**Linear Regression Assignment**

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

1.Perform EDA-1st Moment,2nd Moment,3rd Moment & 4th Moment.

2. Scatter Diagram-Direction, Strength & Linear.

3. Correlation Coefficient Value =r (-1 to +1) thumb rule if r >0.85 then it is strong. Model 1-r=0.82, Model 2-r=0.84.

4. Coefficient of determination is R2(0 to 1) thumb rule if R2>0.8 then it is strong. Model 1-R2=0.6823, Model 2-R2=0.7109(Successfully increased by using Transformation model.

I have done the analysis with R studio as follows:

R-script

# **Load delivery\_time.csv dataset**

library(readr)

delivery\_time <- read\_csv("Data Science Assignments/Linear Regression/Dataset/delivery\_time.csv")

View(delivery\_time)

attach(delivery\_time)

# **Exploratory data analysis**

summary(delivery\_time)

#**Scatter plot**

plot(delivery\_time$`Delivery Time`, delivery\_time$`Sorting Time`) # plot(X,Y)

#Correlation Coefficient (r)

cor(`Delivery Time`,`Sorting Time`) # cor(X,Y)

# **Simple Linear Regression model**

reg <- lm(delivery\_time$`Sorting Time` ~ delivery\_time$`Delivery Time`) # lm(Y ~ X)

summary(reg)

confint(reg,level=0.95)

predict(reg,interval="predict")

# **Logarithmic Transformation Model**

# x = log (Delivery Time); y = Sorting Time

plot(log(`Delivery Time`), `Sorting Time`)

cor(log(`Delivery Time`), `Sorting Time`)

reg\_log <- lm(`Sorting Time` ~ log(`Delivery Time`)) # lm(Y ~ X)

summary(reg\_log)

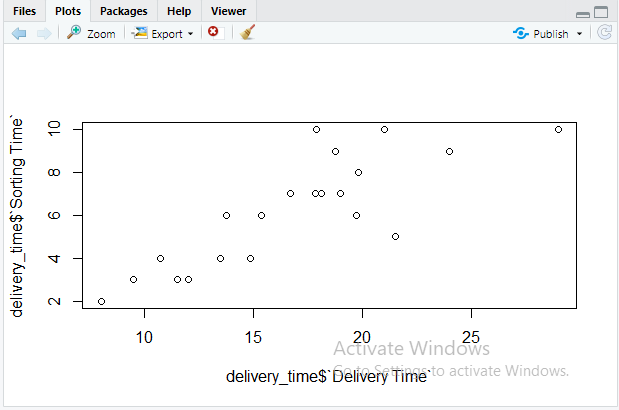
confint(reg\_log,level = 0.95)

predict(reg\_log,interval="predict")

Console Window of R

|  |
| --- |
| > # Load delivery\_time.csv dataset  > library(readr)  > delivery\_time <- read\_csv("Data Science Assignments/Linear Regression/Dataset/delivery\_time.csv")  Parsed with column specification:  cols(  `Delivery Time` = col\_double(),  `Sorting Time` = col\_double()  )  > View(delivery\_time)  > # Exploratory data analysis  > summary(delivery\_time)  Delivery Time Sorting Time  Min. : 8.00 Min. : 2.00  1st Qu.:13.50 1st Qu.: 4.00  Median :17.83 Median : 6.00  Mean :16.79 Mean : 6.19  3rd Qu.:19.75 3rd Qu.: 8.00  Max. :29.00 Max. :10.00  > #Scatter plot  > plot(delivery\_time$`Delivery Time`, delivery\_time$`Sorting Time`) # plot(X,Y)  > #Correlation Coefficient (r)  > cor(`Delivery Time`,`Sorting Time`) # cor(X,Y)  [1] 0.8259973  > # Simple Linear Regression model  > reg <- lm(delivery\_time$`Sorting Time` ~ delivery\_time$`Delivery Time`) # lm(Y ~ X)  > summary(reg)  Call:  lm(formula = delivery\_time$`Sorting Time` ~ delivery\_time$`Delivery Time`)  Residuals:  Min 1Q Median 3Q Max  -3.1388 -1.0014 -0.1045 0.5521 3.3507  Coefficients:  1.Point Estimation is ST=-0.75667+0.41374(DT)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -0.75667 1.13395 -0.667 0.513  delivery\_time$`Delivery Time` 0.41374 0.06477 6.387 3.98e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.47 on 19 degrees of freedom  Multiple R-squared: 0.6823, Adjusted R-squared: 0.6655  F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06  > confint(reg,level=0.95)  2. Lower Prediction Equation is ST=-3.1300583+0.2781691(DT)  3. Upper Prediction Equation is ST= 1.6167115+0.5493182(DT)  2.5 % 97.5 %  (Intercept) -3.1300583 1.6167115  delivery\_time$`Delivery Time` 0.2781691 0.5493182  > predict(reg,interval="predict")  fit lwr upr  1 7.931943 4.7313012 11.132584  2 4.828866 1.6480569 8.009674  3 7.414763 4.2399538 10.589573  4 9.173174 5.8756419 12.470706  5 11.241892 7.6840436 14.799740  6 5.594291 2.4388768 8.749706  7 7.104456 3.9408862 10.268025  8 3.173891 -0.1269490 6.474731  9 6.649338 3.4963890 9.802286  10 7.001020 3.8404788 10.161561  11 7.447863 4.2716645 10.624061  12 3.691071 0.4369599 6.945181  13 6.144570 2.9951731 9.293968  14 4.001378 0.7713594 7.231397  15 4.220662 1.0058373 7.435488  16 5.399832 2.2398322 8.559831  17 4.932302 1.7560698 8.108533  18 6.736224 3.5817893 9.890658  19 2.553276 -0.8140578 5.920609  20 6.620376 3.4678653 9.772886  21 8.138815 4.9253950 11.352234  Warning message:  In predict.lm(reg, interval = "predict") :  predictions on current data refer to \_future\_ responses  **Logarithmic Transformation Model**  > plot(log(`Delivery Time`), `Sorting Time`)  > cor(log(`Delivery Time`), `Sorting Time`)  [1] 0.8431773  > reg\_log <- lm(`Sorting Time` ~ log(`Delivery Time`)) # lm(Y ~ X)  > summary(reg\_log)  Call:  lm(formula = `Sorting Time` ~ log(`Delivery Time`))  Residuals:  Min 1Q Median 3Q Max  -3.1658 -0.9513 0.0949 0.5029 3.0685  Coefficients:  1.Point Estimate equation is ST=-12.4992+6.7355 log(DT)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -12.4992 2.7510 -4.543 0.000222 \*\*\*  log(`Delivery Time`) 6.7355 0.9853 6.836 1.59e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.402 on 19 degrees of freedom  Multiple R-squared: 0.7109, Adjusted R-squared: 0.6957  F-statistic: 46.73 on 1 and 19 DF, p-value: 1.593e-06  > confint(reg\_log,level = 0.95)  2.5 % 97.5 %  (Intercept) -18.257239 -6.741226  log(`Delivery Time`) 4.673307 8.797790  2.Lower Prediction equation is ST=-18.257239+4.673307 log (DT)  3.Upper Prediction Equation is ST=-6.741226+8.797790 log (DT)  > predict(reg\_log,interval="predict") [Predictive value-Actual Value=Error)  fit lwr upr  1 8.007296 4.9523462 11.062245  2 5.031310 2.0065374 8.056082  3 7.593942 4.5594842 10.628400  4 8.906703 5.7898284 12.023577  5 10.181351 6.9384633 13.424239  6 5.896328 2.8910989 8.901557  7 7.333179 4.3089936 10.357364  8 2.664452 -0.5275325 5.856437  9 6.931482 3.9190476 9.943917  10 7.243965 4.2228180 10.265112  11 7.621170 4.5855206 10.656820  12 3.497060 0.3820418 6.612078  13 6.456018 3.4510385 9.460997  14 3.951314 0.8701946 7.032433  15 4.254794 1.1930092 7.316579  16 5.686869 2.6790356 8.694703  17 5.154901 2.1343351 8.175467  18 7.010043 3.9957011 10.024384  19 1.506947 -1.8216505 4.835543  20 6.905090 3.8932536 9.916927  21 8.165786 5.1016297 11.229943  Warning message:  In predict.lm(reg\_log, interval = "predict") :  predictions on current data refer to \_future\_ responses |
|  |
| |  | | --- | |  | |

Scatter Diagram



1. Direction-Moderate Positive Correlation
2. Strength of Correlation -Moderate
3. Linear

Multiple R2=0.7109 is successfully increases after using Logarithmic Transformation model. Prediction model of predict delivery time using sorting time is good.