**Simple Linear Regression**

1) Calories\_consumed-> predict weight gained using calories consumed.

Assuming EDA has done and it does not have any outlier.

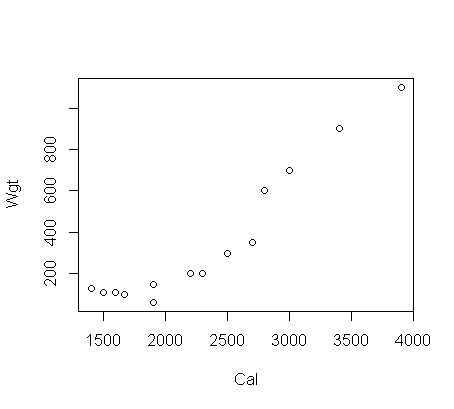
**Data Description :**

X = Calories - > Input

Y = Weights Gained -> Output

Both X and Y are in Continuous Data. We are proceeding with Simple Linear Regression.

**Scatter Diagram: (Calories, Weights)**



**Correlation coefficient (r) :**

Correlation of Weight and Calories is **0.94,** It is above 85% and it has strong correlation between the Weight and Calories.

**Building the Model-1 :**

model1 <- lm(Weight ~ Calories)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : -845.4266546 -406.0780569

Beta1 : 0.3305064 0.5098069

R-squared: 0.89

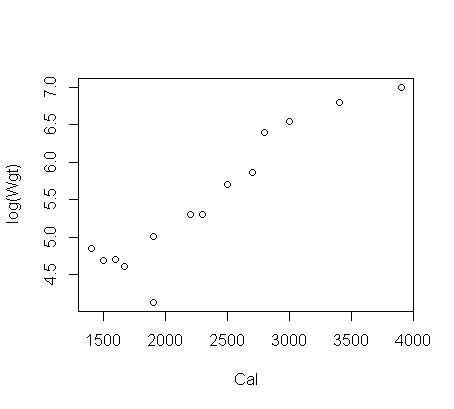
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 103.30

**Exponential transformation:**

**Scatter Diagram: (Calories, log(Weights) )**

****

**Correlation coefficient (r) :**

Correlation of log of Weight and Calories is **0.93,** It is above 85% and it has strong correlation between the log of Weight and Calories.

**Building the Model-2 :**

modl2 <- lm(log(Wgt) ~ Cal)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 2.1862091856 3.491135698

Beta1 : 0.0008673238 0.001399871

R-squared: 0.87

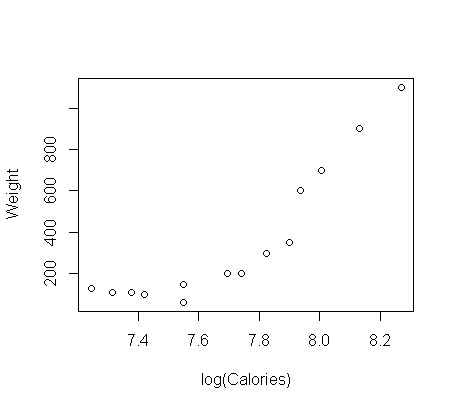
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

**RMSE = 215.64**

**Log Transformation:**

**Scatter Diagram: (log(Calories), Weights )**

****

**Correlation coefficient (r) :**

Correlation of Weight and log of Calories is **0.89,** it is above 85% and it has strong correlation between the Weight and log of Calories.

**Building the Model-3 :**

model3 <- lm(Weight ~ log(Calories))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : -9201.8063 -4709.494

Beta1 : 657.3251 1239.418

R-squared: 0.80

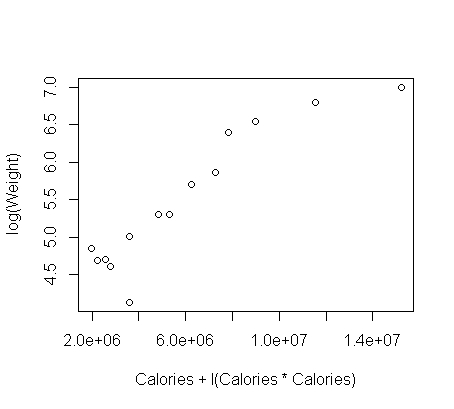
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

**RMSE = 141.00**

**Polynomial Transformation:**

**Scatter Diagram: (Calories + I(Calories \* Calories),log(Weight))**

****

**Correlation coefficient (r) :**

Correlation of Weight and Calories is **0.92,** It is above 85% and it has strong correlation between the Weight and Calories.

**Building the Model-4 :**

model4 <- lm(log(Weight) ~ Calories + I(Calories \* Calories ))del3 <- lm(Weight ~ log(Calories))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 4.920101e-01 5.165429e+00

Beta1 : -7.962832e-04 3.080576e-03

R-squared: 0.87

**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

**RMSE = 294.39**

**Conclusion:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1st Question** | **Transformation** | **Correlation Value** | **R Squared** | **RMSE** |
| 1 | Normal | 94% | 89% | 103.3 |
| 2 | Exponential Transformation | 93% | 87% | 215.6 |
| 3 | Log transformation | 89% | 80% | 141 |
| 4 | Polynomial transformation | 92% | 87% | 294.39 |

**2.** **Delivery\_time -> Predict delivery time using sorting time.**

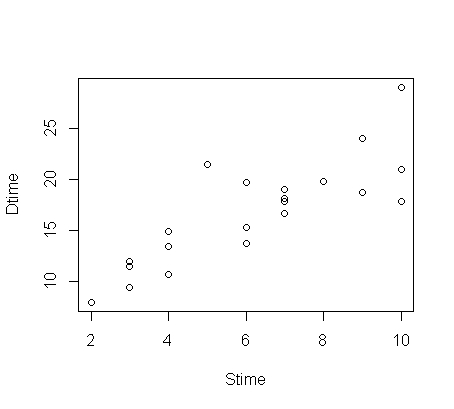
**Data Description:**

X = Sorting Time - > Input

Y = Delivery Time -> Output

Both X and Y are in Continuous Data. We are proceeding with Simple Linear Regression.

**Scatter Diagram: (Stime, Dtime)**



**Correlation coefficient (r) :**

Correlation of Delivery Time and Sorting Time is **0.825,** It is not above 85% and it has little strong correlation between the Delivery Time and Sorting Time.

**Building the Model-1 :**

model1 <- lm(Dtime ~ Stime)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 2.979134 10.186334

Beta1 : 1.108673 2.189367

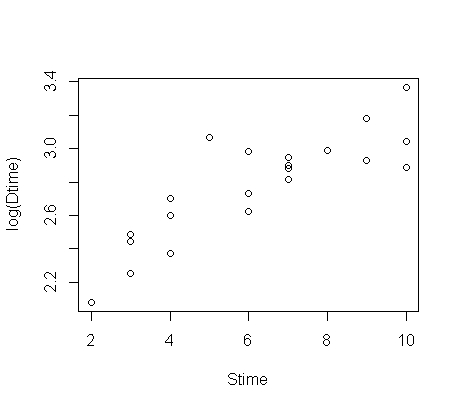
R-squared: 0.68

**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 2.79

**Exponential Transformation:**

**Scatter Diagram: (Stime,log(Dtime))** 

**Correlation coefficient (r) :**

Correlation of Log of Delivery Time and Sorting Time is **0.843,** It is not above 85% and it has little strong correlation between the Log of Delivery Time and Sorting Time.

**Building the Model-2 :**

model2 <- lm(log(Dtime) ~ Stime)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 1.90584807 2.3368956

Beta1 : 0.07323457 0.1378686

R-squared: 0.71

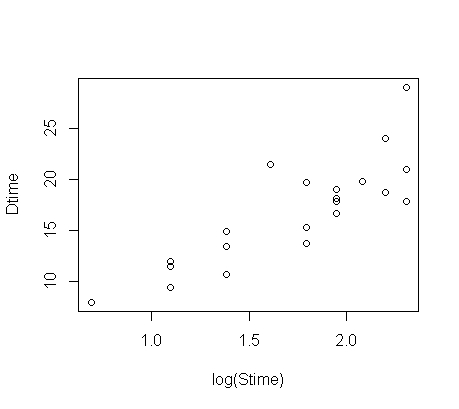
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 3.377

**Log Transformation:**

**Scatter Diagram: (Log(Stime),Dtime)**



**Correlation coefficient (r) :**

Correlation of Delivery Time and log of Sorting Time is **0.833,** It is not above 85% and it has little strong correlation between the Delivery time and log of Sorting time.

**Building the Model-3 :**

model3 <- lm(Dtime ~ log(Stime))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : -3.97778 6.297147

Beta1 : 6.16977 11.917057

R-squared: 0.69

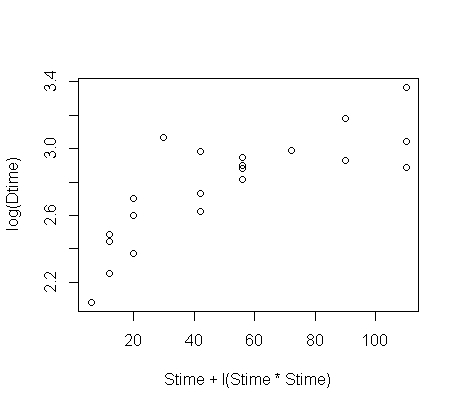
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 2.73

**Polynomial Transformation:**

**Scatter Diagram: (Stime + I(Stime \* Stime),log(Dtime))**

****

**Correlation coefficient (r) :**

Correlation of Delivery Time and Sorting Time is **0.793,** It is not above 85% and it has little strong correlation between the Delivery time and Sorting time.

**Building the Model-4 :**

model4 <- lm(log(Dtime) ~ Stime + I(Stime \* Stime ))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 1.21978583 2.179622708

Beta1 : 0.09738139 0.4344624477

R-squared: 0.76

**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 3.326

**Conclusion**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2nd Question** | **Transformation** | **Correlation Value** | **R Squared** | **RMSE** |
| 1 | Normal | 82% | 68.23% | 2.79 |
| 2 | Exponential Transformation | 84% | 71% | 3.377 |
| 3 | Log transformation | 83% | 69% | 2.73 |
| 4 | Polynomial transformation | 79.30% | 76% | 3.326 |

**3. Emp\_data -> Build a prediction model for Churn\_out\_rate.**

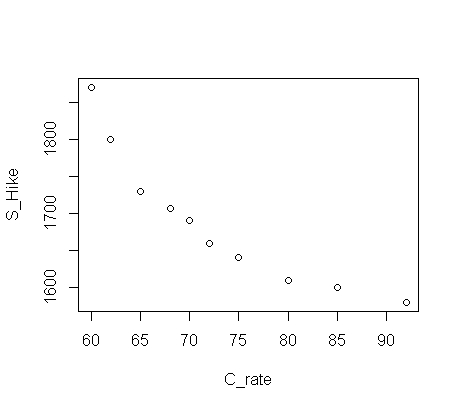
**Data Description:**

X = Sorting Time - > Churn\_out\_rate

Y = Delivery Time -> Salary\_Hike

Both X and Y are in Continuous Data. We are proceeding with Simple Linear Regression.

**Scatter Diagram: (C\_rate, S\_Hike )**



**Correlation coefficient (r) :**

Correlation of Salary Hikes and Churn\_out\_rate is **0.911,** It is above 85% and it has strong correlation between the Salary Hikes and Churn\_out\_rate.

**Building the Model-1 :**

model1 <- lm(S\_Hike ~ C\_rate)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 2064.19292 2506.537671

Beta1 : -11.19332 -5.178839

R-squared: 0.8312

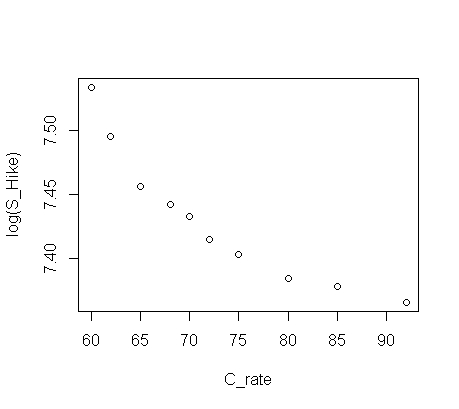
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 35.89

**Exponential Transformation:**

**Scatter Diagram: (C\_rate,log(S\_Hike ))**

****

**Correlation coefficient (r) :**

Correlation of Log of Salary Hikes and Churn\_out\_rate is **0.92,** It is above 85% and it has strong correlation between the Log of Salary Hikes and Churn\_out\_rate.

**Building the Model-2 :**

model2 <- lm(log(S\_Hike ) ~ C\_rate)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 7.659596362 7.903664318

Beta1 : -0.006477996 -0.003159446

R-squared: 0.85

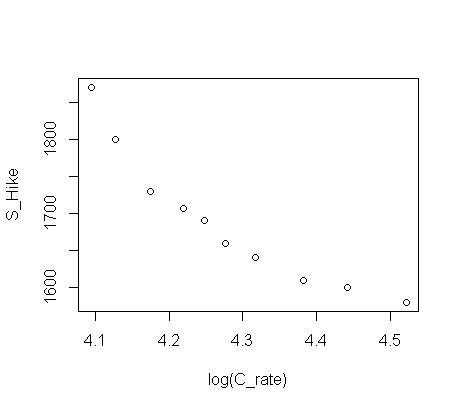
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 46.21

**Log Transformation:**

**Scatter Diagram: (Log(C\_rate),S\_Hike )**

****

**Correlation coefficient (r) :**

Correlation of Salary Hikes and log of Churn\_out\_rate is **0.934,** It is above 85% and it has strong correlation between the Salary Hikes and log of Churn\_out\_rate.

**Building the Model-3 :**

model3 <- lm(S\_Hike ~ log(C\_rate))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 3535.3898 5197.4928

Beta1 : -819.6586 -431.5379

R-squared: 0.8735

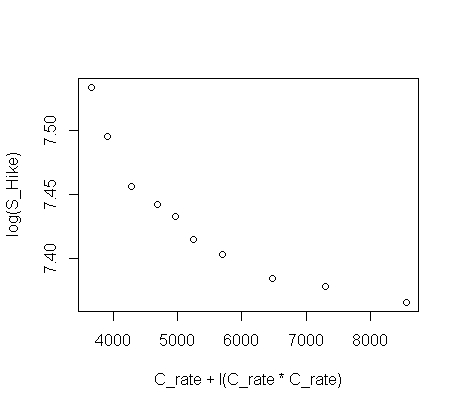
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 31.06

**Polynomial Transformation**

**Scatter Diagram: (C\_rate + I(C\_rate \* C\_rate),log(S\_Hike ))**

****

**Correlation coefficient (r) :**

Correlation of Salary Hikes and Churn\_out\_rate is **0.896,** It is above 85% and it has strong correlation between the Salary Hikes and Churn\_out\_rate.

**Building the Model-4 :**

model4 <- lm(log(S\_Hike ) ~ C\_rate + I(C\_rate \* C\_rate ))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 8.4811570475 9.2874633295

Beta1 : -0.0453390521 -0.0237348663

R-squared: 0.9786

**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 20.53

**Conclusion**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **3rd Question** | **Transformation** | **Correlation Value** | **R Squared** | **RMSE** |
| 1 | Normal | 91% | 83.12% | 35.89 |
| 2 | Exponential Transformation | 92% | 85% | 46.21 |
| 3 | Log transformation | 93% | 87% | 31.006 |
| 4 | Polynomial transformation | 89.60% | 98% | 20.53 |

4. Salary\_hike -> Build a prediction model for Salary\_hike.

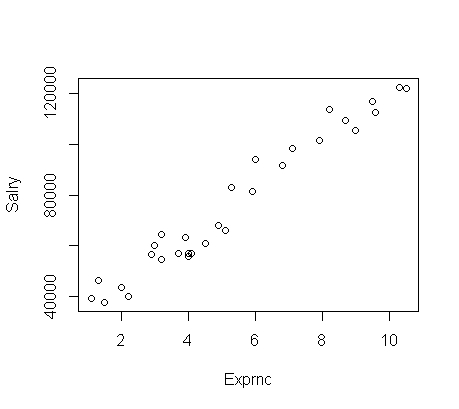
**Data Description:**

X = Sorting Time - > Yearsof experience

Y = Delivery Time -> Salary

Both X and Y are in Continuous Data. We are proceeding with Simple Linear Regression.

**Scatter Diagram: (Exprnc, Salry)**



**Correlation coefficient (r) :**

Correlation of Salary Hikes and Years of experience is **0.978,** It is above 85% and it has strong correlation between the Salary Hikes and Years of experience.

**Building the Model-1 :**

model1 <- lm(Salry ~ Exprnc)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 21136.061 30448.34

Beta1 : 8674.119 10225.81

R-squared: 0.957

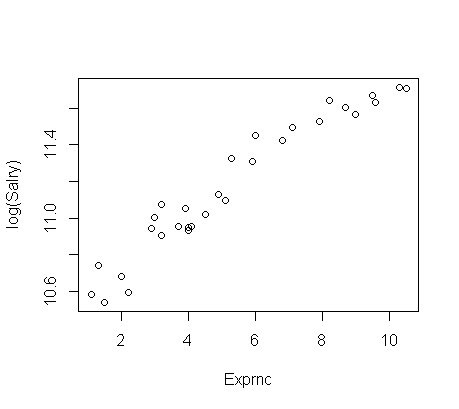
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 5592.04

**Exponential Transformation:**

**Scatter Diagram: (Exprnc,log(Salry))**

****

**Correlation coefficient (r) :**

Correlation of log of Salary Hikes and Years of experience is **0.965,** It is above 85% and it has strong correlation between the log of Salary Hikes and Years of experience.

**Building the Model-2 :**

model2 <- lm(log(Salry ) ~ Exprnc)

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 10.4286558 10.5861480

Beta1 : 0.1123316 0.1385742

R-squared: 0.932

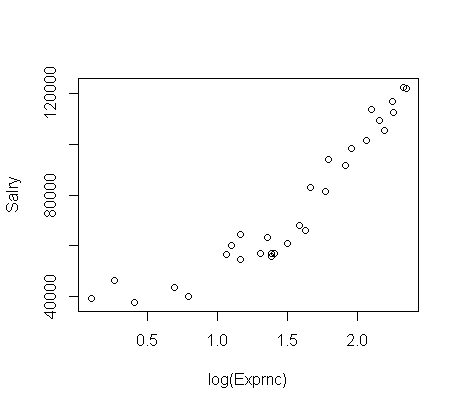
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 8177.02

**Log Transformation :**

**Scatter Diagram: (Log(Exprnc),Salry)**

****

**Correlation coefficient (r) :**

Correlation of Salary Hikes and log of Years of experience is **0.924,** It is above 85% and it has strong correlation between the Salary Hikes and log of Years of experience.

**Building the Model-3 :**

model3 <- lm(Salry ~ log(Exprnc))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 3535.3898 5197.4928

Beta1 : -819.6586 -431.5379

R-squared: 0.853

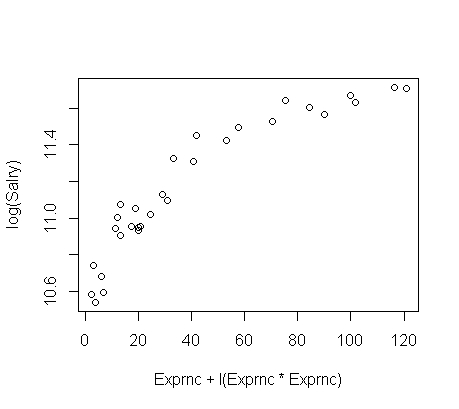
**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 10302.89

**Polynomial Transformation**

**Scatter Diagram: (Exprnc + I(Exprnc \* Exprnc),log(Salry))**

****

**Correlation coefficient (r) :**

Correlation of Salary Hikes and Years of experience is **0.921,** It is above 85% and it has strong correlation between the Salary Hikes and Years of experience.

**Building the Model-4 :**

model4 <- lm(log(Salry ) ~ Exprnc + I(Exprnc \* Exprnc ))

Lower and Upper Parameters for above defined model with 95% confidence.

Lower Upper

Beta0 : 10.19945640 10.474246649

Beta1 : 0.14775127 0.257011990

R-squared: 0.948

**Checking the Accuracy:**

Take the Square Root of Mean of Residuals (RMSE)

RMSE = 6676.884

**Conclusion**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **4th Question** | **Transformation** | **Correlation Value** | **R Squared** | **RMSE** |
| 1 | Normal | 97.8% | 95.7% | 5592.04 |
| 2 | Exponential Transformation | 96.5% | 93.2% | 8177.02 |
| 3 | Log transformation | 92.4% | 85.3% | 10302.89 |
| 4 | Polynomial transformation | 92.1% | 94.8% | 6676.88 |