**Exercise-1**

**Import the dataset in your R system – X50\_Startups**

> library(readxl)

> X50\_Startups <-

> read\_excel("E:/kamagyana/Computing/DARET/Assignments/50\_Startups.xlsx")

> View(X50\_Startups)

**Check the correlation coefficient of the numeric variables**

> nustart <- X50\_Startups %>% select("R&D Spend", "Administration", "Marketing Spend", "Profit")

> options(digits=10)

> rcorr (as.matrix(numstart)

R&DSpend Administration MarketingSpend Profit

R&DSpend 1.00 0.24 0.72 0.97

Administration 0.24 1.00 -0.03 0.20

MarketingSpend 0.72 -0.03 1.00 0.75

Profit 0.97 0.20 0.75 1.00

n= 50

P

R&DSpend Administration MarketingSpend Profit

R&DSpend 0.0905 0.0000 0.0000

Administration 0.0905 0.8246 0.1622

MarketingSpend 0.0000 0.8246 0.0000

Profit 0.0000 0.1622 0.0000

*NOTE: The correlations coefficients show that profit and R&D spending and marketing spending are highly correlated, and the correlation of these three variables with the administration is quite weak. R&D spend Marketing Spending are also reasonably fairly correlated. The correlation coefficients calculated are also statistically significant*

**Split the dataset into two parts – Training and testing datasets**

> trsize <- floor(0.75 \* nrow(X50\_Startups))

> set.seed(123)

> trdata <- sample(seq\_len(nrow(X50\_Startups)), size = trsize)

> train <- X50\_Startups[trdata,]

> test<- X50\_Startups[-trdata,]

**Make dummy variables for character variables**

Starup <- mutate(X50\_Startups,DONE = ifelse(X50\_Startups$State == "Florida",1,0))

Starup <- mutate(Starup,DTWO = ifelse(Starup$State == "New York",1,0))head(Starup)

*Note: Since there are three unique states, we create two dummy variables, to avoid the dummy variable trap. By introducing a dummy without any interaction effect with the explanatory variables, the intercept or the average profit (which is going to be the response variable)is expected to differ depending on the state analysed.*

**Fit the regression model on the dataset provided (ENTIRE DATASET)**

> Starup <- as.data.frame(Starup)

> regresuls <- lm(Starup$Profit ~ Starup$'r&dspend'+ Starup$Administration +Starup$MarketingSpend + Starup$DONE+Starup$DTWO)

> summary(regresuls)

Call:

lm(formula = Starup$Profit ~ Starup$"r&dspend" + Starup$Administration +

Starup$MarketingSpend + Starup$DONE + Starup$DTWO)

Residuals:

Min 1Q Median 3Q Max

-33503.639 -4735.510 89.919 6671.951 17337.715

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.012534e+04 6.884820e+03 7.28056 4.4442e-09

Starup$"r&dspend" 8.060231e-01 4.640697e-02 17.36858 < 2.22e-16

Starup$Administration -2.700432e-02 5.223155e-02 -0.51701 0.60774

Starup$MarketingSpend 2.697986e-02 1.714216e-02 1.57389 0.12268

Starup$DONE 1.987888e+02 3.371007e+03 0.05897 0.95324

Starup$DTWO -4.188702e+01 3.256039e+03 -0.01286 0.98979

(Intercept) \*\*\*

Starup$"r&dspend" \*\*\*

Starup$Administration

Starup$MarketingSpend

Starup$DONE

Starup$DTWO

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439.207 on 44 degrees of freedom

Multiple R-squared: 0.9507525, Adjusted R-squared: 0.9451562

F-statistic: 169.8892 on 5 and 44 DF, p-value: < 2.2204e-16

**Check the VIF of the highly correlated variables (ENTIRE DATASET)**

> vif(regresuls)

Starup$"r&dspend" Starup$Administration Starup$MarketingSpend

2.495510922 1.177766051 2.416796757

Starup$DONE Starup$DTWO

1.387641303 1.335060877

*Note: The VIF values of the two highly correlated variables R&D spend and Marketing spend are lesser than the acceptable range of 3 or 5. Their values are 2.49 and 2.41 respectively. Hence if they were to be considered together in a regression equation the problem of muliticollinearity would not arise, to a disturbing extent that they cannot be used to predict a particular response variable.*

**What inference you can draw from the dataset (ENTIRE DATASET)**

*NOTE: The correlation and regression results are moving in tune. Based on the dataset, we hypothesise based on common business understanding that Profit is the target variable of any business, and one would like to find out in a multi-state business enterprise, what influences profit. With the variables in the dataset, it appears, that R&D expenditure, Adminsitration expense, and Marketing Expenditure should be the predictor variables. All of these are numerical type variables. Additionally in case there is something called the “place” effect, then the impact of a particular state also can examined by introducing them as dummy variables.*

*Intially the analysis is performed on the entire analysis to find out which of the given variables would emerge as robust predictors. The regression results above first show based on F-Static, that the regression is not invalid, and atleast one of the considered variables definitely predicts the response variable profit. Further the results also show that only R&D Spend strongly predicts Profit with almost 100% accuracy. Administration and Marketing Spends do not seem to have any significant influence on profit. Further irrespective of any of the given predictor variables there is an average level of profit which is not influenced by the belongingness to a state.*

*In search of a much more robust and accurate model to predict the following backward elimination method is attempted and the results are as follows.*

**Use backward elimination method to get the final model(ENTIRE DATASET)**

> regresuls1 <- lm(Starup$Profit ~ Starup$'r&dspend'+Starup$Administration+Starup$MarketingSpend)

> summary(regresuls1)

Call:

lm(formula = Starup$Profit ~ Starup$"r&dspend" + Starup$Administration +

Starup$MarketingSpend)

Residuals:

Min 1Q Median 3Q Max

-33533.734 -4795.026 62.653 6606.496 17275.430

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.012219e+04 6.572353e+03 7.62622 1.0574e-09

Starup$"r&dspend" 8.057150e-01 4.514727e-02 17.84637 < 2.22e-16

Starup$Administration -2.681597e-02 5.102878e-02 -0.52551 0.60176

Starup$MarketingSpend 2.722806e-02 1.645123e-02 1.65508 0.10472

(Intercept) \*\*\*

Starup$"r&dspend" \*\*\*

Starup$Administration

Starup$MarketingSpend

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232.335 on 46 degrees of freedom

Multiple R-squared: 0.950746, Adjusted R-squared: 0.9475338

F-statistic: 295.9781 on 3 and 46 DF, p-value: < 2.2204e-16

> regresuls2 <- lm(Starup$Profit ~ Starup$'r&dspend'+ Starup$MarketingSpend)

> summary(regresuls2)

Call:

lm(formula = Starup$Profit ~ Starup$"r&dspend" + Starup$MarketingSpend)

Residuals:

Min 1Q Median 3Q Max

-33645.494 -4632.426 -414.014 6483.898 17096.506

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.697586e+04 2.689933e+03 17.46358 < 2e-16 \*\*\*

Starup$"r&dspend" 7.965840e-01 4.134758e-02 19.26556 < 2e-16 \*\*\*

Starup$MarketingSpend 2.990788e-02 1.552001e-02 1.92705 0.06003 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9160.966 on 47 degrees of freedom

Multiple R-squared: 0.9504503, Adjusted R-squared: 0.9483418

F-statistic: 450.7713 on 2 and 47 DF, p-value: < 2.2204e-16

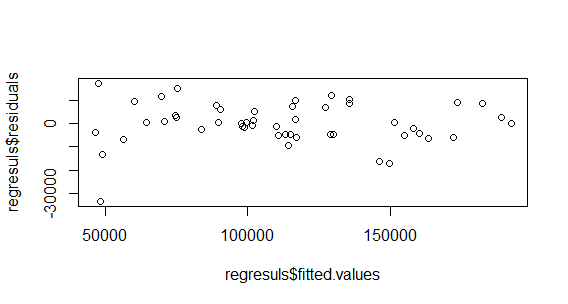
*Note: Gradually removing the State Dummies, and then the weakly correlated Administration related expense the final model displays a high F statistic with nearly 100% confidence and the two strongly correlated variables R&D Spend and Marketing Spend predict better Profits. R&D Spend is the most robust predictor variable, Marketing Spend is significant only at 6% level. The results convey that an amount of 46795 is the average profitability of a startup irrespective of any predictor variable and the state of its location. Further every 1 unit increase in the R&D expenditure leads to 0.7 units of increase in profits, and every 1 unit increase in Marketing Expense increases profits by 0.02 units.*

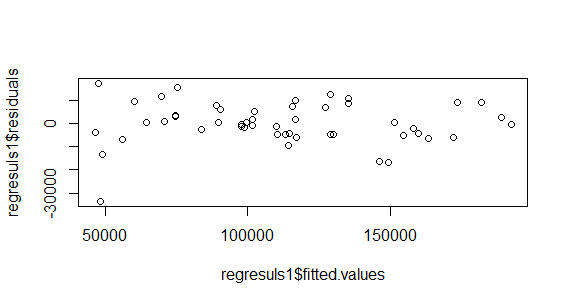
**Check the residual plots for the diagnosis (ENTIRE DATASET)**

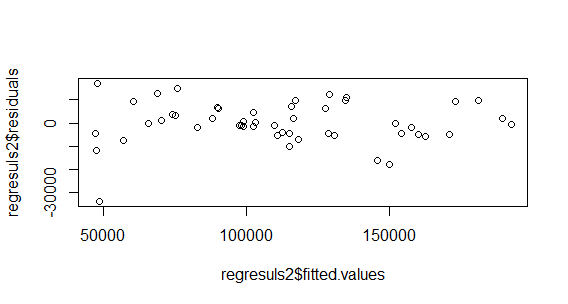
> plot(regresuls$fitted.values, regresuls$residuals)

> plot(regresuls1$fitted.values, regresuls1$residuals)

> plot(regresuls2$fitted.values, regresuls2$residuals)







*NOTE: All the residual plots seem to be randomly distributed. It means that there is no pattern in them, so the dependent variable might also be normally distributed and homoscedastic. In this way we can conclude that the regression results are dependable.*

**Fitting the best model with Training dataset**

> trdata <- sample(seq\_len(nrow(Starup)), size = trsize)

> train <- X50\_Startups[trdata,]

> test<- X50\_Startups[-trdata,]

> trainreg <- lm(train$Profit ~ train$'r&dspend'+ train$MarketingSpend)

> summary(trainreg)

Call:

lm(formula = train$Profit ~ train$"r&dspend" + train$MarketingSpend)

Residuals:

Min 1Q Median 3Q Max

-18546.2 -5146.0 -548.4 5836.3 15152.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.893e+04 2.611e+03 18.740 <2e-16 \*\*\*

train$"r&dspend" 8.085e-01 3.801e-02 21.272 <2e-16 \*\*\*

train$MarketingSpend 2.045e-02 1.401e-02 1.459 0.154

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7656 on 34 degrees of freedom

Multiple R-squared: 0.9655, Adjusted R-squared: 0.9635

F-statistic: 476.3 on 2 and 34 DF, p-value: < 2.2e-16

*NOTE: The entire dataset of 50 observations is classified into 75% training dataset and 25% testing dataset. This is best practice to fit a model and predict any response variable.*

**Predict the value of the test dataset with the model**

> test <- mutate(test, preprofit = (48930 + 0.8085\*(test$'r&dspend')) )

> head(test)

r&dspend Administration MarketingSpend State Profit preprofit

1 142107.34 91391.77 366168.4 Florida 166187.9 163823.8

2 131876.90 99814.71 362861.4 New York 156991.1 155552.5

3 123334.88 108679.17 304981.6 California 149760.0 148646.3

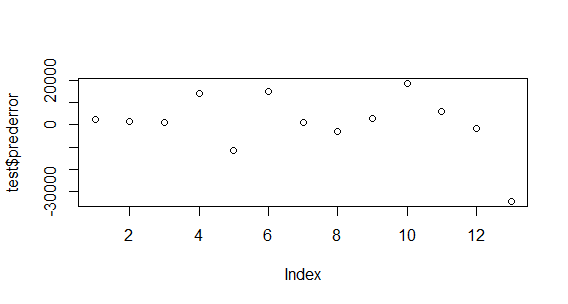
4 100671.96 91790.61 249744.5 California 144259.4 130323.3

5 114523.61 122616.84 261776.2 New York 129917.0 141522.3

6 78013.11 121597.55 264346.1 California 126992.9 112003.6

> test <- mutate(test, prederror = (test$Profit - test$preprofit))

> plot(test$prederror)



> library(stats)

> shapiro.test(x = test$prederror)

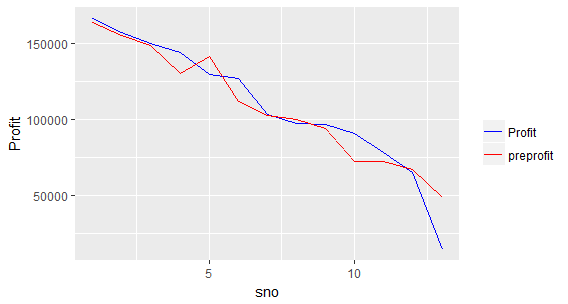
Shapiro-Wilk normality test

data: test$prederror

W = 0.86115, p-value = 0.0399

> test$sno <- 1:nrow(test)

> ggplot(test,aes(x = sno)) + geom\_line(aes(y = Profit, colour = "Profit")) + geom\_line(aes(y = preprofit, colour = "preprofit")) + scale\_colour\_manual("", breaks = c("Profit","preprofit"), values = c("red","blue")) + ylab(label = "Profit") + xlab(label = "sno")



**What you can infer from the prediction**

*NOTE: After the training data set has resulted into an intercept of 48930 and the coefficient of R&D Spend is 0.8085. The Marketing Spend variable loses its significance due the reduction in number of variables from 50 to 35 (it appears). So considering only the R&D Spend prediction of Profits with the data in the testing dataset is undertaken.*

*The prediction error is plotted and it is not showing any specific pattern and when tested for normality, it’s distribution is not normal. So the predictive accuracy is not absolutely 100%, there are definitely some errors in prediction. The model which we have chosen has am adjusted r@ of 96.35%, which means that on an average 96.35% of the variance in profits of a company are definitely explained by the R&D Spend and Marketing Expenditure. The profitability of a Startup irrespective of any explanatory variable on an average is 48930 units. So it can be concluded that Startups do enjoy profits. Their profits would be further increase due to the R&D expenditure of the firm. It appears from this trend that the data should pertain to some technology or R&D type of businesses.*

*The line charts also show that the predictive accuracy should be further increased. As of now only 96.35% of the profits figure on an average can be predicted. There might be some other variables which are not considering due to which the predictive accuracy is less than 100%. In business and management areas of research this accuracy is considered to be pretty decent.*

**Position salary.csv**

> Salary\_Data <- read.csv("E:/kamagyana/Computing/DARET/Assignments/Salary\_Data.csv", stringsAsFactors=FALSE)

> View(Salary\_Data)

**After importing dataset in your R system, check the scatter plot to decide on the shape of the model**

> plot(Salary\_Data$YearsExperience, Salary\_Data$Salary)



*NOTE: The shape appears to be a linear model with a positive slope and the intercept is not clear due to the scaling of a graph. Meaning that the salary increases as the years of experience increase and a fresher would get quite less salary.*

**Fit a linear regression model**

> reg <- lm(Salary\_Data$Salary ~ Salary\_Data$YearsExperience)

> summary(reg)

Call:

lm(formula = Salary\_Data$Salary ~ Salary\_Data$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 \*\*\*

Salary\_Data$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

> abline(reg, col="red")



*NOTE: The regression results show that number of years of experience robustly predicts the salary. Each of the sample element observed shows an average salary of 25792 and for every 1 unit of increase in years of experience the salary would increase by 9450. Around 95.54% of the changes in salary is being predicted by number of years of experience.*

**Form a polynomial regression with on the dataset**

> reg1 <- lm(Salary\_Data$Salary ~ poly(Salary\_Data$YearsExperience,3))

> summary(reg1)

Call:

lm(formula = Salary\_Data$Salary ~ poly(Salary\_Data$YearsExperience,

3))

Residuals:

Min 1Q Median 3Q Max

-7468 -4286 -1100 2639 10412

Coefficients:

Estimate Std. Error t value

(Intercept) 76003.0 1008.6 75.358

poly(Salary\_Data$YearsExperience, 3)1 144419.0 5524.1 26.144

poly(Salary\_Data$YearsExperience, 3)2 635.1 5524.1 0.115

poly(Salary\_Data$YearsExperience, 3)3 -12013.4 5524.1 -2.175

Pr(>|t|)

(Intercept) <2e-16 \*\*\*

poly(Salary\_Data$YearsExperience, 3)1 <2e-16 \*\*\*

poly(Salary\_Data$YearsExperience, 3)2 0.9094

poly(Salary\_Data$YearsExperience, 3)3 0.0389 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5524 on 26 degrees of freedom

Multiple R-squared: 0.9636, Adjusted R-squared: 0.9594

F-statistic: 229.4 on 3 and 26 DF, p-value: < 2.2e-16