**import** numpy **as** np

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

**import** statsmodels.api **as** sm

**from** numpy.polynomial.polynomial **import** polyfit

**from** sklearn.linear\_model **import** LinearRegression

**1 - Business Problem**

**\_Salary\_hike -> Build a prediction model for Salary*hike***

**2 - Data collection and description**

In [2]:

df **=** pd**.**read\_csv("C:/Users/Sohail/OneDrive/Data Science/Assignments/Simple Linear Regression/Raw Data/Salary\_Data.csv")

**Scatter Plot**

In [3]:

x **=** df['YearsExperience']

y **=** df['Salary']

In [4]:

b, m **=** polyfit(x, y, 1)

plt**.**scatter(x, y)

plt**.**plot(x, y, '.')

plt**.**plot(x, b **+** m **\*** x, '-')

plt**.**title('Scatter plot Salary Hike')

plt**.**xlabel('Years of Experience')

plt**.**ylabel('Salary Hike')

plt**.**show()

As displayed in the scatter plot,but there is positive correlation between Salary Hike and Years of experience

**Correlation Analysis**

In [5]:

corr **=** np**.**corrcoef(x, y)

Corr  
array([[1. , 0.97824162], [0.97824162, 1. ]])

The correlation between Salary Hike and Years of experience is high (98%)

**3 - Regression Model**

**1 - No transformation**

In [6]:

model **=** sm**.**OLS(y, x)**.**fit()

predictions **=** model**.**predict(x)

In [7]:

model**.**summary()

Out[7]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.973 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.972 |
| **Method:** | Least Squares | **F-statistic:** | 1048. |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 2.56e-24 |
| **Time:** | 20:24:55 | **Log-Likelihood:** | -327.28 |
| **No. Observations:** | 30 | **AIC:** | 656.6 |
| **Df Residuals:** | 29 | **BIC:** | 658.0 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 1.325e+04 | 409.401 | 32.376 | 0.000 | 1.24e+04 | 1.41e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.610 | **Durbin-Watson:** | 0.323 |
| **Prob(Omnibus):** | 0.737 | **Jarque-Bera (JB):** | 0.671 |
| **Skew:** | -0.121 | **Prob(JB):** | 0.715 |
| **Kurtosis:** | 2.308 | **Cond. No.** | 1.00 |

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

**2 - Log Transformation of X**

In [8]:

x\_log **=** np**.**log(df['YearsExperience'])

In [9]:

model **=** sm**.**OLS(y, x\_log)**.**fit()

predictions **=** model**.**predict(x\_log)

In [10]:

model**.**summary()

Out[10]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.979 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.978 |
| **Method:** | Least Squares | **F-statistic:** | 1338. |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 8.06e-26 |
| **Time:** | 20:26:07 | **Log-Likelihood:** | -323.70 |
| **No. Observations:** | 30 | **AIC:** | 649.4 |
| **Df Residuals:** | 29 | **BIC:** | 650.8 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 4.909e+04 | 1341.796 | 36.583 | 0.000 | 4.63e+04 | 5.18e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 10.249 | **Durbin-Watson:** | 0.421 |
| **Prob(Omnibus):** | 0.006 | **Jarque-Bera (JB):** | 8.950 |
| **Skew:** | 1.106 | **Prob(JB):** | 0.0114 |
| **Kurtosis:** | 4.507 | **Cond. No.** | 1.00 |

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

**3 - Log Transformation of Y**

In [11]:

y\_log **=** np**.**log(df['Salary'])

In [12]:

model **=** sm**.**OLS(y\_log, x)**.**fit()

predictions **=** model**.**predict(x)

In [13]:

model**.**summary()

Out[13]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.809 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.802 |
| **Method:** | Least Squares | **F-statistic:** | 122.8 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 6.09e-12 |
| **Time:** | 20:26:10 | **Log-Likelihood:** | -90.160 |
| **No. Observations:** | 30 | **AIC:** | 182.3 |
| **Df Residuals:** | 29 | **BIC:** | 183.7 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 1.6755 | 0.151 | 11.083 | 0.000 | 1.366 | 1.985 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 3.609 | **Durbin-Watson:** | 0.016 |
| **Prob(Omnibus):** | 0.165 | **Jarque-Bera (JB):** | 2.045 |
| **Skew:** | -0.389 | **Prob(JB):** | 0.360 |
| **Kurtosis:** | 1.985 | **Cond. No.** | 1.00 |

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

**4 - Log Transformation of X & Y**

In [14]:

model **=** sm**.**OLS(y\_log, x\_log)**.**fit()

predictions **=** model**.**predict(x\_log)

In [15]:

model**.**summary()

Out[15]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.878 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.874 |
| **Method:** | Least Squares | **F-statistic:** | 209.1 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 8.60e-15 |
| **Time:** | 20:26:14 | **Log-Likelihood:** | -83.410 |
| **No. Observations:** | 30 | **AIC:** | 168.8 |
| **Df Residuals:** | 29 | **BIC:** | 170.2 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 6.4461 | 0.446 | 14.461 | 0.000 | 5.534 | 7.358 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.550 | **Durbin-Watson:** | 0.026 |
| **Prob(Omnibus):** | 0.279 | **Jarque-Bera (JB):** | 2.163 |
| **Skew:** | 0.640 | **Prob(JB):** | 0.339 |
| **Kurtosis:** | 2.697 | **Cond. No.** | 1.00 |

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

**5 - Sq Root Transformation of X**

In [16]:

x\_sqrt **=** np**.**sqrt(df['YearsExperience'])

In [17]:

model **=** sm**.**OLS(y, x\_sqrt)**.**fit()

predictions **=** model**.**predict(x\_sqrt)

In [18]:

model**.**summary()

Out[18]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.989 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.989 |
| **Method:** | Least Squares | **F-statistic:** | 2697. |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 3.62e-30 |
| **Time:** | 20:26:18 | **Log-Likelihood:** | -313.35 |
| **No. Observations:** | 30 | **AIC:** | 628.7 |
| **Df Residuals:** | 29 | **BIC:** | 630.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 3.48e+04 | 670.056 | 51.932 | 0.000 | 3.34e+04 | 3.62e+04 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 5.654 | **Durbin-Watson:** | 0.734 |
| **Prob(Omnibus):** | 0.059 | **Jarque-Bera (JB):** | 1.849 |
| **Skew:** | -0.040 | **Prob(JB):** | 0.397 |
| **Kurtosis:** | 1.786 | **Cond. No.** | 1.00 |

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

**6 - Square Root Transformation of Y**

In [19]:

y\_sqrt **=** np**.**sqrt(df['Salary'])

In [20]:

model **=** sm**.**OLS(y\_sqrt, x)**.**fit()

predictions **=** model**.**predict(x)

In [21]:

model**.**summary()

Out[21]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.906 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.902 |
| **Method:** | Least Squares | **F-statistic:** | 278.1 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 2.12e-16 |
| **Time:** | 20:26:23 | **Log-Likelihood:** | -175.75 |
| **No. Observations:** | 30 | **AIC:** | 353.5 |
| **Df Residuals:** | 29 | **BIC:** | 354.9 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 43.7142 | 2.621 | 16.676 | 0.000 | 38.353 | 49.076 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.437 | **Durbin-Watson:** | 0.046 |
| **Prob(Omnibus):** | 0.296 | **Jarque-Bera (JB):** | 1.859 |
| **Skew:** | -0.445 | **Prob(JB):** | 0.395 |
| **Kurtosis:** | 2.167 | **Cond. No.** | 1.00 |

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

**7 - Square Root Transformation of X & Y**

In [22]:

model **=** sm**.**OLS(y\_sqrt, x\_sqrt)**.**fit()

predictions **=** model**.**predict(x\_sqrt)

In [23]:

model**.**summary()

Out[23]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Salary | **R-squared (uncentered):** | 0.988 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.987 |
| **Method:** | Least Squares | **F-statistic:** | 2338. |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 2.81e-29 |
| **Time:** | 20:26:26 | **Log-Likelihood:** | -145.12 |
| **No. Observations:** | 30 | **AIC:** | 292.2 |
| **Df Residuals:** | 29 | **BIC:** | 293.6 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **YearsExperience** | 118.8652 | 2.458 | 48.352 | 0.000 | 113.837 | 123.893 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.762 | **Durbin-Watson:** | 0.231 |
| **Prob(Omnibus):** | 0.251 | **Jarque-Bera (JB):** | 2.152 |
| **Skew:** | 0.653 | **Prob(JB):** | 0.341 |
| **Kurtosis:** | 2.881 | **Cond. No.** | 1.00 |

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

**4 - Output Interpretation**

We will use Model 5 as it has the best R square value

1 - p-value < 0.01  
Thus the model is accepted

2 - coefficient == 3.48e+04  
Thus if the value of years of experience is increased by 1, the predicted value of Salary hike will increase by 3.48e+04

3 - Adj. R-sqared == 0.989  
Thus the model explains 98.9% of the variance in dependent variable