**SIMPLE LINEAR RIGRESSION**

**Business Problem** = Build a prediction model for Churn out rate.

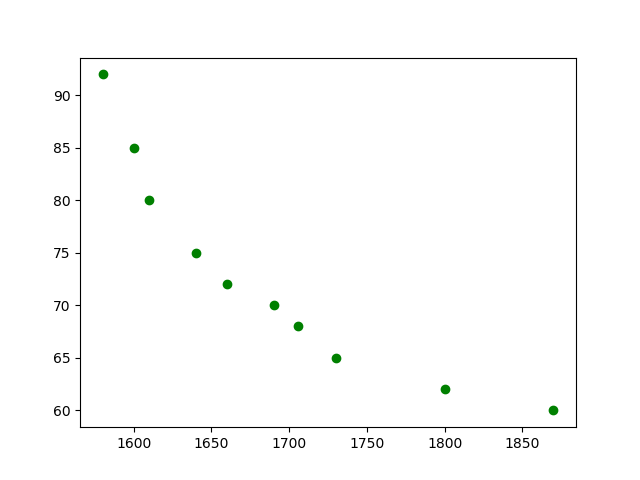
* **Name of the File: -** emp\_data.csv
* **Size of the File: -** 118 bytes
* **Data: -** 10 Observation, 2 Variable
* **Column Name: -** Salary\_hike, Churn\_out\_rate

**Exploratory data Analysis** =

* **Skewness: -** Both variable have Right Skewed data
* **Outliers: -**  Data don’t have Outliers
* **Missing Value: -** Data don’t have Missing Values
* **Normality: -** Data are not normal
* **Transformation: -**  Required

**Liner Regression Model =**

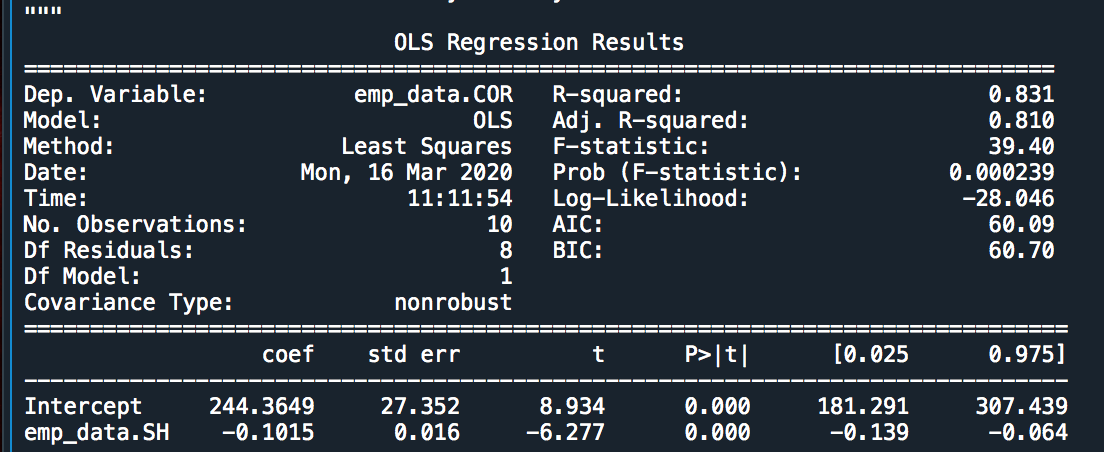
**Scatter plot =** From below scatter plot we can assume that the direction of correlation of data is negative and curvilinear linear in nature.



**Correlation Coefficient (r) =** Correlation Coefficient In between Salary hike and Churn out rate is equal to -0.91 this means the data have strong negative correlation.

**Model Building =**

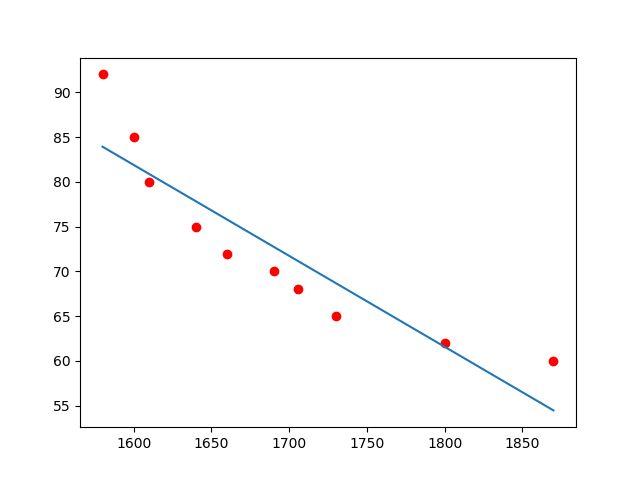
* **Summary: -**

****

* **Intersect(B0): -** 244.36
* **Slope(B1): -** -0.1015
* **Coefficient of**

**Determination (R2): -** 0.831-Moderate Correlation

* **Pvalue: -** P value less than 0.05 so we can use B1 and B0 for model building.
* **Model: - Churn out rate** = 244.36 – 0.1015(Salary hike)
* **RMSE:** - 3.99

**Final Scatter Plot with Best Fit line**: -

As P value is significant and the Determination of coefficient is moderate but we can observe from above plot the distribution of data is curvilinear and the line of base fit is straight due which our model might not give accurate result. Therefore, for increasing accuracy we have to try several transformations.

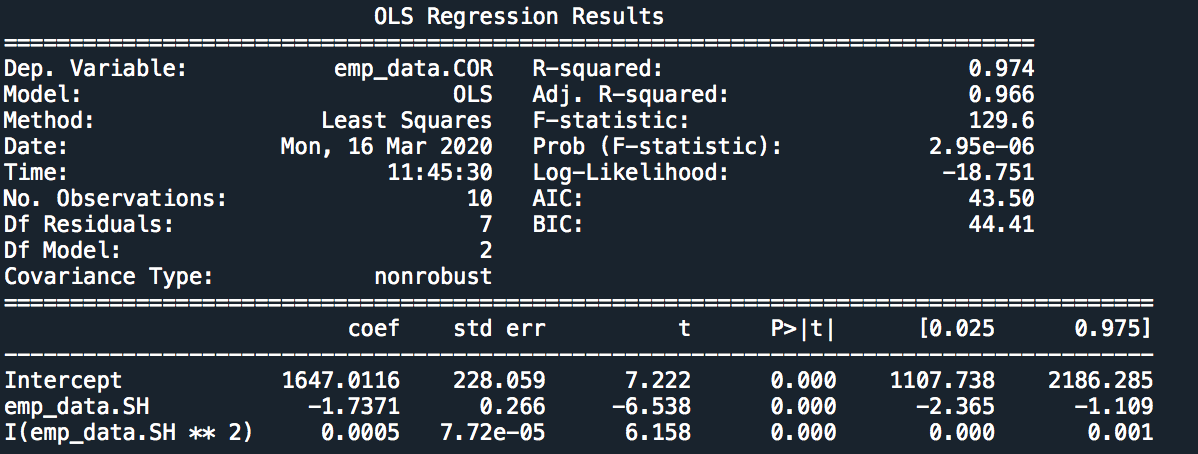
**Transformation Table**: -

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Output** | **Input** | **Cor (r)** | **R2** | **RMSE** | **Problem** | **Model** |
| Churn out rate | Log(Salary hike) | -0.92 | 0.85 | 3.78 | Curvilinear | Logarithmic Model |
| Log(Churn out rate) | Salary hike | -0.93 | 0.87 | 3.54 | Curvilinear | Exponential Model |
| Churn out rate | Polynomial  (Salary hike) | -0.90 | 0.97 | 1.57 | Base fit | Polynomial Model |

From above table we can conclude that the Polynomials transformation give us accurate model as compare to other transformations so we use Polynomials transformation for model building.

**Polynomials Liner Regression Model =**

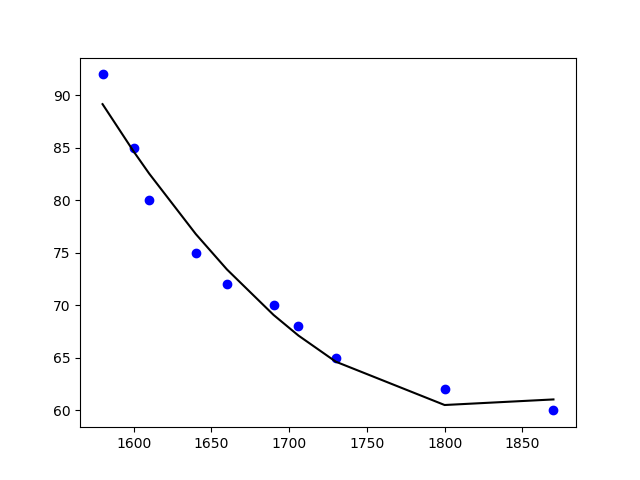
**Correlation Coefficient (r) =** Correlation Coefficient In between Salary hike and square of Churn out rate is equal to -0.90 this means the data have strong negative correlation.

**Model Building =**

* **Summary: -**
* **Intersect(B0): -** 1647.011
* **Slope(B1, B2): -** -1.7371, 0.005
* **Coefficient of**

**Determination (R2): -** 0.974 - Strong Correlation

* **Pvalue: -** P value less than 0.05 so we can use B1 and B0 for model building.
* **Model: - Churn out rate** = 1647.01 – 1.737 (Salary hike) + 0.0005 (Salary hike)2
* **RMSE:** - 1.57

**Final Scatter Plot with Best Fit Polynomial line**: -

As P value is significant and the Determination of coefficient is strong also we can see from above scatter plot that the new polynomial line is base fit to the distribution of data after considering all we can use this model for Prediction.

**Final Models with 95% confidence interval: -**

* **Churn out rate** = 1107.8 – 2.365 (Salary hike) + 0.000 (Salary hike)2
* **Churn out rate** = 2186.3 – 1.109 (Salary hike) + 0.001 (Salary hike)2

**Python code file**: - [Churn\_out Rate Analysis .py](https://github.com/nilaydeshmukh0/Simple-Linear-Regression-With-EDA/blob/master/Churn_out%20Rate%20Analysis/Churn_out%20Rate%20Analysis%20.py)