**Analyze the report of Swedish Motor Insurance**

***Business Analytic Foundation with R Tools- Question***

**Question**

The data gives the details of third party motor insurance claims in Sweden for the year 1977. In Sweden, all motor insurance companies apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined. The data were compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The Committee was asked to look into the problem of analyzing the real influence on the claims of the risk arguments and to compare this structure with the actual tariff.

**Answers:**

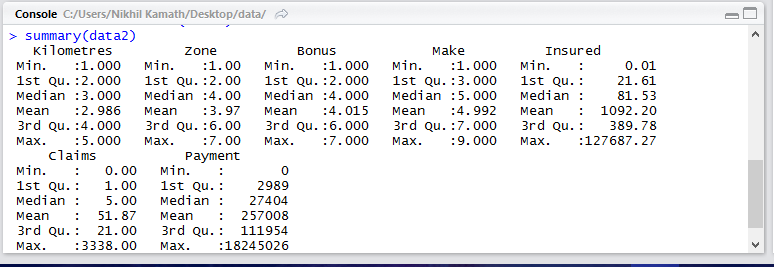
* Data understanding:
* Industry: Insurance
* 2182 observations, 7 variables
* Data is related to car insurance claims
* Data includes kilometers, zone, bonus, make, insured, claims, payment
* The goals of the project are detailed as below with solutions:
  1. The committee is interested to know each field of the data collected through descriptive analysis to gain basic insights into the data set and to prepare for further analysis.

**Code:**

**data1<-read.csv("C:/Users/Nikhil Kamath/Desktop/data/SwedishMotorInsurance.csv", header= TRUE,sep=",")**

**data2 <- data.frame(data1)**

**summary(data2)**



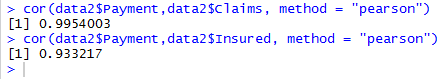
The above result shows the summary of the dataset. This summary details on the maximum, minimum, 1st and 3rd quartile, and mean values for the variables in the dataset in question.

* 1. The total value of payment by an insurance company is an important factor to be monitored. So the committee has decided to find whether this payment is related to number of claims and the number of insured policy years. They also want to visualize the results for better understanding.

**Code:**

**cor(data2$Payment,data2$Claims, method = “pearson”)**

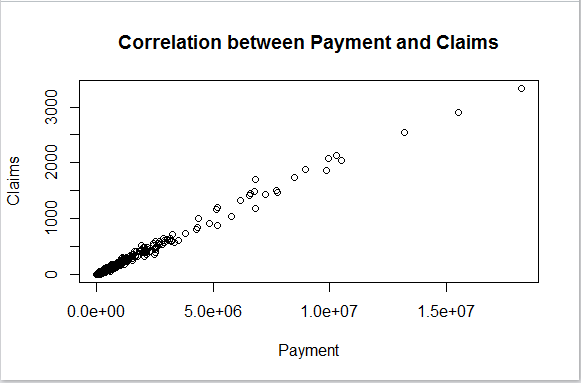
**cor(data2$Payment,data2$Insured, method = “pearson”)**



The above result shows the relation between the payment and the number of claims and insured policy years. So basically, here, we can see that there is a 99.5% correlation between payment and claims, and there is a 93.3% correlation between payment and insured policy years. As shown above, we use the cor( ) function to produce correlations. Both these correlations are positive. This means that the payment is directly proportional to both claims and insured amount. These relations are illustrated below with the help of a scatterplot.

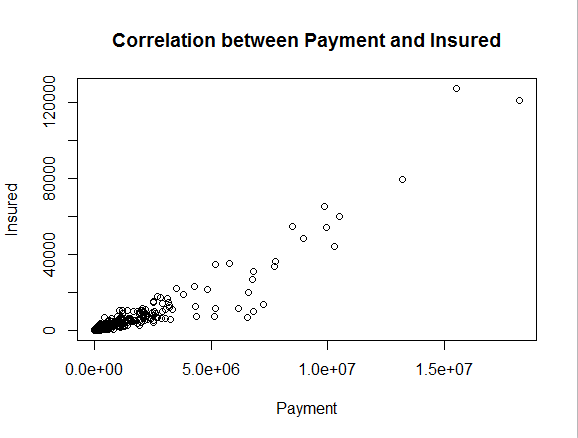
* Relation between payment and Claims:

**Code: plot(data2$Payment,data2$Claims, main = "Correlation between Payment and Claims",xlab = "Payment", ylab = "Claims")**



* Relation between payment and number of Insured policy years:

**Code**: **plot(data2$Payment,data2$Insured, main = "Correlation between Payment and Insured",xlab = "Payment", ylab = "Insured")**

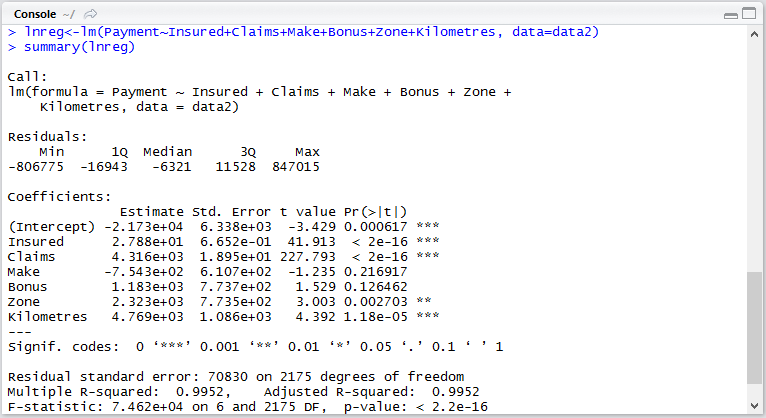


* 1. The committee wants to figure out the reasons for insurance payment increase and decrease. So they have decided to find whether distance, location, bonus, make, and insured amount or claims are affecting the payment or all or some of these are affecting it.

**Code:**

**lnreg<-lm(Payment~Insured+Claims+Make+Bonus+Zone+Kilometres, data=data2)**

**summary(lnreg)**



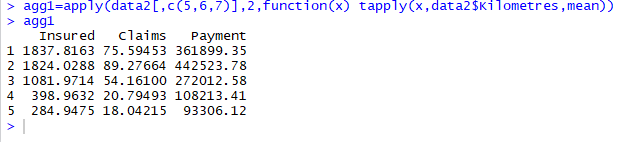
Here, the dependent variable is Payment, and the remaining variables are independent. To find out whether the payment variable is dependent on the remaining variables as asked above, we need to use Predictive Modeling, i.e. Linear Regression. Since there is only one dependent variable, i.e. payment, while the remaining variables are independent, we use multiple linear regression here. Linear regression in R is used by using the lm function as shown above. We can infer the following from the model:

* Since the t-values for the insured amount, claims, zone, and kilometers are less than 0.05, these are considered significant variables in the model, and we can say that these variables are the ones affecting the payment.
* Slopes of the significant variables are all positive, which indicated a positive relationship with the dependent variable, i.e. payment (Direct proportionality)
* The R^2 value is 0.9952, i.e. 99.52%, which is greater than the required 70%, which means that this model is an ACCURATE model.
* The adjusted R^2 value is the same as the R^2 value.
  1. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometer, and bonus level their insured amount, claims, and payment get increased. (Hint: Aggregate Dataset)

**Code:**

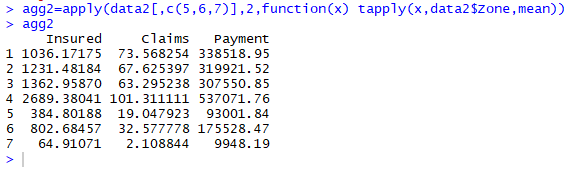
**agg1=apply(data2[,c(5,6,7)],2,function(x) tapply(x,data2$Kilometres,mean))**

**agg1**



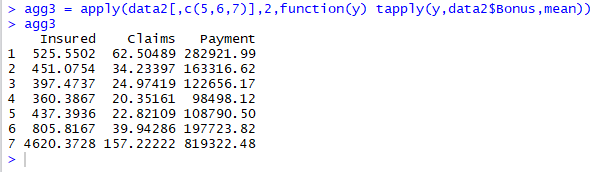
**agg2=apply(data2[,c(5,6,7)],2,function(x) tapply(x,data2$Zone,mean))**

**agg2**



**agg3 = apply(data2[,c(5,6,7)],2,function(y) tapply(y,data2$Bonus,mean))**

**agg3**



Here, we find the mean value of value of insured, payment, and claims based on zone, kilometre, and bonus variables, group all the result variables based on individual categorical variables. We have used the ‘apply’ and ‘tapply’ functions here, which are used in data manipulation. The apply function is used to manipulate any changes in rows and columns of the data. The tapply function needs 3 arguments – vector, factor of vector, and functions.

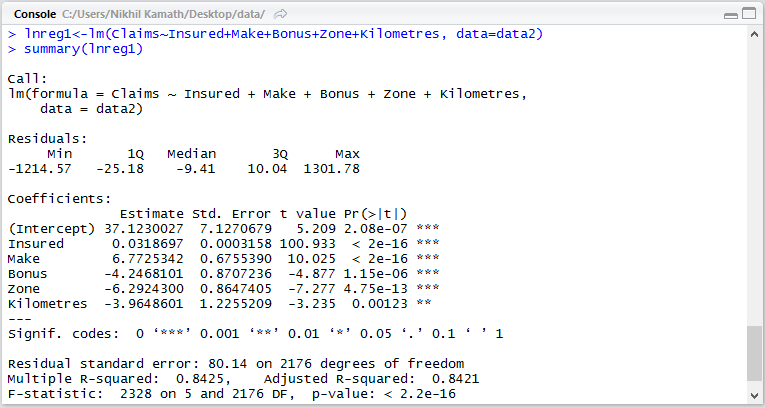
We can observe from the results that

1. **Zone 1-4 have more insured years.**
2. **The group 2 kilometer has maximum payments though are insured lesser number of years than Group1 and even have higher claims than group1.**
3. **Zone 4 has highest claims hence higher payment.**
4. **Group 7 has extremely high amount of insured years, Payments, claims in groups of bonus.**
   1. The committee wants to understand what affects their claim rates so as to decide the right premiums for a certain set of situations. Hence, they need to find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent.

**Code:**

**lnreg1<-lm(Claims~Insured+Make+Bonus+Zone+Kilometres, data=data2)**

**summary(lnreg1)**



Here, the dependent variable is Claims, and the remaining variables are independent. To find out whether the payment variable is dependent on the remaining variables as asked above, we need to use Predictive Modeling, i.e. Linear Regression. Since there is only one dependent variable, i.e. claims, while the remaining variables are independent, we use multiple linear regression here. Linear regression in R is used by using the lm function as shown above. We can infer the following from the model:

* Since the t-values for all the insured amount, make, bonus, zone, and kilometers are less than 0.05, these are all considered significant variables in the model, and we can say that these variables are the ones affecting the claims.
* Slopes of the significant variables bonus, zone, and kilometers are negative, which means these variables have a negative or inverse relationship with respect to claims.
* The R^2 value is 0.8425, i.e. 84.25%, which is greater than the required 70%, which means that this model is an ACCURATE model.
* The adjusted R^2 value is 0.8421, or 84.21%, which is less than the R^2 value.
* From this model, we can infer that all insured amount, bonus, zone and kilometers have a significant impact on the claim rates.