Assignment 4

**Multi Linear Regression**

Part 1 of 3

**To predict Profit of 50 startups**

Submitted

To

logo



Submitted

By

Munawwara Shakila

(29-Jan-2020)

**Q1) Prepare a prediction model for profit of 50\_startups data**

**1. Business Problem**

The business objective is to predict the profit. There are 50 startup firms from 3 major cities of USA. Their major expenses have been incurred in a few departments. Based on these we have to predict Profit of startups. We will be using the versatile R program to predict Profit

**2. Data acquisition**

As per the business problem we have to predict profit of 50 startups. The dataset has 50 records and 5 variables. Below table shows first 6 records of the dataset. The first 3 variables give details of expenses that have been incurred on RD (Research and Development department), Admin (Administration department) and Marketing department. The states of USA from where these startups function are also mentioned. Last variable is the Profit that has been made by these firms.

We have to predict profit using these variables.

|  |
| --- |
| > head(startups) |
| RD Admin Marketing State Profit |
| 1 165349.2 136897.80 471784.1 New York 192261.8 |
| 2 162597.7 151377.59 443898.5 California 191792.1 |
| 3 153441.5 101145.55 407934.5 Florida 191050.4 |
| 4 144372.4 118671.85 383199.6 New York 182902.0 |
| 5 142107.3 91391.77 366168.4 Florida 166187.9 |
| 6 131876.9 99814.71 362861.4 New York 156991.1 |

|  |
| --- |
| > names(startups) |
| [1] "R.D.Spend" "Administration" "Marketing.Spend" "State" "Profit" |
| > dim(startups) |
| [1] 50 5 |

**3. Exploratory Data Analysis (EDA)**

Let us now understand the dataset. Only variable ‘State’ is of categorical type, others are all continuous in nature. To know about our dataset, we have to do univariate analysis: first we will find the four business moment decisions and then do visualizations.

|  |
| --- |
| > str(startups) |
| 'data.frame': 50 obs. of 5 variables: |
| $ RD : num 165349 162598 153442 144372 142107 ... |
| $ Admin : num 136898 151378 101146 118672 91392 ... |
| $ Marketing: num 471784 443899 407935 383200 366168 ... |
| $ State : Factor w/ 3 levels "California","Florida",..: 3 1 2 3 2 3 1 2… |
| $ Profit : num 192262 191792 191050 182902 166188 ... |

Business moment decisions

Let us first find the business moments. Moments are popularly used to describe the characteristics of a distribution. The below table gives the 4 business moments for the 4 continuous variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RD | Admin | Marketing | Profit |
| Mean | 73721.62 | 121344 | 211025 | 112012 |
| Median | 73051.08 | 122700 | 212716 | 107978 |
| Mode | 0 | --- | 0 | -- |
| Variance | 2107017150 | 784997271 | 14954920097 | 1624588173 |
| Std deviation | 45902.26 | 28017.8 | 122290 | 40306 |
| Range | 0 to 165349  165349 | 51283 to 182645  131362 | 0 to 471784  471784 | 14681 to 192261 |
| Skewness | 0.1590405 | - 0.4600745 | - 0.04372111 | 0.02191219 |
| Kurtosis | 2.194932 | 3.085538 | 2.275967 | 2.824704 |

First business moment decisions tell us about central tendency. Second business moment decisions tell us about the spread of data. The third and fourth business moments tell us about skewness and kurtosis. Skewness tells about asymmetry of the data and kurtosis about how peaked or flat the data is from normal distribution.

The mean and median are almost similar for all variables. There may not be any outliers and also the data may be normal.

Of all variables, Admin has the least spread (standard deviation of 28017). Marketing seems to have highest spread, but this can be better said by looking at boxplots.

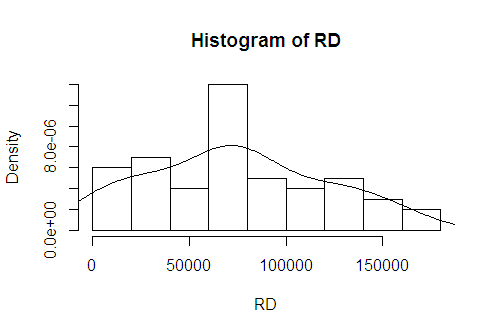
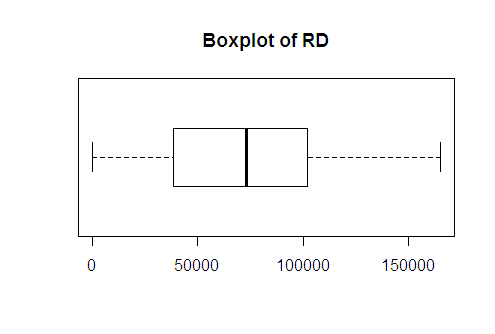
When we look at range, we see that Marketing expenses range upto 4,70,000 while for others upper limit is only 2,00,000. Since the values have different ranges, we can also standardize or normalize the data before running the model.

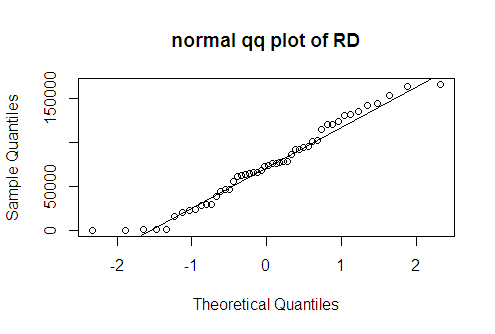
For all skewness is within ±0.5, indicating that the 4 variables have normal distribution. Even kurtosis varies from 2 to 3, suggests normality of data. More visualizations and normality tests are performed later to check normality. OLS method requires only the residuals to be normally distributed. But if we are using Maximum Likelihood method, then it requires that Y and all Xs are normally distributed.

**Visualizations**

R and D expenses

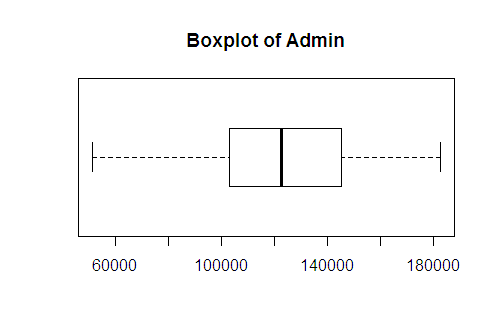
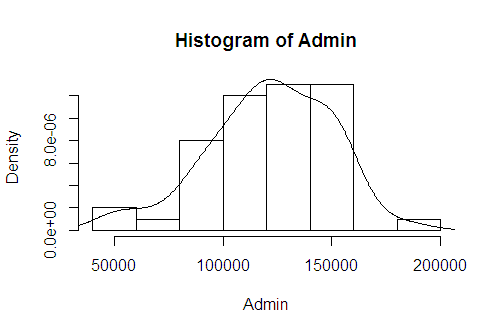
The boxplot shows that RD is slightly left skewed. There are no outliers. Histogram shows normal data. The qq-plot shows normal data. The Shapiro-Wilk and Anderson-Darling normality tests give a p-value more than 0.05, confirming normality of RD variable.

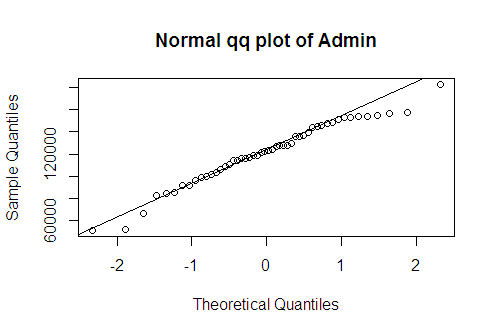




Administration expenses

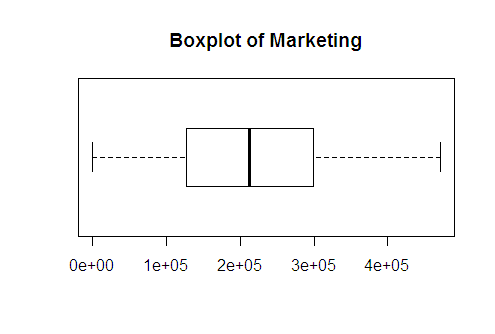
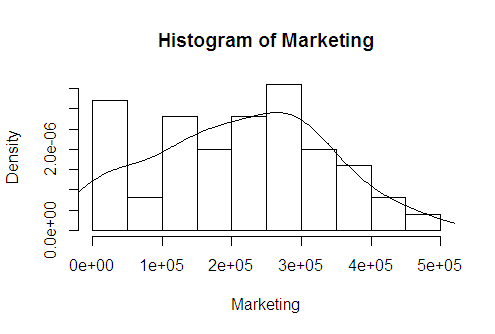
The histogram and boxplot of admin expenses show slightly right skewed data. Normal qqplot and normality tests show that the variable is normally distributed. There are no outliers.

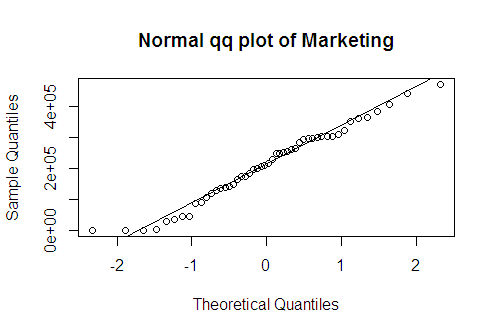




Marketing expenses

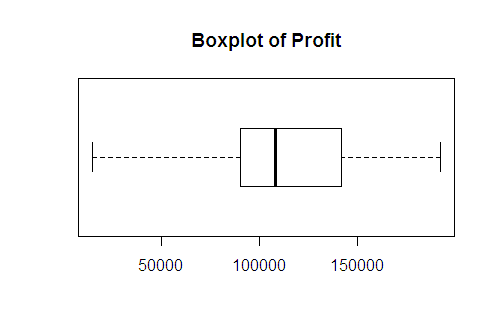
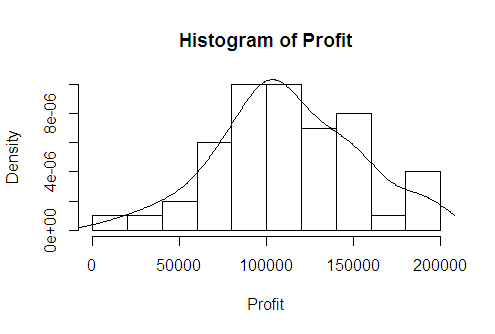
All the three plots show that Marketing expenses follow normal distribution. This is confirmed by Shapiro-Wilk and Anderson-Darling normality tests. There are no outliers.

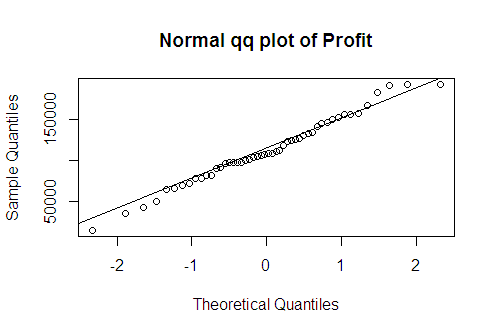




Profit

The histogram and boxplot show slightly left-skewed Profit variable. As per Shapiro and Anderson normality tests the Profit is normally distributed. There are no outliers.





State variable

The startups are operating in three states of USA *viz*, California, Florida and New-york.

|  |
| --- |
| > table(State) |
| State |
| California Florida New York |
| 17 16 17 |

Dummy variable creation:

State variable is of categorical nature. To use it for further analysis, we can convert it to dummy variables. In the below table we have converted State into 3 dummy variables *viz*, California, Florida and New York. For analysis, we will be dropping State variable (below 2nd table) and using these 3 dummy variables.

|  |
| --- |
| RD Admin Marketing State Profit California Florida New York |
| 1 165349.2 136897.80 471784.1 New York 192261.8 0 0 1 |
| 2 162597.7 151377.59 443898.5 California 191792.1 1 0 0 |
| 3 153441.5 101145.55 407934.5 Florida 191050.4 0 1 0 |
| 4 144372.4 118671.85 383199.6 New York 182902.0 0 0 1 |
| 5 142107.3 91391.77 366168.4 Florida 166187.9 0 1 0 |
| 6 131876.9 99814.71 362861.4 New York 156991.1 0 0 1 |

|  |
| --- |
| RD Admin Marketing Profit California Florida New York |
| 1 165349.2 136897.80 471784.1 192261.8 0 0 1 |
| 2 162597.7 151377.59 443898.5 191792.1 1 0 0 |
| 3 153441.5 101145.55 407934.5 191050.4 0 1 0 |

Finding missing values

The dataset has no missing values.

|  |
| --- |
| > # finding missing values in dataset |
| > startups[!complete.cases(startups),] |
| [1] RD Admin Marketing State Profit |
| <0 rows> (or 0-length row.names) |

We have pre-processed our dataset. The dataset has no missing values and there are no outliers. The variables Profit, RD, Admin and Marketin are continuous variables. They are all normally distributed. State variable which was categorical in nature has been converted to 3 dummy variables.

Our dataset is now ready for further analysis.

**4 Model Building and Interpretation**

The dataset is large with 50 records. The target variable is Profit. The predictors are RD, Admin, Marketing, California and Florida (we will be using n-1 categories for analysis: hence, New york is not included).

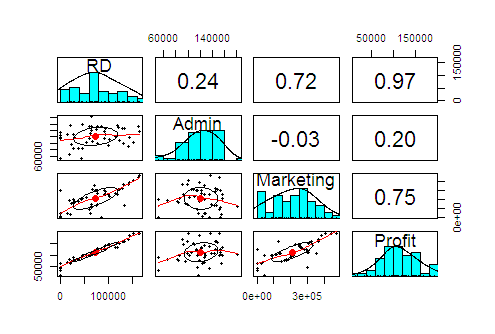
The business problem is to predict Profit. We need to build a prediction model. There are 6 predictor variables. So we will be using Multiple Linear Regression technique.

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 and correlation coefficient, we can say that the RD has a very strong positive correlation (0.973) with Profit. Marketing has strong positive correlation (0.748) with Profit. Admin has a very weak positive correlation (0.201) with Profit.

|  |
| --- |
| > # Correlation Coefficient matrix - Strength & Direction of Correlation |
| > cor(startups[,c(1,2,3,4)]) |
| RD Admin Marketing Profit |
| RD 1.0000000 0.24195525 0.72424813 0.9729005 |
| Admin 0.2419552 1.00000000 -0.03215388 0.2007166 |
| Marketing 0.7242481 -0.03215388 1.00000000 0.7477657 |
| Profit 0.9729005 0.20071657 0.74776572 1.0000000 |



# Partial correlation

> cor2pcor(cor(startups[,c(1,2,3,4)]))

[,1] [,2] [,3] [,4]

RD Admin Marketing profit

[1,] RD 1.00000000 0.20852619 0.03890336 0.93477127

[2,] Admin 0.20852619 1.00000000 -0.28192506 -0.07725021

[3,] Marketing 0.03890336 -0.28192506 1.00000000 0.23707116

[4,] profit 0.93477127 -0.07725021 0.23707116 1.00000000

Multicollinearity

The variables RD and Marketing have a strong positive correlation (0.724) among themselves. But since these 2 are predictors, this creates collinearity problem. When we calculate partial correlation, we find that RD and marketing have a very weak positive correlation (effects of other variables is removed when finding partial correlation). So collinearity may not be a problem in this regression analysis.

**I Standard Multi-Regression Model**

Let us now build our model using all variables and then check for collinearity. Since 3 categories of state variable have been converted to 3 dummy variables, only 2 dummy variables will be used for analysis. We will remove New\_York as it is an aliased variable.

|  |
| --- |
| > reg.profit <- lm(Profit ~ . - New\_York,data = startups) |
| > summary(reg.profit) |
|  |
| Call: |
| lm(formula = Profit ~ . - New\_York, data = startups) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -33504 -4736 90 6672 17338 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 5.008e+04 6.953e+03 7.204 5.76e-09 \*\*\* |
| RD 8.060e-01 4.641e-02 17.369 < 2e-16 \*\*\* |
| Admin -2.700e-02 5.223e-02 -0.517 0.608 |
| Marketing 2.698e-02 1.714e-02 1.574 0.123 |
| California 4.189e+01 3.256e+03 0.013 0.990 |
| Florida 2.407e+02 3.339e+03 0.072 0.943 |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 9439 on 44 degrees of freedom |
| Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452 |
| F-statistic: 169.9 on 5 and 44 DF, p-value: < 2.2e-16 |

|  |  |  |
| --- | --- | --- |
| **Coefficients for reg.profit** | | |
| intercept | 5.008346e+04 | 50083.46 |
| RD | 8.060231e-01 | 0.8060231 |
| Admin | -2.700432e-02 | 0.02700432 |
| Marketing | 2.697986e-0 | 0.02697986 |
| California | 4.188702e+01 | 41.88702 |
| Florida | 2.406758e+02 | 240.6758 |
| New\_York | 0.00 | 0.00 |

Coefficients:

We have now built a prediction model. The coefficients are the estimates of regression equation.

The Regression equation is :

ŷ = B0 + B1\*X1 + B2\*X2 + ……………… + Bn\*Xn + 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,

For nth variable, Xn is the variable value and Bn is its coefficient

ε is the error term

Profit = B0 + B1\* RD + B1\* Admin + B1 \* Marketing + B4 \* California + B5 \* Florida

Using above coefficients we can write this equation as:

Profit = 50083.46 + 0.806 \* RD + 0.027\*Admin + 0.027\*Marketing + 0 \*New\_york

+ 41.887\*California + 240.679\*Florida

Interpretation of coefficients:

When there are no variables, then the average profit will be 50,083 units.

If RD expenses is increased by 1 unit (ceteris paribus), then predicted Profit increases by 0.806 units or if there is 1000 units increase in RD, the Profit increases by 806 units.

If Admin expenses is increased by 1 unit (ceteris paribus), then predicted Profit increases by 0.027 units or if there is 1000 units increase in Admin, the Profit increases by 27 units.

If Marketing expenses is increased by 1 unit (ceteris paribus), then predicted Profit increases by 0.027 units or if there is 1000 units increase in RD, the Profit increases by 27 units.

For dummy variables, the interpretation is slightly different. When the firm operates from New York, keeping other conditions same, there is no change in profit. When firm operates from California, the profit increases by 42 units. When firm operates from Florida, the profit increases by 240 units.

Pvalue of t-stat and F-statistic: Only intercept and RD are highly significant in estimating Profit. Other variables are not significant. The F-statistic suggests that the model is significant in estimating Profit.

R2 and Adjusted R2: R2 is the coefficient of determination. It indicates that about 95.08% of the variance of Profit 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 the adj. R2 value (0.9452) is close to R2 value, indicating that the model is good.

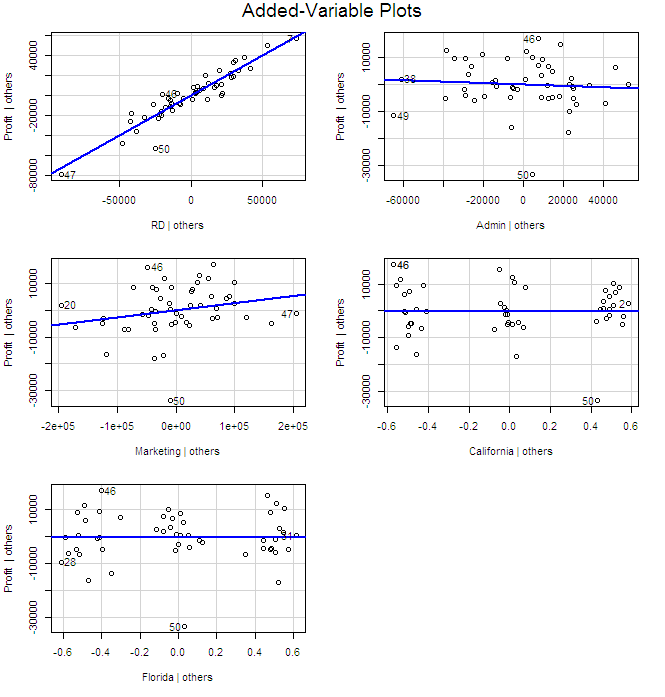
Now let us plot the regression line on the scatter plot of actual salary vs YearsExperience. The straight line gives the predicted Y values. The line passes close to almost all points. The model looks good. Above R2 value also suggests that the model is good. Next let us look at the residuals to confirm if it is a good model.

Multicollinearity: we have seen that RD and Marketing are correlated ( r = 0.724). Let us use variance inflation factor measure to know if there is multicollinearity among the variables. The value of vif is less than 4 for all variab;es, so there is no multicollinearity.

|  |
| --- |
| > library(car) |
| > vif(reg.profit) |
| RD Admin Marketing California Florida |
| 2.495511 1.177766 2.416797 1.335061 1.361299 |

Added Variable Plot

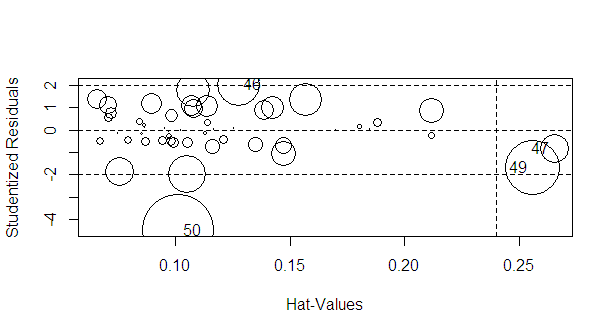
In the below plot, we see that amongst RD, Marketing and Admin, the variable that is least significant in predicting profit is Admin. So we can build another model by removing Admin variable.



**II Other Multi Linear Regression models**

We have built a few more models.

* reg.profit: This was the model built previously using all variables
* reg.profit.std : here we have standardized the variables and then used in analysis
* reg.profit.noadmin : here we have built model using all variables except Admin
* reg.profit.nomarketing : here we have built model using all variables except Marketing
* Final\_model : here we have built the model by using all variables and remove 50th record which showed highest influence as shown in below influence plot.



**Comparison of models**

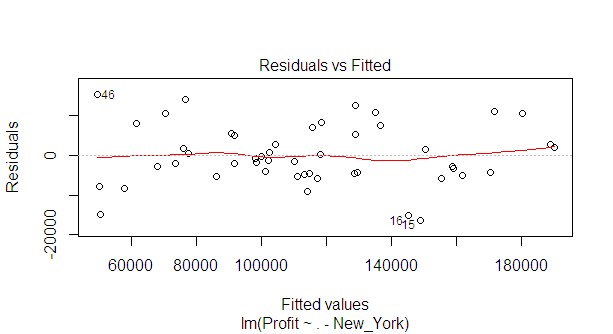
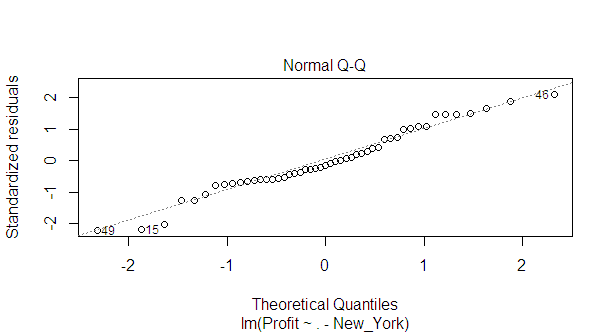
We find that R2 is more than 90% for all models. It is the highest for final-model i.e., 96.29. Also, intercept, RD and Marketing (<0.10) are all significant for final\_model.

We will select the final model as the better model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Reg.profit | reg.profit.std | Reg.profit.noadmin | reg.profit.  nomarketing | Final\_model |
|  |  |  |  |  |  |
| R2 | 0.9508 | 0.9508 | 0.9505 | 0.948 | 0.9618 |
| Adj R2 | 0.9452 | 0.9452 | 0.946 | 0.9434 | 0.9574 |
| F-stat | sig | sig | sig | Sig | Sig |
| RMSE | 8854.761 |  |  |  | 7383.181 |
|  |  |  |  |  |  |
| Intercept | Sig | Not sig | sig | Sig | Sig |
| RD | Sig | Sig | sig | Sig | Sig |
| Admin | Not sig | Not sig | --- | Not sig |  |
| Marketing | Not sig | Not sig | Sig < 0.10 | --- | Sig < 0.10 |
| California | Not sig | Not sig | Not sig | Not sig | Not sig |
| Florida | Not sig | Not sig | Not sig | Not sig | Not sig |
| New York | -- | -- | -- | -- | -- |

Residuals

The residuals are normally distributed and they do not follow any pattern. RMSE for final-model is 7383 which is less than the RMSE of reg-profit model (8854).

**Conclusions**

As per the business problem we have predicted Profit for 50 startup firms. Target variable is normally distributed. Some predictors are continuous and one is categorical in nature. We have converted categorical input variable into dummy variables and used for further analysis. There are no missing values in the dataset.

We are using multiple linear regreession technique to predict Profit. We have built different models by deleting variables, by deleting observations and standardizing the variables.

The different models are compared on the basis of R2, adjusted R2 and significance of inputs. We have obtained a good model (named final model) by using all variables in analysis and removing influential observations. This model has the highest R2 and least RMSE values. The residual analysis also shows that this is a good model.

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