Assignment 4

**Simple Linear Regression**

Part 2 of 4

**To predict salary hike using work experience**

Submitted

To

logo



Submitted

By

Munawwara Shakila

(21\_Nov-2019)

**Q2. Salary\_hike -> Build a prediction model for Salary\_hike**

**1 Business Problem: Predicting hike in salary**

We need to predict the salary based on work experience of the employee.

**2 Dataset acqusition**

We will be using the Salary\_Data dataset. It has 30 records and 2 variables. The variables are YearsExperience and Salary. Below table gives first 6 records of our dataset.

|  |
| --- |
| YearsExperience Salary |
| 1 1.1 39343 |
| 2 1.3 46205 |
| 3 1.5 37731 |
| 4 2.0 43525 |
| 5 2.2 39891 |
| 6 2.9 56642 |

|  |
| --- |
| > # Details of dataset |
| > dim(salary\_hike) |
| [1] 30 2 |
| > names(salary\_hike) |
| [1] "YearsExperience" "Salary" |
| > # varaibles are "YearsExperience" "Salary" |
| > class(salary\_hike) |
| [1] "data.frame" |

**3 Exploratory Data Analysis**

Let us understand our data. Both are continuous variables. Also we have seen that the dataset has 30 records. 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 20. So we can say that our dataset is large enough.

|  |
| --- |
| > str(salary\_hike) |
| 'data.frame': 30 obs. of 2 variables: |
| $ YearsExperience: num 1.1 1.3 1.5 2 2.2 2.9 3 3.2 3.2 3.7 ... |
| $ Salary : num 39343 46205 37731 43525 39891 ... |

**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(YearsExperience) |
| [1] 5.313333 |
| > median(YearsExperience) |
| [1] 4.7 |
| > library(modeest) |
| > mlv(YearsExperience,method='mfv') |
| [1] 3.2 4.0 |
| > mean(Salary) |
| [1] 76003 |
| > median(Salary) |
| [1] 65237 |
| > mlv(Salary,method='mfv') |

|  |  |  |
| --- | --- | --- |
|  | YearsExperience | Salary |
| Mean | 5.313333 | 76003 |
| Median | 4.7 | 65237 |
| Mode | 3.2 , 4.0 | --- |
| Variance | 8.053609 | 751550960 |
| Std deviation | 2.837888 | 27414.43 |
| Range | 1.1 to 10.5  9.4 | 37731 to 122391  84660 |
| Skewness | 0.3603123 | 0.3361619 |
| Kurtosis | 1.955248 | 1.717087 |

* 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 experience is 5.313 years. Median experience is 4.7 years. The data is bimodal with values of 3.2 and 4.0 and each is repeated twice.

Average salary is around 76,000 and the median is around 65,000. No value is repeated in salary i.e., it has no mode.Since mean and median differ, it indicates that the data may have outliers.

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

These are measures of dispersion.and show the spread of the data. For experience the vairance is 8.05 and standard deviation is 2.84.The years of experience ranges from 1.10 to 10.5 years. The range value is 9.4 years.

The variance for salary is 751550960 and standard deviation is 27414. Salary ranges from around 37,000 to 122,000. The range value is 84660.

* 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 experience is 0.3603123 and for salary it is 0. 3361619. 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 experience is 1.955248 and for salary it is 1.717087. 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 experience variable. Below 4.70 (median value), 50% of the data lie. Below 3.2 (Q1), 25% of the data lie and below7.7 (Q3), 75% of the data points lie.

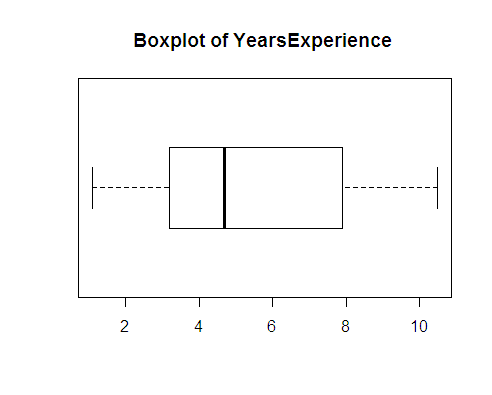
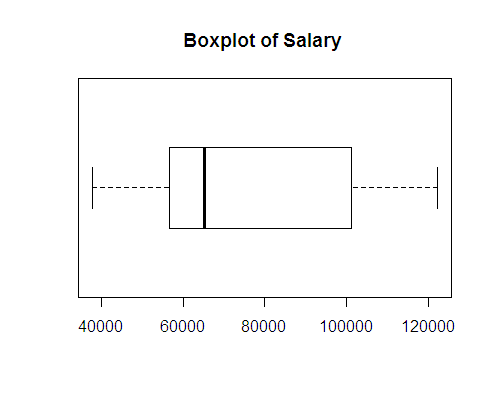
|  |
| --- |
| > summary(salary\_hike) |
| YearsExperience Salary |
| Min. : 1.100 Min. : 37731 |
| 1st Qu.: 3.200 1st Qu.: 56721 |
| Median : 4.700 Median : 65237 |
| Mean : 5.313 Mean : 76003 |
| 3rd Qu.: 7.700 3rd Qu.:100545 |
| Max. :10.500 Max. :122391 |

**Visualizations**

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

Boxplot

The median for both variables is towards the left indicating that they are not normal and the data are left skewed.

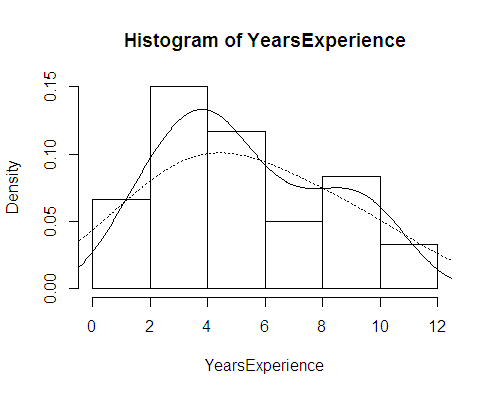
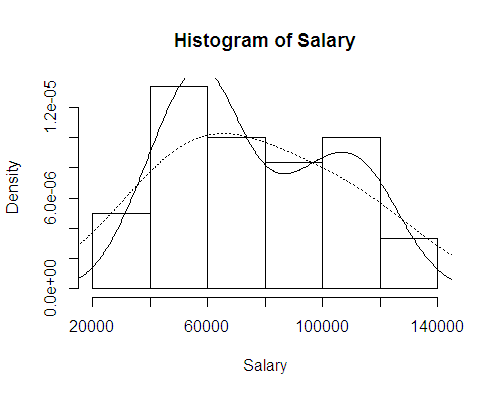
 

Finding outliers:

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

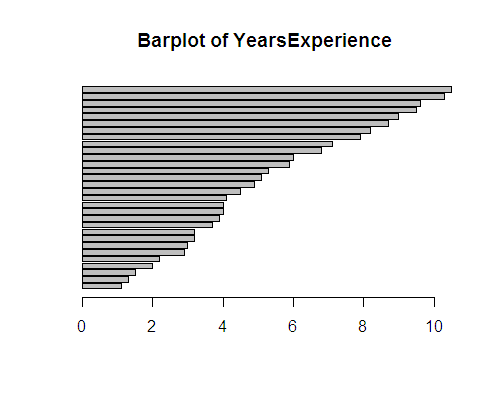
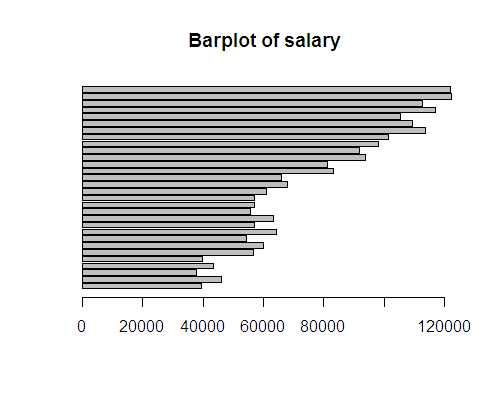
|  |
| --- |
| # finding outliers |
| > boxplot\_exp$out |
| numeric(0) |
| > boxplot\_sal$out |
| numeric(0) |

Histogram: the histogram and the density lines show that the data of both variables is slightly right skewed.

**Barplot**

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

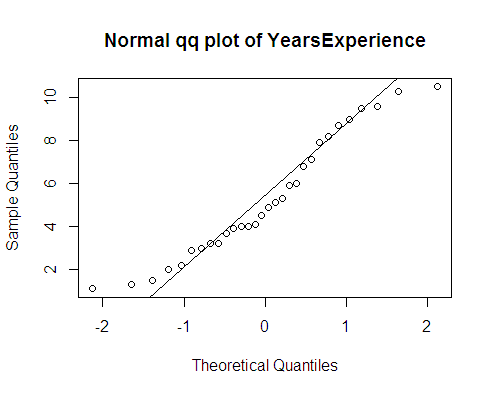
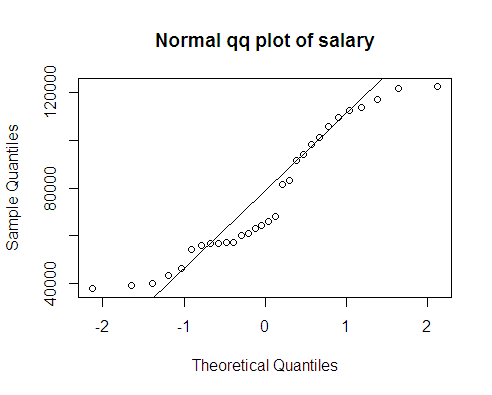
**Finding Normality of data**

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 plot of YearsExperience looks normal but for salary it is not normal. The Shapiro-Wilk and Anderson\_Darling normality tests give a p-value greater than 0.05 for YearsExperience. The same tests give a p-value less than 0.05 for salary.

We can conclude that YearsExperience is normally distributed. Salary is not normal.

Finding Missing values

There are no missing values in the entire dataset.

|  |
| --- |
| > salary\_hike[!complete.cases(salary\_hike),] |
| [1] YearsExperience Salary |
| <0 rows> (or 0-length row.names) |

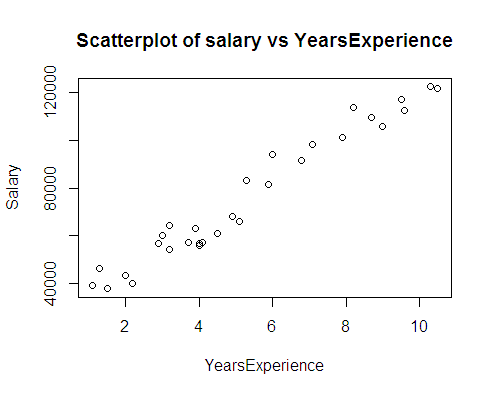
**4 Model Building and Interpretation**

We have pre-processed our dataset. The dataset has no missing values and there are no outliers. The target variable YearsExperience is not normal and predictor variable salary is normally distributed.

The business problem is to predict salary hike using experience in years. We need to build a prediction model. As seen, both are continuous variables. Here target variable is salary. We need to predict the outcome using only one independent variable YearsExperience. So we will be using 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 positive correlation. The correlation coefficient is 0.9782416. Both variables have a strong linear relationship.



**I Standard Linear Regression Model using Salary and YearsExperience**

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

|  |
| --- |
| > # Regression model |
| > reg\_simple <- lm(Salary ~ YearsExperience) |
| > summary(reg\_simple) |
|  |
| Call: |
| lm(formula = Salary ~ YearsExperience) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -7958.0 -4088.5 -459.9 3372.6 11448.0 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\* |
| YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 5788 on 28 degrees of freedom |
| Multiple R-squared: 0.957, Adjusted R-squared: 0.9554 |
| F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16 |

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

Salary = B0 + B1\* (YearsExperience)

Using above coefficients we can write this equation as:

Salary = 25792.2 + 9450.0 \* YearsExperience

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | Coeff of YearsExperience | YearsExperience | Predicted(y) | If X1 increases by 1 unit, increase in Y is |
| 1 | 2 | 3 | 4 | 5 |
| 25792.2 | 9450.0 | 2 | 44692.2 | 54142.2  -- 44692.2  = **9450** |
| 25792.2 | 9450.0 | 2.1 | 45637.2 |  |
| 25792.2 | 9450.0 | 3 | 54142.2 |  |

If there are no variables, then the average salary is 25792.2. We are now adding a new variable YearsExperience.

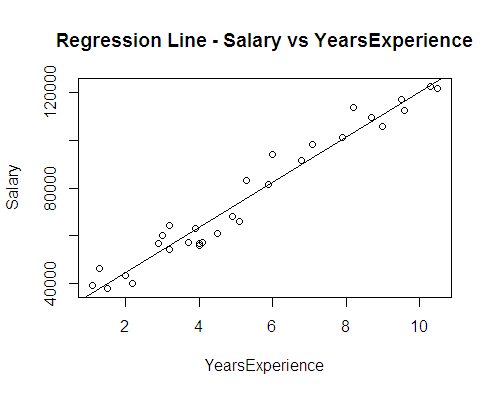
If YearsExperience is increased by 1 year, then predicted salary increases by 9450 (as shown in 5th column of above table) or if there is 0.1 unit increase in YearsExperience, the salary increases by 945.

If we have a new experience value of 2.1 years, then we can predict the salary as 45637.2.

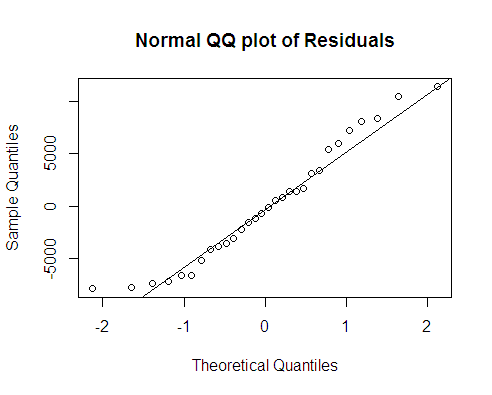
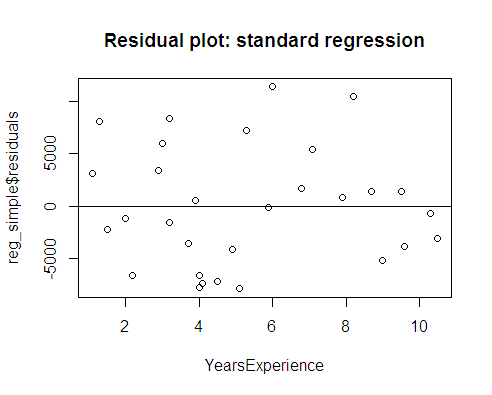
Pvalue of t-stat and F-statistic: For the t-value, the pvalue is = 0.000 for both intercept and YearsExperience. This indicates that both are highly significant in estimating salary. F-statistic suggests if the model is good or not. Its p-value is <0.05 suggesting that the model is significant in predicting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

R2 and Adjusted R2: R2 is the coefficient of determination. It indicates that about 95.7% of the variance of salary 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 (0.9554) is close to R2 value, indicating that the predictor is significant in predicting the target.

Now let us plot the regression line on the scatter plot of actual salary vs YearsExperience. 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 5592.044. We can compare this value with different models and select the model with least RMSE. The Residual plot shows random pattern. This suggests that our model is good. From below QQ plot of residuals we can say that they are normally distributed. Also, the p-value of Shapiro-Wilk and Anderson-Darling tests suggests normal distribution of residuals.

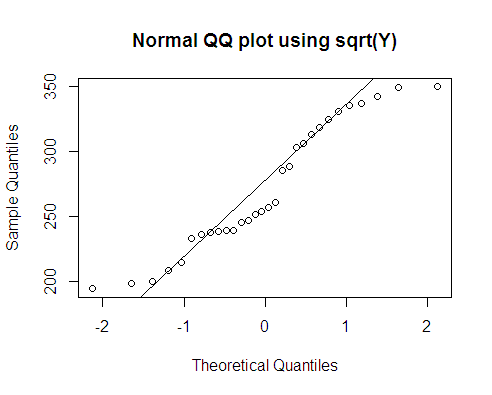


From above results, we can conclude that the model obtained is a good one.

But we have seen that the target variable was not normal. Let us try squareroot(Y) transformation and build the model.

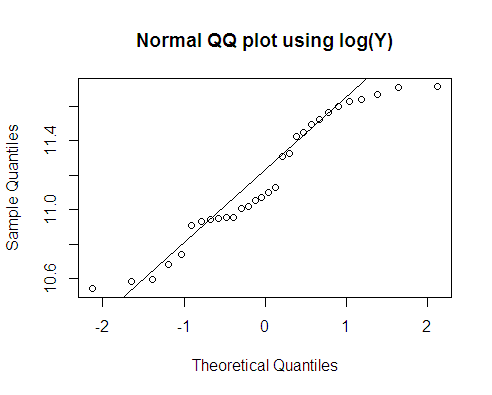
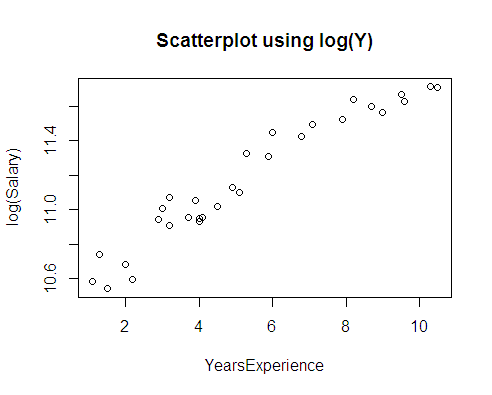
**II Quadratic Model using squareroot(Y)**

The basic concern here was that the target variable is not normal. Let us use sqrt(Y) and see if its normal. The value of Shapiro-Wilk and Anderson\_Darling tests are less than 0.05. The QQ plot and the test values suggest that the data is still not normal. We will not proceed further using sqrt(Y).



**III Exponential Model using log(Y)**

Let us now use log transformation of salary and find if its normal or not.The QQ plot and the normality tests suggest that log(salary) is now normally distributed. The scatterplot of log(salary) and YearsExperience shows strong positive linear relation. This is confirmed by correlation coefficient which is 0.9653844.

Regression Model

Let us now build regression model. The regression coefficients are significant as revealed by t-statistic and F-statistic. R2 value is 0.932, though it is good but less than the standard regression model.

|  |
| --- |
| > # regression - log(Y) |
| > reg\_log <- lm(log(Salary) ~ YearsExperience) |
| > summary(reg\_log) |
|  |
| Call: |
| lm(formula = log(Salary) ~ YearsExperience) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -0.18949 -0.06946 -0.01068 0.06932 0.19029 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 10.507402 0.038443 273.33 <2e-16 \*\*\* |
| YearsExperience 0.125453 0.006406 19.59 <2e-16 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 0.09789 on 28 degrees of freedom |
| Multiple R-squared: 0.932, Adjusted R-squared: 0.9295 |
| F-statistic: 383.6 on 1 and 28 DF, p-value: < 2.2e-16 |

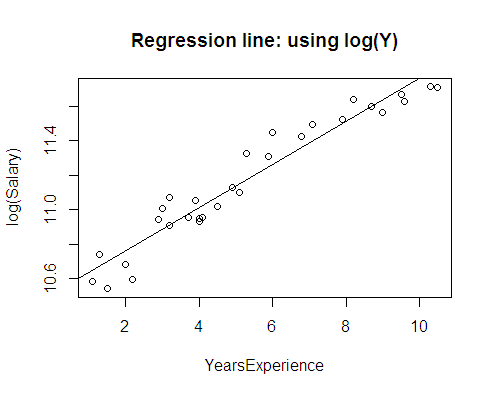
We can write the regression equation as follows:

Log(Y) = B0 + B1\* (X1)

log(salary) = 10.507402 + 0.125453 \* (YearsExperience)

The coefficients of intercept and YearsExperience are 10.507402 and 0.125453 respectively. We can interpret B1 by exponentiating, exp(0.125453) is 1.133662. If we increase experience by 1 year, then salary will increase by 13.37%.

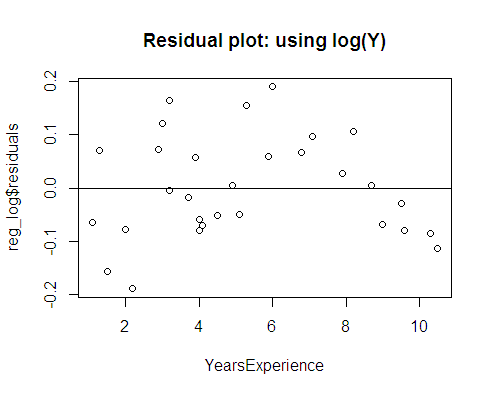
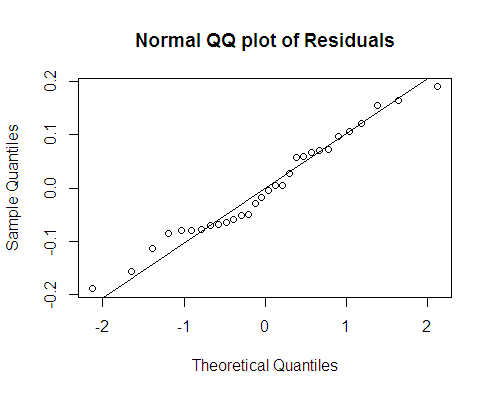
Now let us plot the regression line. The line passes close to almost all points, suggesting that this is a good model.



Residuals:

Residuals are the errors. The sum and mean of residuals is zero. RMSE is 0.0945. The residual plot suggests that the errors are random i.e., following homoscedasticity (constant variance) across the values of independent variable.

The residuals are normally distributed as shown in the above QQ plot and also the Anderson-Darling test gives a p-value of 0.337, which is > 0.05.

**Conclusion:**

The dataset has 2 variables salary\_hike and YearsExpereince. Both variables are continuous. As per the business problem we have predicted salary\_hike based on years of work experience. The predictor is normally distributed. Target variable is not normal and so we have used log transformation to normalise it.

We are using simple linear regreession technique to predict Salary. We have built two models: standard model using only Y and exponential model using log(Y).

The standard model has highest R2 and the exponential model has least RMSE. The correlation coefficient is highest with simple Y as compared to log(Y). The Residual plot of standard model was more random than the log(Y) model. We can conclude that the standard model is a better one.

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