Assignment 4

**Simple Linear Regression**

Part 4 of 4

**To predict delivery time using sorting time**

Submitted

To

Description: logo



Submitted

By

Munawwara Shakila

(21\_Nov-2019)

**Q4) Delivery\_time -> Predict delivery time using sorting time**

**1 Business Problem: Predict delivery time**

We need to predict the delivery time of goods using the time taken to sort them.

**2 Dataset acquisition**

The dataset is delivery\_time. It has 21 records and 2 variables viz delivery time and sorting time. The below table shows first 5 records of the dataset.

|  |
| --- |
| Delivery Time Sorting Time |
| 0 21.00 10 |
| 1 13.50 4 |
| 2 19.75 6 |
| 3 24.00 9 |
| 4 29.00 10 |

**3 Exploratory Data Analysis (EDA)**

Let us understand our data. Both variables are of continuous type. The dataset is small as n\*10 times of kurtosis gives a value of 26.

|  |
| --- |
| # large dataset or not |
|  |
| from scipy.stats import kurtosis, skew |
|  |
| 10 \* kurtosis(delivery.dt) |
| Out[20]: -0.2558576894549036 |

Business moment decisions

Let us first find the business moments. Moments are popularly used to describe the characteristics of a distribution.

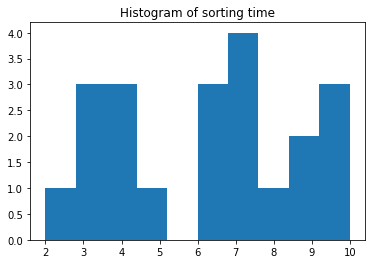
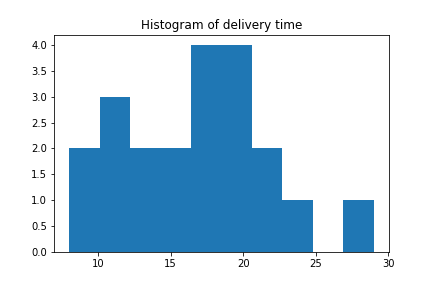
|  |  |  |
| --- | --- | --- |
|  | Deliverytime | sortingtime |
| Mean | 16.790952 | 6.190476 |
| Median | 17.83 | 6.0 |
| Mode | --- | 7 |
| Variance | 24.5282 | 6.154195 |
| Std deviation | 4.952 | 2.48076 |
| Range | 29 to 8  21 | 10 to 8  2 |
| Skewness | -0.0255 | -1.1653 |
| Kurtosis | -0.02558577 | -1.165 |

First business moment decisions tell us about central tendency. They are mean, median and mode. For delivery time, mean is 16.8 and it is 6.19 for sorting time. Second business moment decisions tell us about the spread of data. third and fourth business moments tell us about skewness and kurtosis. As seen both these are negative for both variables. They may not be normal and left skewed and also more flat.

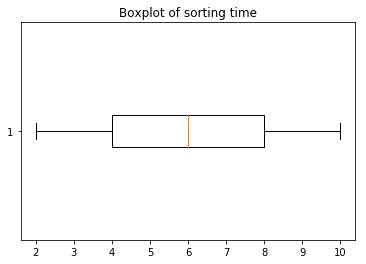
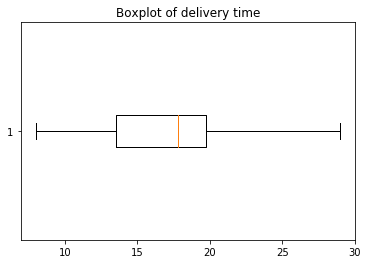
**Visualizations**

Below histograms and boxplots are plotted. From histogram and boxplot of delivery time the data looks slightly left skewed. Sorting time data looks normal. Also there seem to be no outliers in both variables.

Histogram

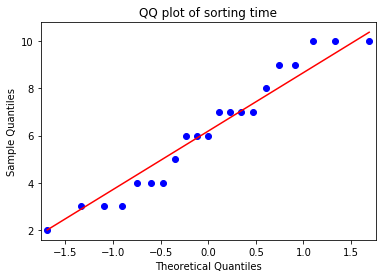


Boxplot



**Normality of data**

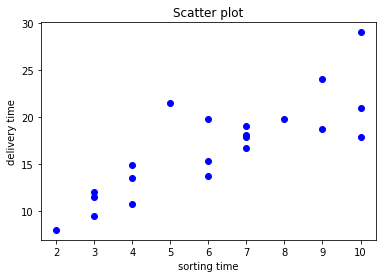
Below qq plots show that the variables are normal. This is also confirmed by the Shapiro test and normaltest. The p-values for these tests are more than 0.05.



|  |
| --- |
| normaltest(delivery.deliverytime) |
| Out[14]: NormaltestResult(statistic=0.8639105397907696, pvalue=0.6492384165379566) |
| normaltest(delivery.sortingtime) |
| Out[15]: NormaltestResult(statistic=2.7066189560535268, pvalue=0.2583837288180371) |
| shapiro(delivery.deliverytime) # (0.9781284928321838, 0.8963273763656616) |
| Out[16]: (0.9781284928321838, 0.8963273763656616) |
| shapiro(delivery.sortingtime) # (0.9367821216583252, 0.1881045252084732) |
| Out[17]: (0.9367821216583252, 0.1881045252084732) |

**Scatter plot**

The scatter plot and correlation coefficient values suggest that both variables have moderate positive linear relationship.



|  |
| --- |
| delivery.deliverytime.corr(delivery.sortingtime) # 0.82599726 |
| Out[22]: 0.8259972607955327 |

**4 Model Buildiing**

We have seen that delivery time and sorting time are normally distributed. They have moderate correlationship. Now let us build our model to predict delivery time.

**I Standard Regression model using Y ~ X**

We have run the regression model using sorting time as predictor and delivery time as target variable. Below table gives the summary of results.

|  |
| --- |
| import statsmodels.formula.api as smf |
| reg\_simple=smf.ols("deliverytime~ sortingtime", data=delivery).fit() |

|  |
| --- |
| <class 'statsmodels.iolib.summary.Summary'> |
| OLS Regression Results |
| ================================================================== |
| Dep. Variable: deliverytime R-squared: 0.682 |
| Model: OLS Adj. R-squared: 0.666 |
| Method: Least Squares F-statistic: 40.80 |
| Date: Wed, 20 Nov 2019 Prob (F-statistic): 3.98e-06 |
| Time: 10:56:59 Log-Likelihood: -51.357 |
| No. Observations: 21 AIC: 106.7 |
| Df Residuals: 19 BIC: 108.8 |
| Df Model: 1 |
| Covariance Type: nonrobust |
| ================================================================== |
| coef std err t P>|t| [0.025 0.975] |
| ------------------------------------------------------------------------------- |
| Intercept 6.5827 1.722 3.823 0.001 2.979 10.186 |
| sortingtime 1.6490 0.258 6.387 0.000 1.109 2.189 |
| ================================================================== |
| Omnibus: 3.649 Durbin-Watson: 1.248 |
| Prob(Omnibus): 0.161 Jarque-Bera (JB): 2.086 |
| Skew: 0.750 Prob(JB): 0.352 |
| Kurtosis: 3.367 Cond. No. 18.3 |
| ================================================================== |

t-stat and F-statistic: the p-value (of t-stat) for intercept and sorting time are < 0.05 suggesting that both these are significant in predicting delivery time. The p-value of F-statistics suggests that the model is significant in predicting Y.

R2 and adjusted R2: the R2 is a goodness of fit measure, its value is 0.682, indicating that sorting time is able to explain about 68% of the variance in delivery time. Adj. R2 is close to R2 value. The model is moderately good.

We now that the total variance = Sum of squares Total

SST = sum of squares regression (SSR) + Sum of squares error (SSE)

Sum of squares of Regression (SSR)

R2 = ---------------------------------------------------

Sum of squares Total (SST)

Confidence interval at 95%: the summary table also gives the confidence interval of B0 and B1. Also we can notice that the B1 coefficient value (1.649) lies between 1.109 and 2.189 (C.I.) values.

Coefficients: These are the estimates of regression equation.

The Regression equation is :

ŷ = B0 + B1\*X1 + ε

Where, ŷ is the predicted value of dependent variable, delivery time

B0 is Y intercept,

X1 is independent variable and B1 is regression coefficient of X1 (sorting time),

ε is the error term

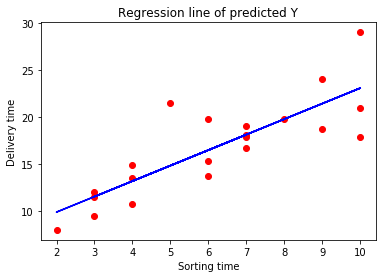
Deliverytime = B0 + B1\* (sortingtime)

Using above coefficients we can write this equation as:

Deliverytime = 6.5827 + 1.6490 \* sortingtime

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | Coeff of sortingtime | sortingtime | Predicted(y) | If X1 increases by 1 unit, increase in Y is |
| 1 | 2 | 3 | 4 | 5 |
| 6.5827 | 1.6490 | 9 | 21.4237 | 23.0727 -- 21.4237  = **1.649** |
| 6.5827 | 1.6490 | 10 | 23.0727 |  |

We have plotted the regression line; it gives the predicted values of Y. We find that the line is not a very good fit. Also the model was able to explain only 68% of variance.

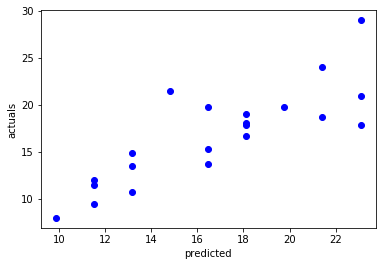
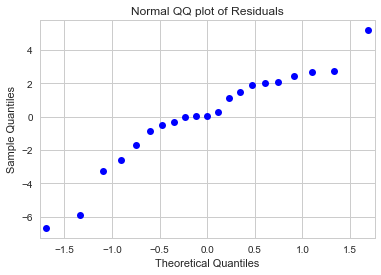


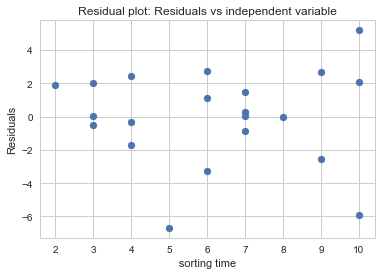
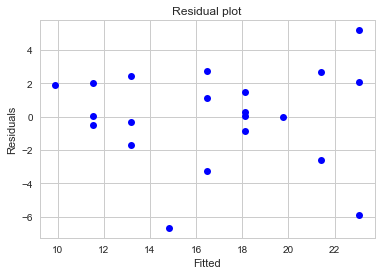
Residuals: The sum and mean of the residuals is zero. RMSE is 2.79. The below qqplot and Shapiro normality test values suggest normal distribution of residuals.

Let us examine the predicted vs actual values. Their correlation is 0.8259.There is moderate correlation between model’s prediction and actual results, and hence chance for improvement. The plot also shows moderate relation.

We have plotted 2 residual plots. In first, residuals are plotted against fitted values and in second, they are plotted against independent variable (sorting time). Both are similar. Here we just have one independent variable, so it’s easy to plot just one plot. But when we have many explanatory variables, it will be difficult plotting for each and then we can go for first residual plot.

In the residual plot, predicted (or fitted) values are plotted against residuals. The residuals are pretty symmetrically distributed. They are clustered around the lower single digits of y-axis (within 0 ± 6). In general, there are no clear patterns; there is randomness, which shows there is no autocorrelation.



From above results, we have seen that our standard regression model can be improved. Let us try transformations to increase accuracy of our model.

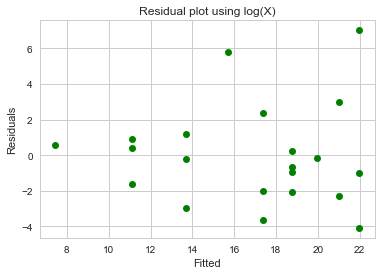
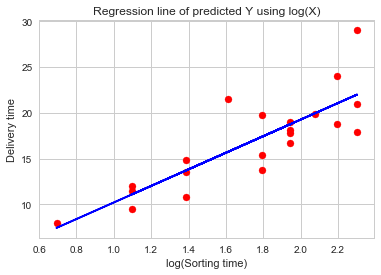
**II Model building using log(X): log(sorting-time)**

We have built until now standard regression model. To improve accuracy let us try some transformations and build the models. In the below table, a summary of different transformations used and the results of models built are displayed.

The log transformed variable is normally distributed. The correlation of log(sorting-time) and delivery-time is 0.834 which is slightly more than correlation between only sorting-time and delivery-time.

The regression coefficient of X is significant but not significant for intercept. F-statistic is significant suggesting that the overall model is good. R2 value of 0.695 is slightly higher than previous model. Regression line obtained fits the data better.

RMSE is slightly less than from that of previous model. Residuals exhibit homoscedasticity. They are normally distributed.



**Different Models built using X and various transformations of X**

The below table compares the three models: built using only X, log(X) and square-root(X).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | X | log (X) | sqrt(X) |
| 1 | Transformed\_X normality test | normal | normal | normal |
| 2 | Correlation value | 0.825997 | 0.8339 | 0.83415 |
| 3 | R2 | 0.682 | 0.695 | 0.696 |
| 4 | adjusted R2 | 0.666 | 0.679 | 0.680 |
| 5 | Intercept p-value | sig | Not sig | Not sig |
| 6 | p-value of B1 of X or transformed-X | sig | Sig | Sig |
| 7 | F-statistic | sig | Sig | Sig |
| 8 | Regression line | Good | Good | Good |
| 9 | RMSE | 2.79165 | 2.7331 | 2.7315 |
| 10 | Residual plot | Random | Random | random |
| 11 | residual normality test | normal | normal | normal |
| 12 | B0 | 6.5827 | 1.1597 | -2.5188 |
| 13 | B1 | 1.6490 | 9.0434 | 7.9366 |

The transformations tried are:

Log(X), X2, squareroot(X), 1/squareroot(X), 1/X, cuberoot(X)

When we tried 1/X, 1/square-root(X) and X2: the correlation values (delivery time vs transformed-X) were less as compared to standard regression and log(X) models. Hence did not proceed with further analysis.

When we tried square-root(X) and cube-root(X), the model values obtained were only slightly different from log(X) model. The details of square-root(X) are mentioned in above table.

**Conclusion**

As per the business problem we have predicted delivery time using sorting time. The dataset has 2 variables, delivery time and sorting time. Both are continuous. Both variables are normally distributed. We have used transformations to build a better model.

We are using simple linear regression technique to predict delivery time. The different models are built using only X, log(X) and squareroot(X).

The logarithmic model built is better than standard regression model. The

The logarithmic model has the highest R2 and the least RMSE. The correlation coefficient is also highest in log(X) model. For all models the residual plots are random. All three models are good in predicting delivery time as shown by significant F-statistic.

The log(X) model and the square-root(X) model are better than the standard regression model. When we compare log(X) and square-root(X), their difference is negligible. We can go with either of the models.

The limitations are that the dataset is very small and just one predictor variable is used to build the models.

-----------------------------------------X------------X------------X----------------------------------------