Assignment 4

**Simple Linear Regression**

Part 3 of 4

**To predict churn out rate using salary hike**

Submitted

To

Description: logo



Submitted

By

Munawwara Shakila

(21\_Nov-2019)

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

**1 Business Problem: Predicting churn out rate**

We need to predict the churn\_out\_rate based on salary hike of employees.

**2 Dataset acqusition**

We will be using the emp\_data dataset. It has 10 records and 2 variables. The variables are Salary\_hike and Churn\_out\_rate . Below table gives first 6 records of our dataset.

|  |
| --- |
| > head(emp\_data) |
| Salary\_hike Churn\_out\_rate |
| 1 1580 92 |
| 2 1600 85 |
| 3 1610 80 |
| 4 1640 75 |
| 5 1660 72 |

|  |
| --- |
| > # Details of dataset |
| > dim(emp\_data) |
| [1] 10 2 |
| > # 10 obsvn 2 variables |
| > names(emp\_data) |
| [1] "Salary\_hike" "Churn\_out\_rate" |
| > # varaibles are "Salary\_hike" "Churn\_out\_rate" |
| > class(emp\_data) |

**3 Exploratory Data Analysis**

Let us understand our data. Both are continuous variables. Generally if n≥30 we call it as a large dataset. But better measure is if n > 10 times of (kurtosis). Here this value comes to around 25. But we have seen that the dataset has only10 records. So our dataset is small.

|  |
| --- |
| > str(emp\_data) |
| 'data.frame': |
| $ Salary\_hike : int 1580 1600 1610 1640 1660 1690 1706 1730 1800 1870 |
| $ Churn\_out\_rate: int 92 85 80 75 72 70 68 65 62 60 |

**Moments**

To know more about our data we need to find the 4 business moments. Moments are popularly used to describe the characteristics of a distribution. They summarize many of the descriptive statistical measures.

|  |
| --- |
| > # I BM |
| > mean(Salary\_hike) |
| [1] 1688.6 |
| > median(Salary\_hike) |
| [1] 1675 |
| > library(modeest) |
| > mlv(Salary\_hike,method='mfv') |
| # no mode |
| > mean(Churn\_out\_rate) |
| [1] 72.9 |
| > median(Churn\_out\_rate) |
| [1] 71 |
| > mlv(Churn\_out\_rate,method=’mfv’) |
| # no mode |

|  |  |  |
| --- | --- | --- |
|  | Salary\_hike | churn\_out\_rate |
| Mean | 1688.6 | 72.9 |
| Median | 1675 | 71 |
| Mode | ---- | --- |
| Variance | 8481.822 | 105.2111 |
| Std deviation | 92.09681 | 10.25725 |
| Range | 1580 to 1870  290 | 60 to 92  32 |
| Skewness | 0.6180303 | 0.466011 |
| Kurtosis | 2.548327 | 2.268898 |

* First business moment decisions: mean, median and mode

These are measures of central tendency: give measures around which most of the data points lie. Average salary is 1689. Median salary is 1675. There is no repeated value in the data i.e., it has no mode.

Average churn is around 72.9 and the median is around 71. No value is repeated in churn i.e., it has no mode.Since mean and median are similar, it may not have any outliers.

* Second business moment decisions: variance, standard deviation and range

These are measures of dispersion and show the spread of the data. For salary the variance is 8881 and standard deviation is 92.The salary ranges from 1580 to 1870. The range value is 190.

The variance for churn is 105 and standard deviation is 10.26. churn ranges from 60 to 92. The range value is 32.

* Third business moment decision: skewness

Skewness measures the symmetry of the data. Variation tells the amount of variation and skewness tells the direction of variation. The value of skewness for salary is 0.6180303 and for churn it is 0.466011. Since, for both variables, the values are closer to zero, their distributions could be normal. To know more accurately about normality of the distribution, we will be using some more measures in the report.

* Fourth business moment decision: kurtosis

Kurtosis measures the flatness or peakedness of the data. It measures the heaviness of the distribution tails. The value of kurtosis for salary is 2.548327 and for churn it is 2.268898. Since both values are slightly positive, the distributions could be more peaked than normal. More measures will follow to confirm these results.

We can get basic information of each variable from summary statistics. The summary() function in r gives the below results. It summarises some of the descriptive statistics values for all the variables. Below table gives the following values: minimum, first quartile (Q1), mean, median (Q2), third quartile (Q3), maximum.

We can say this about the 3 quartiles of salary variable. Below 1675 (median value), 50% of the data lie. Below 1618 (Q1), 25% of the data lie and below1724 (Q3), 75% of the data points lie.

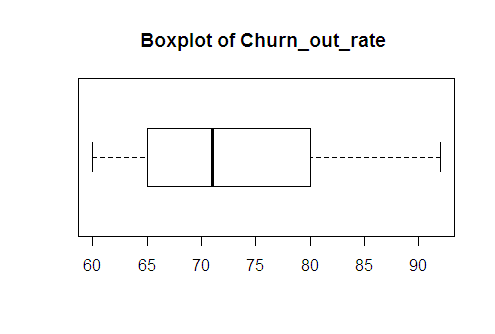
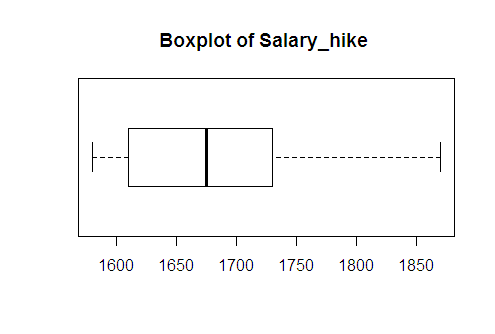
|  |
| --- |
| > summary(emp\_data) |
| Salary\_hike Churn\_out\_rate |
| Min. :1580 Min. :60.00 |
| 1st Qu.:1618 1st Qu.:65.75 |
| Median :1675 Median :71.00 |
| Mean :1689 Mean :72.90 |
| 3rd Qu.:1724 3rd Qu.:78.75 |
| Max. :1870 Max. :92.00 |

**Visualizations**

Visualizations help us to understand the data in an easier way.

Boxplot

The median for salary is slightly towards the right of the box. The right whisker is very long compared to left whisker. The value of skewness as seen before was 0.62. The data is positively skewed. In churn, the median is towards left of the plot, and skewness is 0.46. churn is positively skewed.

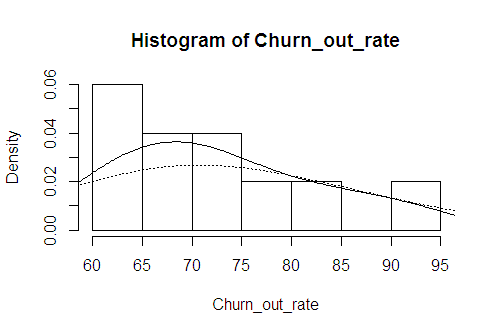
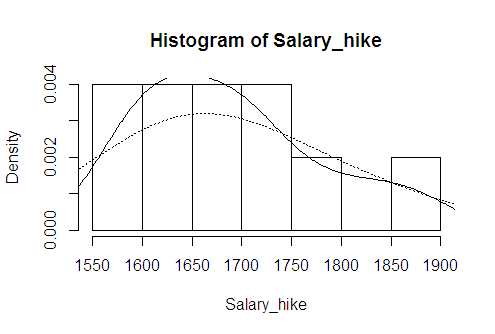


Finding outliers:

The below code can be used to find outliers. Since value is 0, indicates both variables do not have outliers.

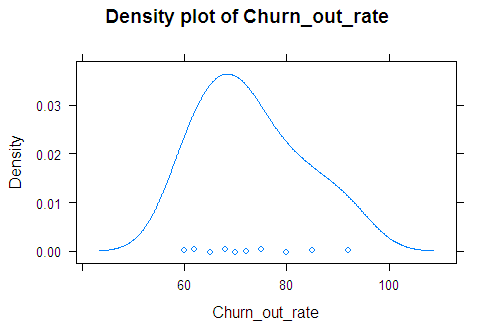
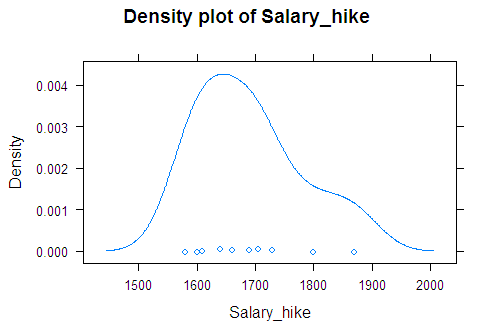
|  |
| --- |
| > # finding outliers |
| > boxplot\_salary$out |
| numeric(0) |
| > boxplot\_churn$out |
| numeric(0) |

Histogram: the histogram and the density lines show that the data of both variables is slightly right skewed. Also, for both variables the distribution is flattened.



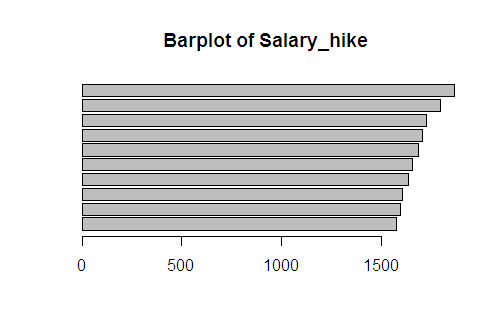
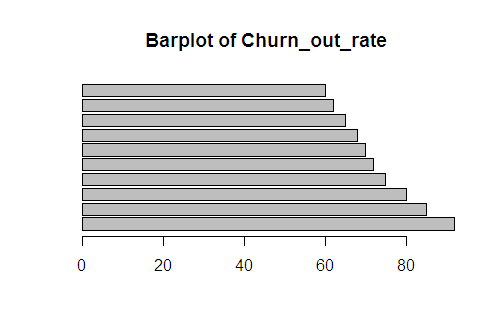
Densityplot

It is a smoothed version of histogram. It shows the probability density function of the variable using kernel density estimate. Both variables seemed to be slightly right skewed.



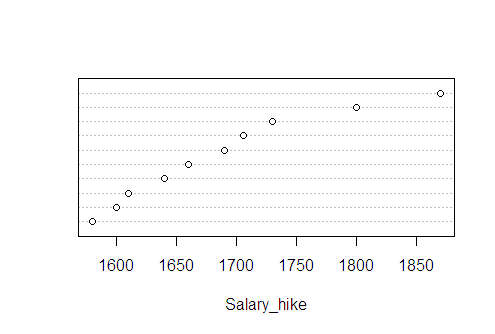
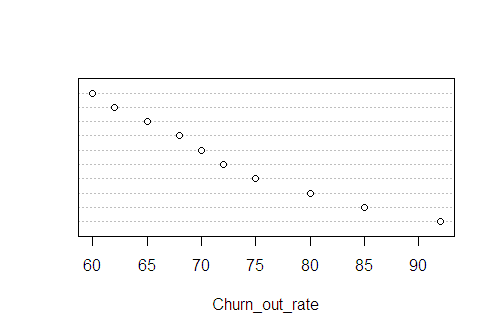
**Barplot**

Barplot can be used for small datasets. The length of the bar gives value of each data point.

**Dot chart**

Dot chart can be used for small datasets. It shows the position of each data point.

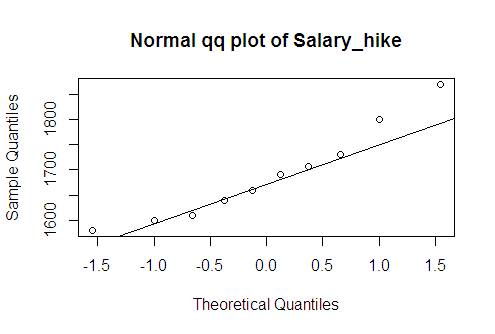
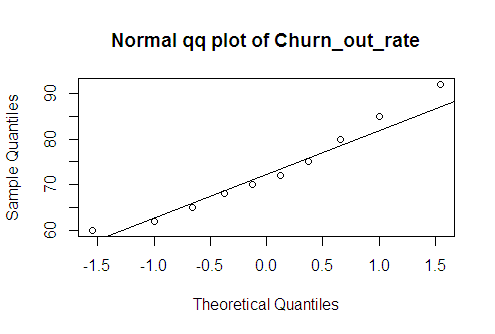
 

**Finding Normality of data**

Most of the analytical models developed work with underlying assumption that the variables are normally distributed. There are different ways to know whether variables are normally distributed or not. We have already seen boxplot and histogram.

Normal Quantile-Quantile plot

QQ plot compares two distributions. It plots the theoretical quantiles of normal distribution against the variable quantiles. QQ plots of both salary and churn variables show normal distribution.The Shapiro-Wilk and Anderson\_Darling normality tests also give a p-value greater than 0.05 for both variables.

Finding Missing values

There are no missing values in the entire dataset.

|  |
| --- |
| > # finding missing values |
| > emp\_data[!complete.cases(emp\_data),] |
| [1] Salary\_hike Churn\_out\_rate |

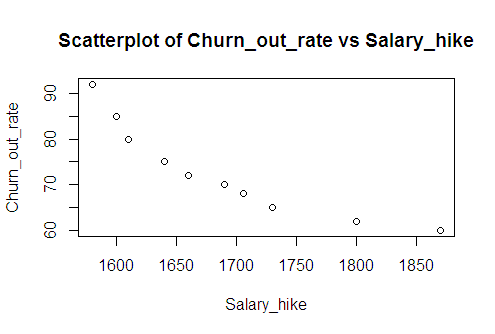
**4 Model Building and Interpretation**

We have pre-processed our dataset. The dataset has no missing values and there are no outliers. The target variable here is churn\_out\_rate and predictor variable is salary\_hike. Both are normally distributed. The dataset is small with just 10 records.

The business problem is to predict churn\_out\_rate using salary\_hike. We need to build a prediction model. As seen, both are continuous variables. There is only one predictor varaible. So we will be using simple linear regression technique.

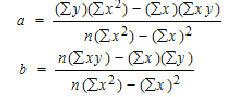
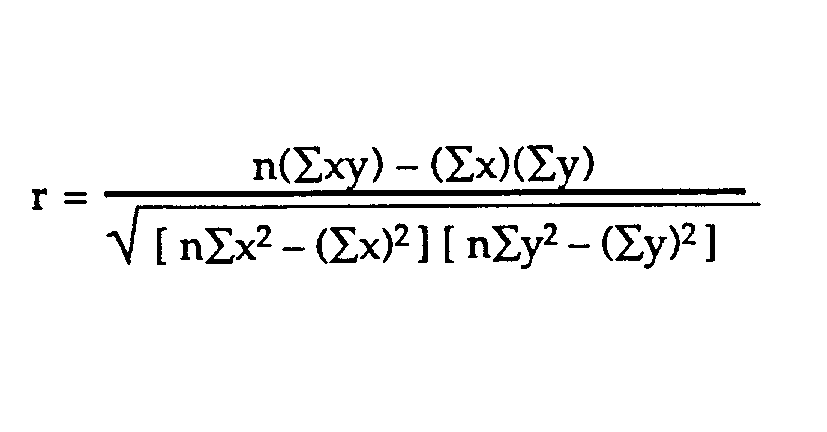
Linear relationship

First, linear regression needs the predictor to be linearly related to the target variable. We can check this visually using scatter plot. Also, we can find magnitude using pearson correlation coefficient. Looking at the scatter plot we can say that the variables have a negative correlation. The correlation coefficient is --0.9117216. Both variables have a strong negative linear relationship. This means that when salary\_hike increases, the churn\_out\_rate decreases in the company. This is also seen in real time scenario.



**Manual calculation of Pearson’s Correlation Coefficient and Regression equation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Salary\_hike X** | **Churn\_out\_rate Y** | **XY** | **X2** | **Y2** |
|  | 1 | 1580 | 92 | 145360 | 2496400 | 8464 |
|  | 2 | 1600 | 85 | 136000 | 2560000 | 7225 |
|  | 3 | 1610 | 80 | 128800 | 2592100 | 6400 |
|  | 4 | 1640 | 75 | 123000 | 2689600 | 5625 |
|  | 5 | 1660 | 72 | 119520 | 2755600 | 5184 |
|  | 6 | 1690 | 70 | 118300 | 2856100 | 4900 |
|  | 7 | 1706 | 68 | 116008 | 2910436 | 4624 |
|  | 8 | 1730 | 65 | 112450 | 2992900 | 4225 |
|  | 9 | 1800 | 62 | 111600 | 3240000 | 3844 |
|  | 10 | 1870 | 60 | 112200 | 3496900 | 3600 |
| **∑** | **10** | **16886** | **729** | **1223238** | **28590036** | **54091** |
|  |  |  |  |  |  |  |
|  | **r** | **-0.91172** |  |  |  |  |
|  | **R2 = r^2** | **0.831236** |  |  |  |  |
|  | **B0 = a** | **244.3649** |  |  |  |  |
|  | **B1 = b** | **-0.10154** |  |  |  |  |



**I Standard Linear Regression Model using churn\_out\_rate and Salary\_hike**

Let us now build the Simple linear regression model in r. Below table gives summary statistics of regression.

|  |
| --- |
| > # standard regression model |
| > reg\_simple <- lm(Churn\_out\_rate ~ Salary\_hike) |
| > summary(reg\_simple) |
|  |
| Call: |
| lm(formula = Churn\_out\_rate ~ Salary\_hike) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -3.804 -3.059 -1.819 2.430 8.072 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\* |
| Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 4.469 on 8 degrees of freedom |
| Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101 |
| F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386 |

Formula call: Shows the syntax used to build the linear model

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,

B0 is Y intercept,

X1 is independent variable and B1 is regression coefficient of X1 variable,

ε is the error term

churn\_out\_rate = B0 + B1\* (Salary\_hike)

Using above coefficients we can write this equation as:

churn\_out\_rate = 244.36491 – 0.10154 \* Salary\_hike

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | Coeff of Salary\_hike | Salary\_hike | Predicted(y) | If X1 increases by 100 units, increase in Y is |
| 1 | 2 | 3 | 4 | 5 |
| 244.36491 | 0.10154 | 1600 | 81.90 | 71.74 -- 81.90  = **- 10.154** |
| 244.36491 | 0.10154 | 1700 | 71.74 |  |
| 244.36491 | 0.10154 | 1800 | 61.59 |  |

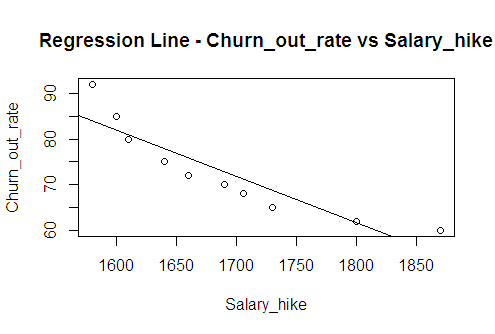
We can interpret the equation in the following manner. If salary\_hike is increased by 1 unit, then predicted churn\_out\_rate decreases by 0.10154 units ( or 0.10 % since the units for churn is %) or if there is 100 units increase in salary, the churn\_out\_rate decreases by 10.154 units (or 10.15% since the units for churn is %) as shown in 5th column of above table.

If the management decides to increase salary to 1700 units then they can see a decrease in the churn\_out\_rate by about 10%.

Pvalue of t-stat and F-statistic: For the t-value, the pvalue is = 0.000 for both intercept and salary\_hike. This indicates that both are highly significant in estimating salary. F-statistic suggests if the model is good or not. Its p-value is <0.05 suggesting that the model is significant in predicting churn\_out\_rate.

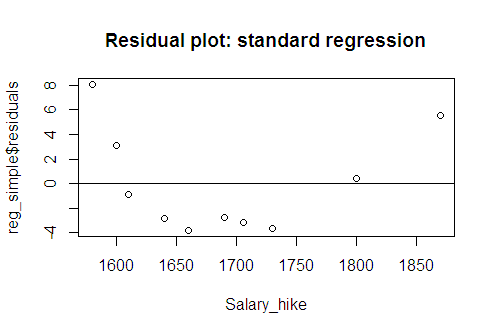
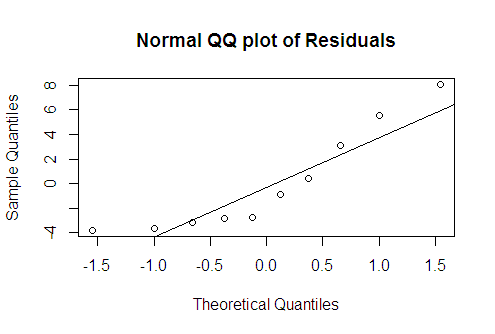
R2 and Adjusted R2: R2 is the coefficient of determination. It indicates that about 83.12% of the variance of churn\_out\_rate is explained by the model. R2  increases with increase in number of predictors even if they are not significant. Adj. R2 penalises if more variables are used in the model. Here we have just one predictor and adj. R2 value (81.01%) is close to R2 value, indicating that the predictor is significant in predicting the target.

Plotting Regression Line: now let us plot the regression line on the scatter plot of actual churn\_out\_rate vs salary\_hike. The straight line gives the predicted Y values. The line passes close to almost all points. The model looks good. Above R2 value also suggests that the model is good. Next let us look at the residuals to confirm if it is a good model.



Residuals: The sum and mean of residuals is zero. RMSE is 3.9975. We can compare this value with different models and select the model with least RMSE. The Residual plot shows some U-shape pattern. This suggests that our model is not good. We can go for transformations. From below QQ plot of residuals we can say that they are almost normally distributed. Also, the p-value of Shapiro-Wilk and Anderson-Darling tests suggests normal distribution of residuals.

|  |
| --- |
| > # Residuals |
| > sum(reg\_simple$residuals) # is 0 |
| [1] -4.440892e-16 |
| > mean(reg\_simple$residuals) # is 0 |
| [1] -4.449566e-17 |
| > sqrt(mean(reg\_simple$residuals^2)) |
| [1] 3.997528 |
|  |
| Shapiro-Wilk normality test |
| data: reg\_simple$residuals  W = 0.85023, p-value = 0.05845 |

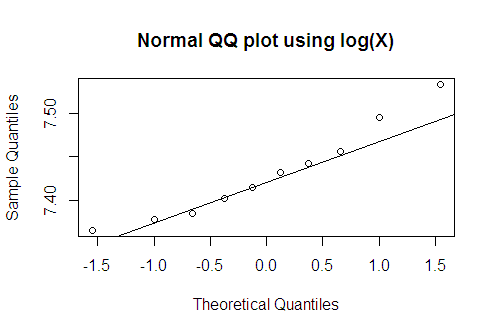
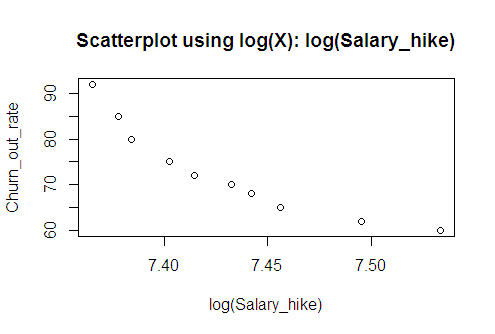
 

But we have seen that the residuals are not random. Let us try different transformations and build the models.

**II Logarithmic Model using log(X): log(salary\_hike)**

Let us now try log transformation of salary\_hike. The QQ plot and the normality tests suggest that log(salary\_hike) is normally distributed. The value of Shapiro\_Wilk normality test is 0.5992, also suggesting normal data.

The scatterplot of log(salary\_hike) and churn\_out\_rate shows strong negative linear relation.This is confirmed by correlation coefficient which is -0.9212077. This implies that as salary\_hike increases, employee churn\_out\_rate decreases.

Regression Model using log(X)

Let us now build regression model. The regression coefficients are significant as revealed by t-statistic. F-statistic indicates that the model is overall significant. R2 value is 0.8486, slightly more than standard regression model. Adj R2 is 0.8297 and close to R2.

|  |
| --- |
| > # regression - log(X) |
| > reg\_log <- lm(Churn\_out\_rate ~ log(Salary\_hike)) |
| > summary(reg\_log) |
|  |
| Call: |
| lm(formula = Churn\_out\_rate ~ log(Salary\_hike)) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -3.678 -2.851 -1.794 2.275 7.624 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 1381.5 195.4 7.070 0.000105 \*\*\* |
| log(Salary\_hike) -176.1 26.3 -6.697 0.000153 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 4.233 on 8 degrees of freedom |
| Multiple R-squared: 0.8486, Adjusted R-squared: 0.8297 |
| F-statistic: 44.85 on 1 and 8 DF, p-value: 0.0001532 |

We can write the regression equation as follows:

churn\_out\_rate = B0 + B1\* log(salary\_hike)

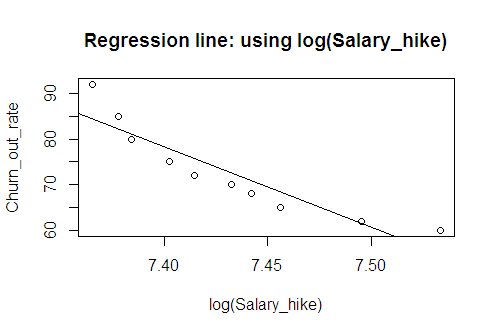
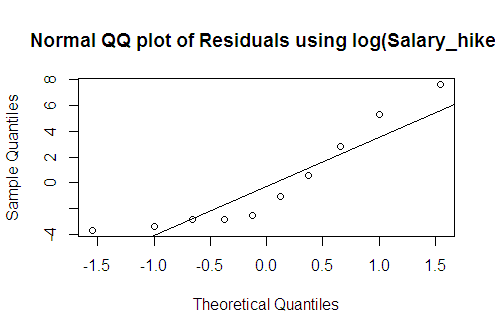
churn\_out\_rate = 1381.5 + (-176.1) \* log(salary\_hike)

The coefficients of intercept and log(salary\_hike) are 1381.5 and -176.1 respectively. Generally log(X) can be interpreted as a one percent increase in the independent variable decreases (or increases) the dependent variable by (coefficient\*0.01) units. So if we increase salary\_hike by one percent, churn\_out\_rate decreases by (176.1\*0.01) 1.761 units.

When we increase salary from 1580 to 1595.8 (one percent increase from 1580), then churn\_out decreases by 1.76 units from 84.38 to 82.62 (82.62 – 84.38 = 1.76).

|  |  |  |
| --- | --- | --- |
| salary\_hike | pred\_Y | if X increases by 1% then Y decreases by |
| 1580 | 84.38 |  |
| 1595.8 | 82.62 | -1.76 |

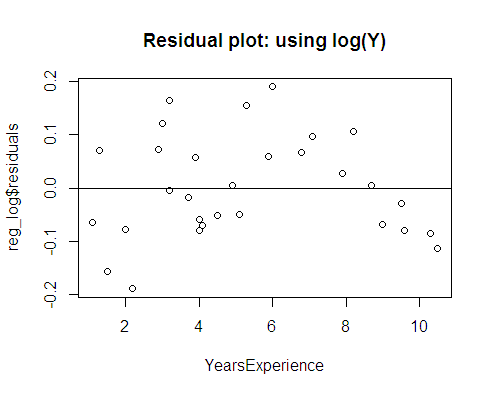
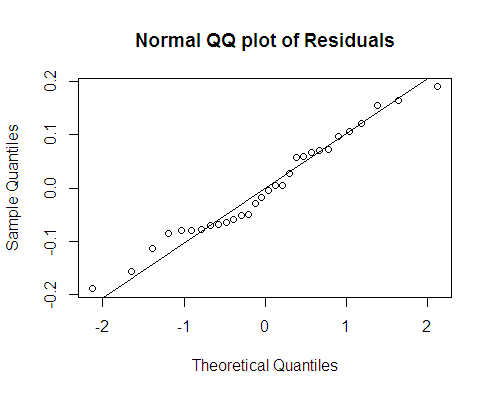
Now let us plot the regression line. The line passes close to almost all points, suggesting that this is a good model.

Residuals:

Residuals are the errors. The sum and mean of residuals is zero. RMSE is 3.786004, less than standard regression residuals. The residual plot is similar to the one obtained using standard regression model. The residuals are somewhat randomi.e., following homoscedasticity (constant variance) across the values of independent variable. All the residuals lie between 0 ± 2, indicating that there are no outliers.

The residuals are normally distributed as shown in the below QQ plot. Also the Shapiro\_Wilk normality test gives a p-value of 0.05982, which is > 0.05, indicating that residuals are normally distributed.

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

We have built until now standard regression model and logarithmic model. As we know we can make many more transformations and build the models. In the below table, a summary of different transformations used and the results of models built are displayed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | X | log (X) | X^2 | sqrt(X) | 1/sqrt(X) | 1/X | cuberoot(X) |
| 1 | Transformed\_X normality test | 0.5018 | 0.5992 | 0.4049 | 0.5509 | 0.6459 | 0.6901 | 0.5671 |
| 2 | Correlation value | -0.91172 | -0.9212077 | -0.90172 | -0.91653 | 0.925747 | -0.99868 | -0.91811 |
| 3 | R2 | 0.8312 | 0.8486 | 0.8312 | 0.84 | na | na | 0.8429 |
| 4 | adjusted R2 | 0.8101 | 0.8297 | 0.8101 | 0.82 | na | na | 0.8233 |
| 5 | Intercept p-value | sig | sig | sig | Sig | Na | na | Sig |
| 6 | p-value of B1 of X or transformed-X | sig | sig | sig | Sig | Na | Na | Sig |
| 7 | F-statistic | sig | sig | sig | Sig | na | Na | Sig |
| 8 | Regr ession line | Good | Good | Good | good | na | na | good |
| 9 | RMSE | 3.99753 | 3.786004 | 3.997528 | 3.891995 | Na | Na | 3.85671 |
| 10 | Residual plot | Slight U pattern | same | Same | Same | na | na | same |
| 11 | residual normality test | 0.05845 | 0.05982 | 0.05845 | 0.05949 | na | na | 0.05968 |
| 12 | B0 |  | 1381.5 | 244.365 | 420.468 | 72.9 | 72.9 | 596.576 |
| 13 | B1 |  | -176.1 | -0.10154 | -8.461 | na | na | -43.989 |

The transformations tried are:

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

When we tried 1/X (inverse of X) and inverse of squareroot of X (1/sqrt(X)), the regression equation had only intercept value suggesting that the transformed variable is not significantly predicting Y. Hence their other results are not discussed further.

The details of log(X) transformations have been explained previously. The X and transformed-X variables were normally distributed as given by Shapiro-Wilk normality test values in row 1 of above table. The correlation values were negative as shown in row2: the X and transformed-X variables have a strong negative correlation with Y variable.

R2 and adj R2 values (row3, row4) are more than 0.80 for all, suggesting that X or transformed-X is able to explain more than 80% of variance in Y. Highest values are for log(X) transformation.

The coefficients of regression equation are highly significant (row5 and row6). F-statistic (row7) is significant suggesting that these models are good.

Regression line obtained is good (row8).

RMSE is the least for log(X) when compared to other models (row9).

The residual plot is almost same for all, it has a slight U-shaped pattern (row10).

The residual normality tests in row11 suggest that the residuals are normally distributed.

The coefficients are given in row12 (intercept) and row13 (B1 coefficient).

**Conclusion**

The dataset has 2 variables and both are continuous. As per the business problem we have predicted churn out rate based on hike in salary. Both variables are normally distributed. we have used transformations to build a better model.

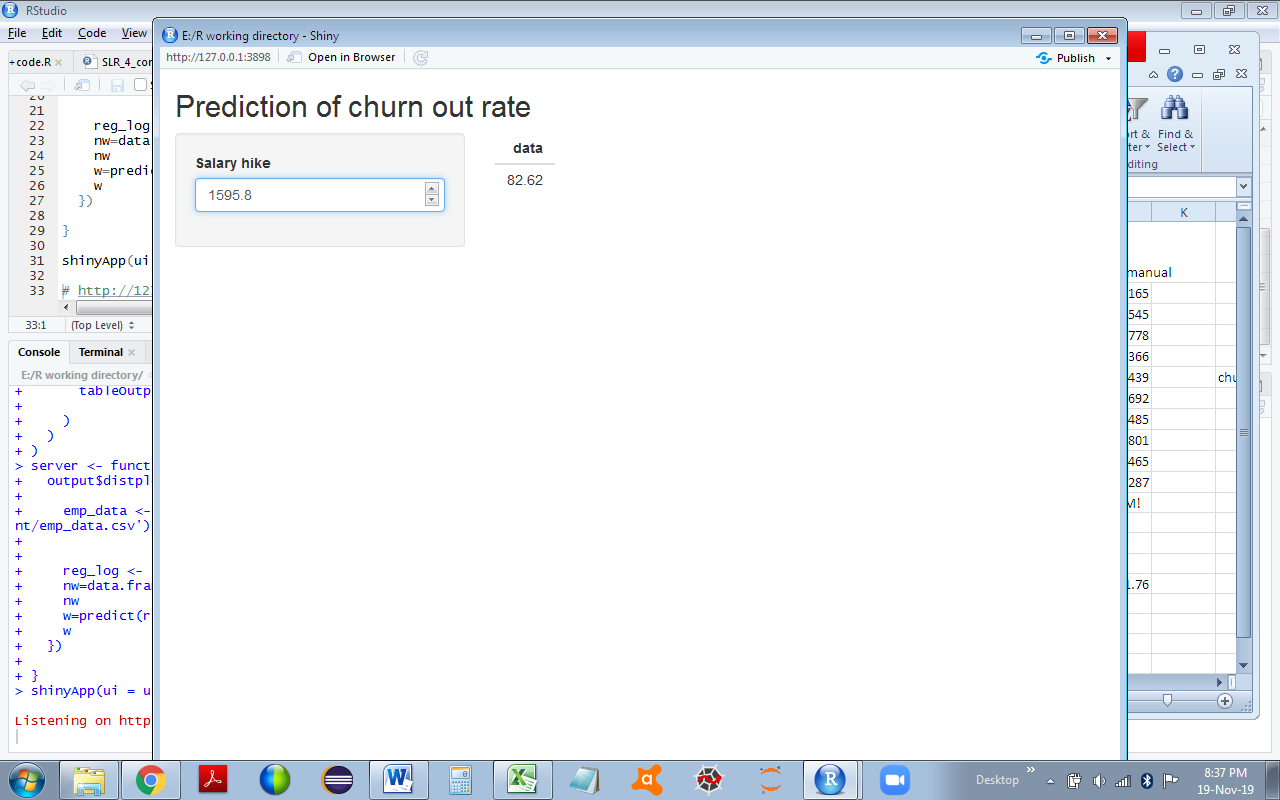
We are using simple linear regression technique to predict churn out rate. The different models are built using only X, log(X), X2, squareroot(X), inverse of squareroot(X), inverse of X, cuberoot(X).

The logarithmic model has the highest R2 and the least RMSE. The correlation coefficient is also highest in log(X) model. We can conclude that the logarithmic model is a better one.

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

**Deployment**

The deployment of this logarithmic regression model is shown below using R-shiny package



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