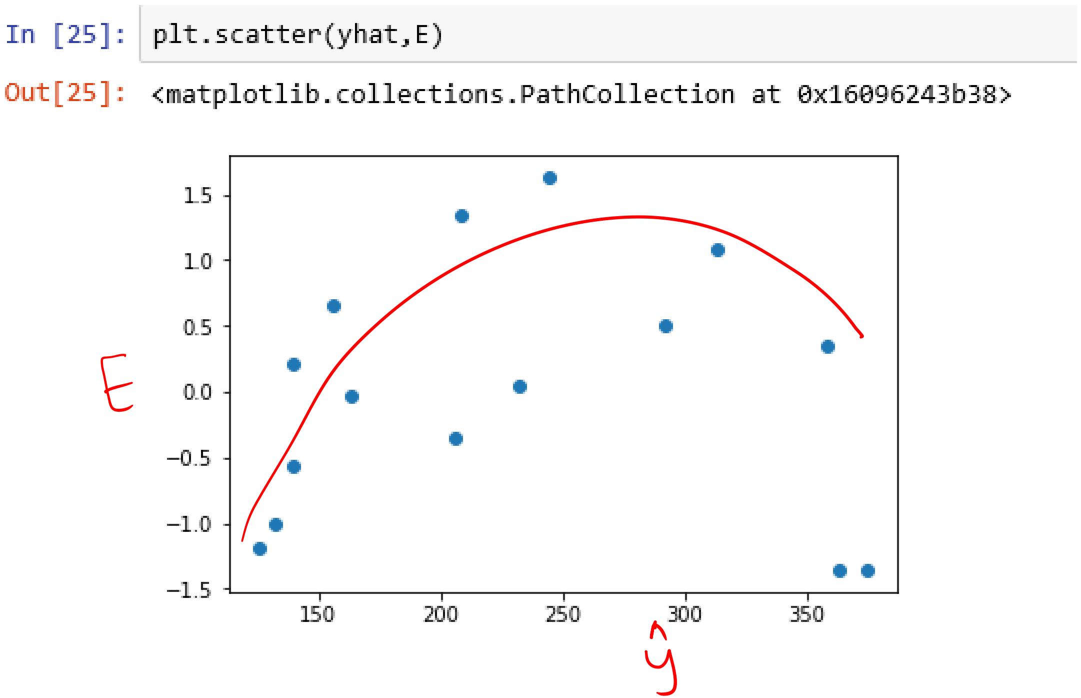
$$\begin{aligned}\text{Sales} &= \text{number of electronic laboratory scales sold}\\ \text{Months} &= \text{the number of months the salesperson has been employed}\end{aligned}$$

Standardized residual plot for the Reynolds example: first-order model



```
In [18]: E=Model.resid_pearson

In [19]: E

Out[19]: array([ 1.33945744, -1.35645713,  0.50765989,  0.35518943, -0.03063607,
                 0.20702037,  1.08543558, -1.35411191, -0.34936157,  0.05163116,
                -1.00208207, -0.56041143, -1.18121025,  1.62923113,  0.65864542])

In [42]: yhat = Model.predict(x2)
         yhat

Out[42]: 0     208.675693
         1     363.166061
         2     291.862814
         3     358.412511
         4     163.516970
         5     139.749221
         6     313.253788
         7     375.049935
         8     206.298918
         9     232.443442
         10    132.618896
         11    139.749221
         12    125.488571
         13    244.327316
         14    156.386645
         dtype: float64
```

Standardized residual plot for the Reynolds example: first-order model



```
In [25]: plt.scatter(yhat,E)

Out[25]: <matplotlib.collections.PathCollection at 0x16096243b38>
```

Need for curvilinear relationship

- Although the computer output shows that the relationship is significant (*p*-value .000) and that a linear relationship explains a high percentage of the variability in sales (R-sq 78.1%), the standardized residual plot suggests that a curvilinear relationship is needed.

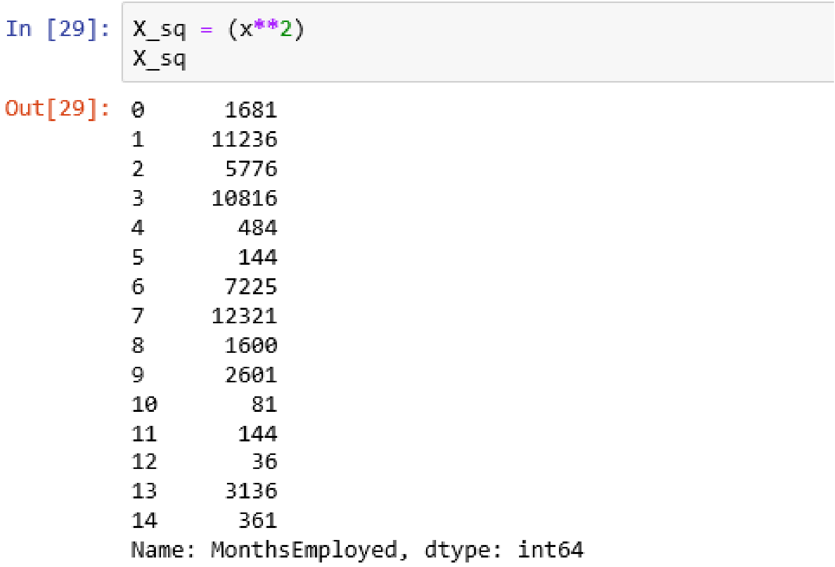
Second-order model with one predictor variable

- Set Z_1= x_1 and Z_2 = X^2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

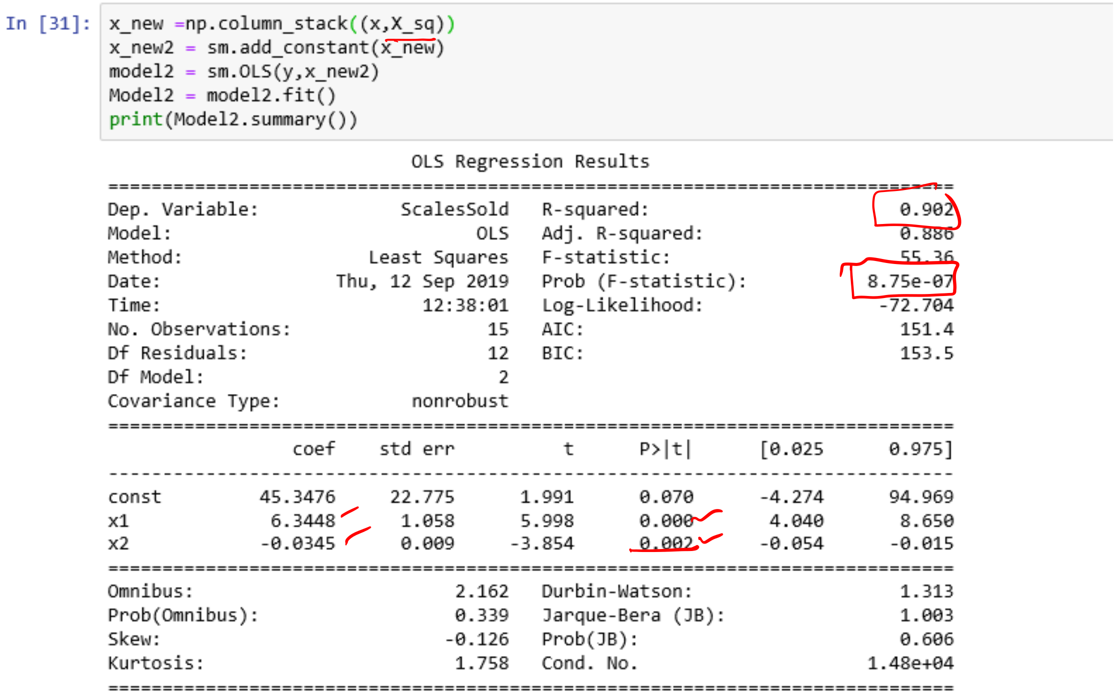
New Data set

- *The data for the* MonthsSq *independent variable is obtained by squaring the values of* Months.



```
In [29]: X_sq = (x**2)
         X_sq

Out[29]: 0      1681
         1     11236
         2      5776
         3     10816
         4       484
         5       144
         6      7225
         7     12321
         8      1600
         9      2601
         10       81
         11      144
         12       36
         13     3136
         14      361
         Name: MonthsEmployed, dtype: int64
```

Python output for the Reynolds example: second-order model



```
In [31]: x_new =np.column_stack((x,X_sq))
         x_new2 = sm.add_constant(x_new)
         model2 = sm.OLS(y,x_new2)
         Model2 = model2.fit()
         print(Model2.summary())
```

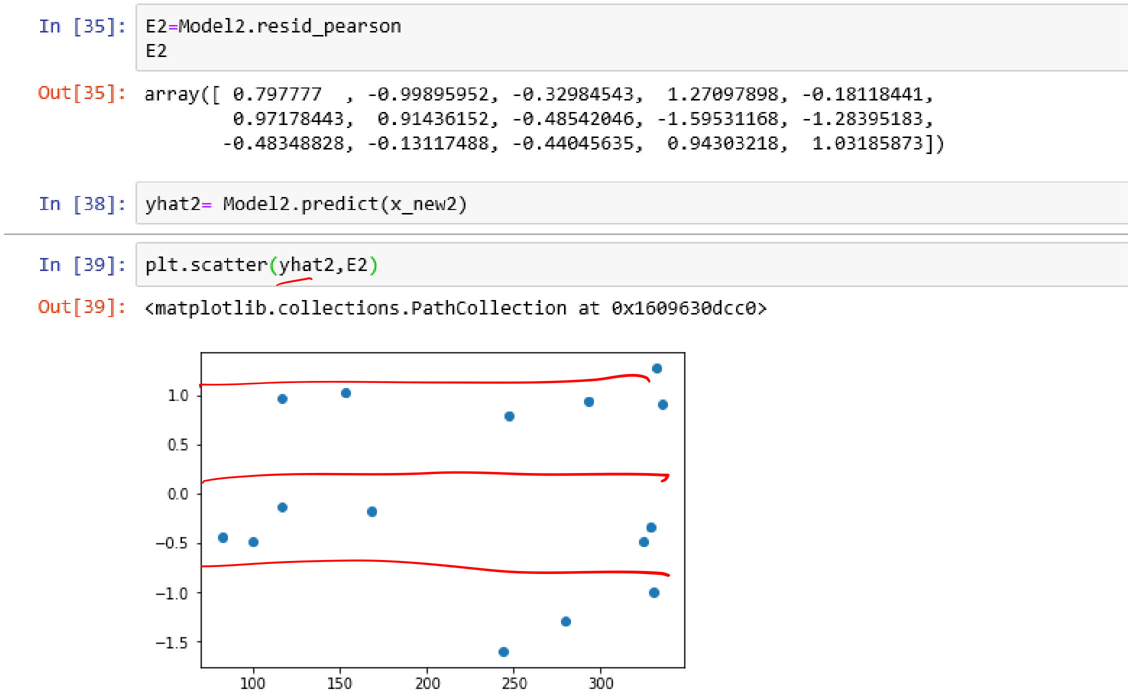
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             ScalesSold   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.886
Method:                 Least Squares   F-statistic:                     55.36
Date:                Thu, 12 Sep 2019   Prob (F-statistic):           8.75e-07
Time:                        12:38:01   Log-Likelihood:                -72.704
No. Observations:                  15   AIC:                             151.4
Df Residuals:                      12   BIC:                             153.5
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         45.3476     22.775      1.991      0.070      -4.274      94.969
x1             6.3448      1.058      5.998      0.000       4.040       8.650
x2            -0.0345      0.009     -3.854      0.002      -0.054      -0.015
==============================================================================
Omnibus:                        2.162   Durbin-Watson:                   1.313
Prob(Omnibus):                  0.339   Jarque-Bera (JB):                1.003
Skew:                          -0.126   Prob(JB):                        0.606
Kurtosis:                       1.758   Cond. No.                     1.48e+04
==============================================================================
```

Second-order regression model

$$\text{Sales} = 45.3 + 6.34\,\text{Months} - .0345\,\text{MonthsSq}$$

MonthsSq = the square of the number of months the salesperson has been employed

Standardized residual plot for the Reynolds example: second-order model



```
In [35]: E2=Model2.resid_pearson
         E2

Out[35]: array([ 0.797777  , -0.99895952, -0.32984543,  1.27097898, -0.18118441,
                 0.97178443,  0.91436152, -0.48542046, -1.59531168, -1.28395183,
                -0.48348828, -0.13117488, -0.44045635,  0.94303218,  1.03185873])

In [38]: yhat2= Model2.predict(x_new2)

In [39]: plt.scatter(yhat2,E2)

Out[39]: <matplotlib.collections.PathCollection at 0x1609630dcc0>
```

Interpretation second order model

- Figure corresponding standardized residual plot shows that the previous curvilinear pattern has been removed.
- At the .05 level of significance, the computer output shows that the overall model is significant (*p*-value for the *F* test is 0.000)
- Note also that the *p*-value corresponding to the *t*-ratio for MonthsSq (*p*-value .002) is less than .05
- Hence we can conclude that adding MonthsSq to the model involving Months is significant.
- With an R-sq(adj) value of 88.6%, we should be pleased with the fit provided by this estimated regression equation.

Meaning of linearity in GLM

- In multiple regression analysis the word *linear* in the term “general linear model” refers only to the fact that β_0, β_1, . . . , β_p all have exponents of β_1
- It does not imply that the relationship between *y* and the *xi*’s is linear.
- Indeed, we have seen one example of how equation general linear model can be used to model a curvilinear relationship.

Thank you