2017

Data Analytics Project Report

LINEAR REGRESSION ANALYSIS ON SERVO DATA

SIDHANT ARORA 101403176
SOHAM GHOSH 101403178
SOUMIK PRADHAN 101403179
SUDHANSHU KAUL 101403180

DATA ANALYTICS

PROJECT REPORT

Introduction

In statistical modelling, **regression** analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

Root Mean Squared Error (RMSE)

The square root of the mean/average of the square of all of the error.

The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions.

Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

R code:

```
RMSE <- sqrt(mean((y-y_pred)^2))
```

Multiple Linear Regression Model

A linear regression model that contains more than one predictor variable is called a *multiple linear regression model*. The following model is a multiple linear regression model with two predictor variables, x_1 and x_2.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The model is linear because it is linear in the parameters β_0, β_1 and β_2. The model describes a plane in the three-dimensional space of Y, x_1 and x_2. The parameter β_0 is the intercept of this plane. Parameters β_1 and β_2 are referred to as *partial regression coefficients*. Parameter β_1 represents the change in the mean response corresponding to a unit change in x_1 when x_2 is held constant. Parameter β_2 represents the change in the mean response corresponding to a unit change in x_2 when x_1 is held constant. Consider the following example of a multiple linear regression model with two predictor variables, x_1 and x_2 :

$$Y = 30 + 5x_1 + 7x_2 + \epsilon$$

This regression model is a first order multiple linear regression model.

About Data:

Data Source: Servo Data Set
Data Analytic Task: Regression
Type: Linear

Abstract: Data was from a simulation of a servo system

Data Set Characteristics:	**Multivariate**	**Number of Instances:**	**167**	**Area:**	**Computer**
Attribute Characteristics:	**Categorical, Integer**	**Number of Attributes:**	**4**	**Date Donated**	**1993-05-01**
Associated Tasks:	**Regression**	**Missing Values?**	**No**	**Number of Web Hits:**	**58245**

Source:

Creator:
Karl Ulrich (MIT)
Donor:
Ross Quinlan

Data Set Information:

Ross Quinlan:
This data was given to me by Karl Ulrich at MIT in 1986. I didn't record his description at the time, but here's his subsequent (1992) recollection:
"This is an interesting collection of data provided by Karl Ulrich. It covers an extremely non-linear phenomenon - predicting the rise time of a servomechanism in terms of two (continuous) gain settings and two (discrete) choices of mechanical linkages."

Attribute Information:

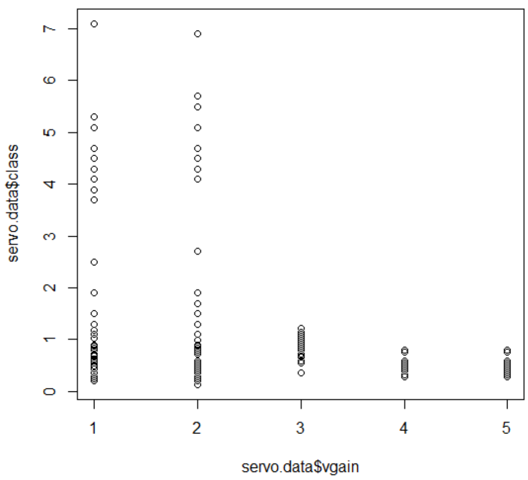
1. motor: A,B,C,D,E
2. screw: A,B,C,D,E
3. pgain: 3,4,5,6
4. vgain: 1,2,3,4,5
5. class: 0.13 to 7.10

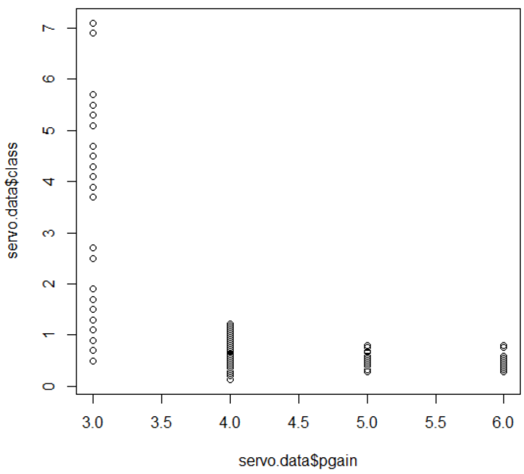
Data Verification

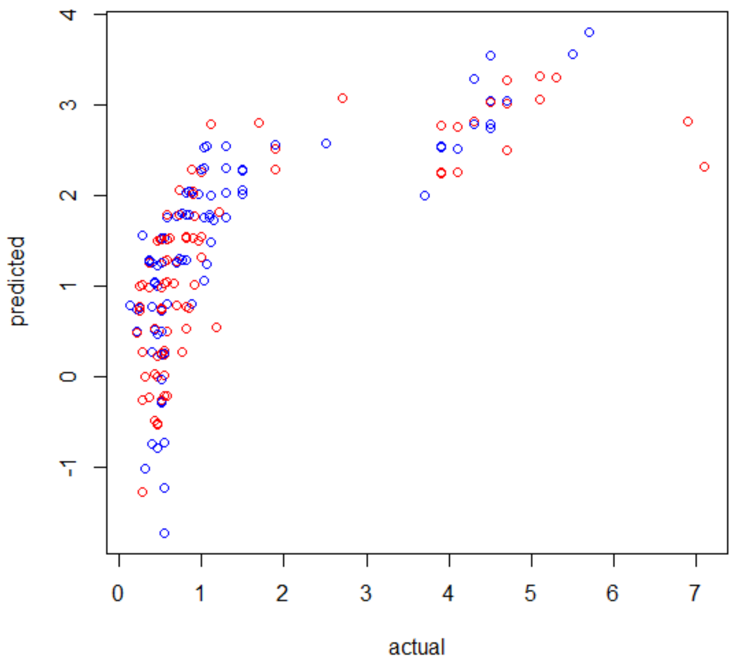
Data was verified whether it was within the constraints as mentioned in the data description.

```
> # verification of attributes
> levels(servo$motor)
[1] "A" "B" "C" "D" "E"
> levels(servo$screw)
[1] "A" "B" "C" "D" "E"
> unique(servo$pgain)
[1] 5 6 4 3
> unique(servo$vgain)
[1] 4 5 3 2 1
>
> # checking if class values are in range of 0.13 and 7.10
> check = 1
> for(i in servo$class){
+ if(i >= 0.13 && i <= 7.10){
+ check = check + 1
+ }
+ }
> print(paste("all ", check, " values are in range"))
[1] "all  167  values are in range"
```

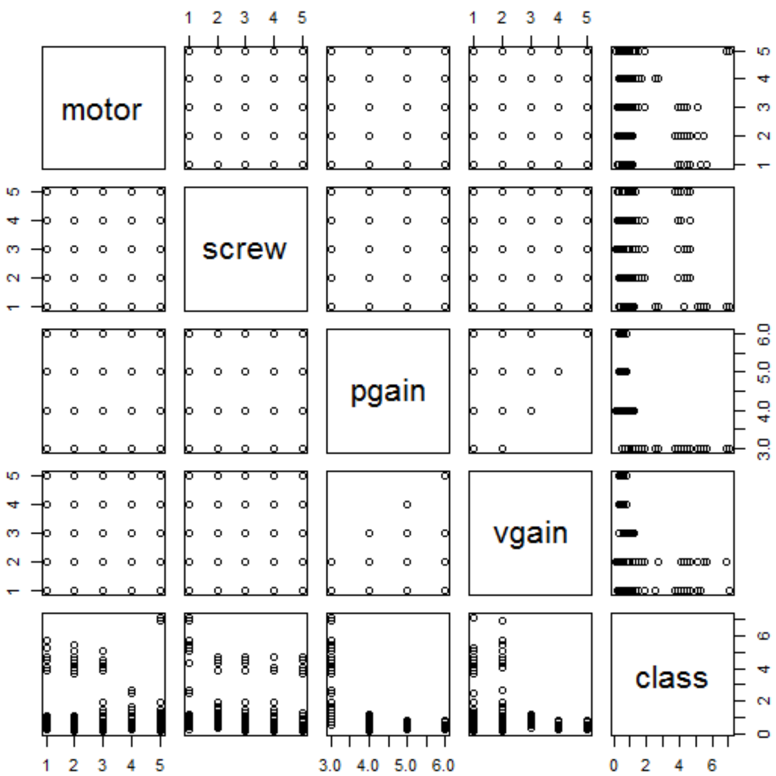
Data Visualisation







The scatter plot shows difference in the actual class and the predicted class



Plot of the Data Set

Formula used :

```
# regression btw pgain and vgain
lm.out = lm(class ~ pgain + vgain + motor + screw , data=servo.data)
lm.out

# Function that returns Root Mean Squared Error
rmse <- function(error)
{
   sqrt(mean(error^2))
}

# Function that returns Mean Absolute Error
mae <- function(error)
{
   mean(abs(error))
}
```

Summary and Coefficients of Regression:

lm(formula = class ~ pgain + vgain + motor + screw, data = servo.data)

Coefficients:

(Intercept)	pgain	vgain	motor	screw
7.8653	-1.5149	0.5016	-0.2453	-0.2654

Root Mean Square Error: 1.098131
Mean Absolute Error: 0.8896696

Conclusion:

After the analysis of Servo data set,we concluded that the class is dependent on four factors namely motor, screw, pgain and vgain.

Using this model we can predict classes for a given set of pgain, vgain, motor and screw.

We calculated root mean square error which came out to be 1.098131 and mean absolute error of 0.8896696.