

Abstract

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.

**Analyze the report of Swedish Motor Insurance**

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

**Presented by: Ankita Agarwal**

**Problem Statement:**

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.

**Detailed description of the given dataset:**

|  |  |
| --- | --- |
| Variable | Description |
| Kilo-meters | Kilo-meters travelled per year  1: < 1000  2: 1000-15000  3: 15000-20000  4: 20000-25000  5: > 25000 |
| Zone | Geographical zone  1: Stockholm, Goteborg, and Malmö with surroundings  2: Other large cities with surroundings  3: Smaller cities with surroundings in southern Sweden  4: Rural areas in southern Sweden  5: Smaller cities with surroundings in northern Sweden  6: Rural areas in northern Sweden  7: Gotland |
| Bonus: | No claims bonus; equal to the number of years, plus one, since the last claim |
| Make: | 1-8 represents eight different common car models. All other models are combined in class 9. |
| Insured: | Number of insured in policy-years |
| Claims: | Number of claims |
| Payment: | Total value of payments in SKR (Swedish Krona) |

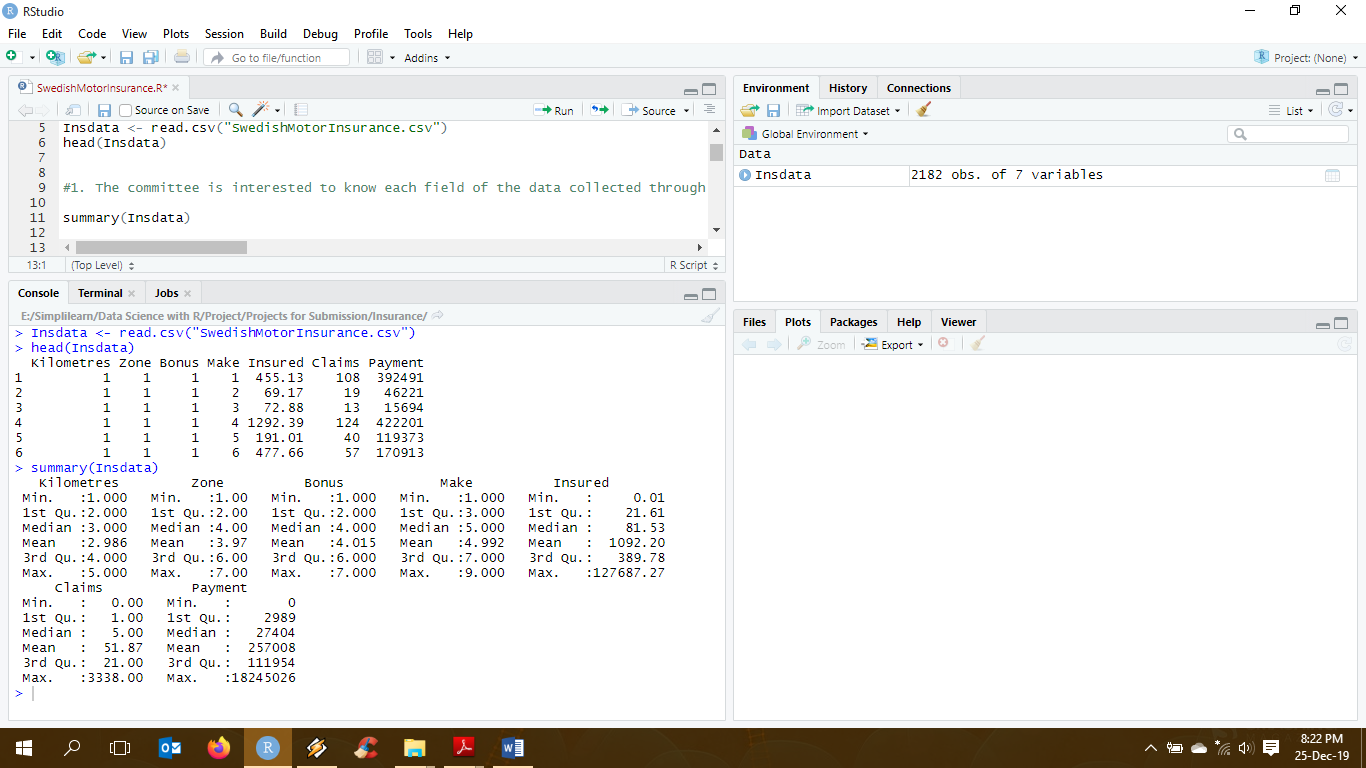
**To Analyze:**

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.
2. 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.
3. 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.
4. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased. (Hint: Aggregate Dataset)
5. 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, kilo meter, bonus, or make affects the claim rates and to what extent.

**Analysis and Interpretations:**

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.**

To get a basic insight into the data, we do a 5-point summary analysis for the data set.

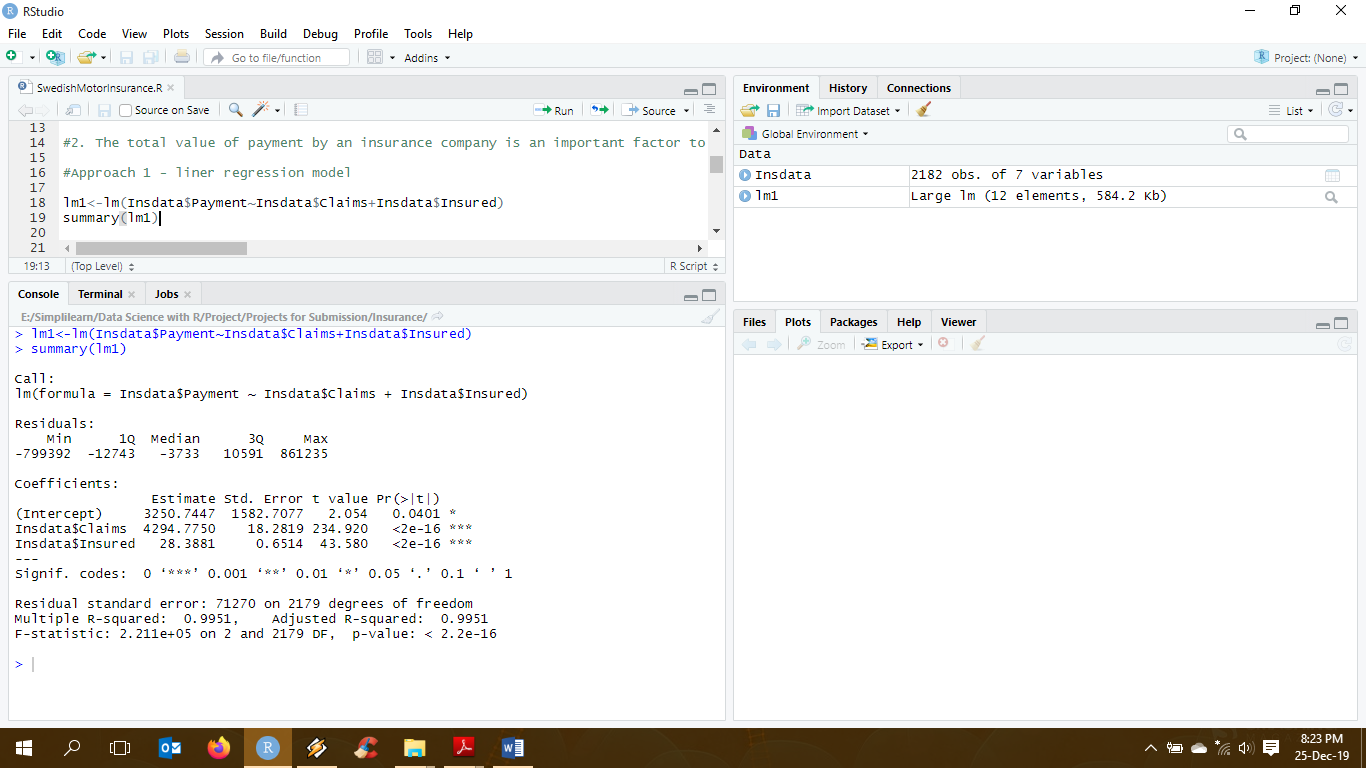


**Interpretation:**

By looking at the 5-point summary of the data, we can safely say that the range between the minimum and maximum data for Claims and Payment is very large with 0 being the least and 3338 and 18245026 being the maximum respectively. We do see a few 0s in these variables though all the cars have been insured telling us that some cars have never claimed for the Insurance.

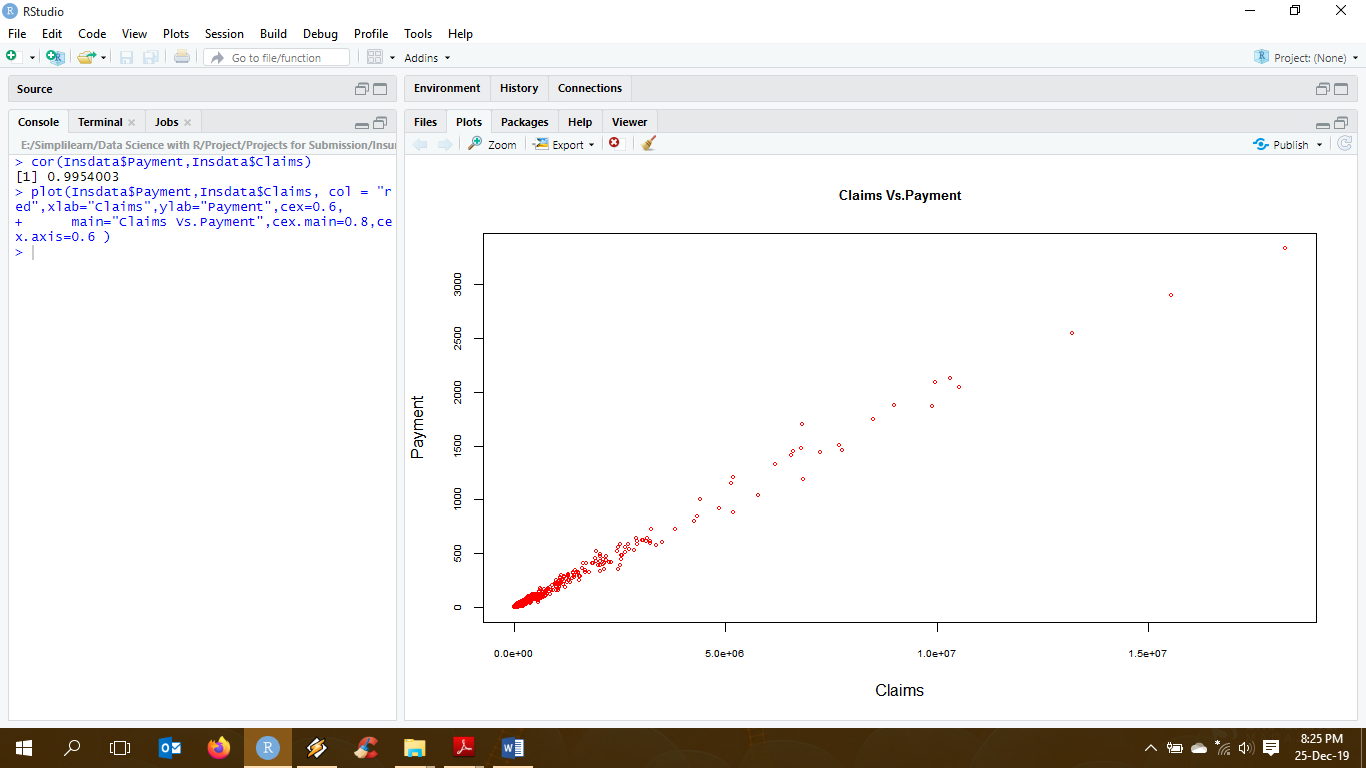
1. **The total value of payment by an insurance company is an important factor to be monitored. So, the committee has decided to find whether this payment is related to number of claims and the number of insured policy years. They also want to visualize the results for better understanding.**

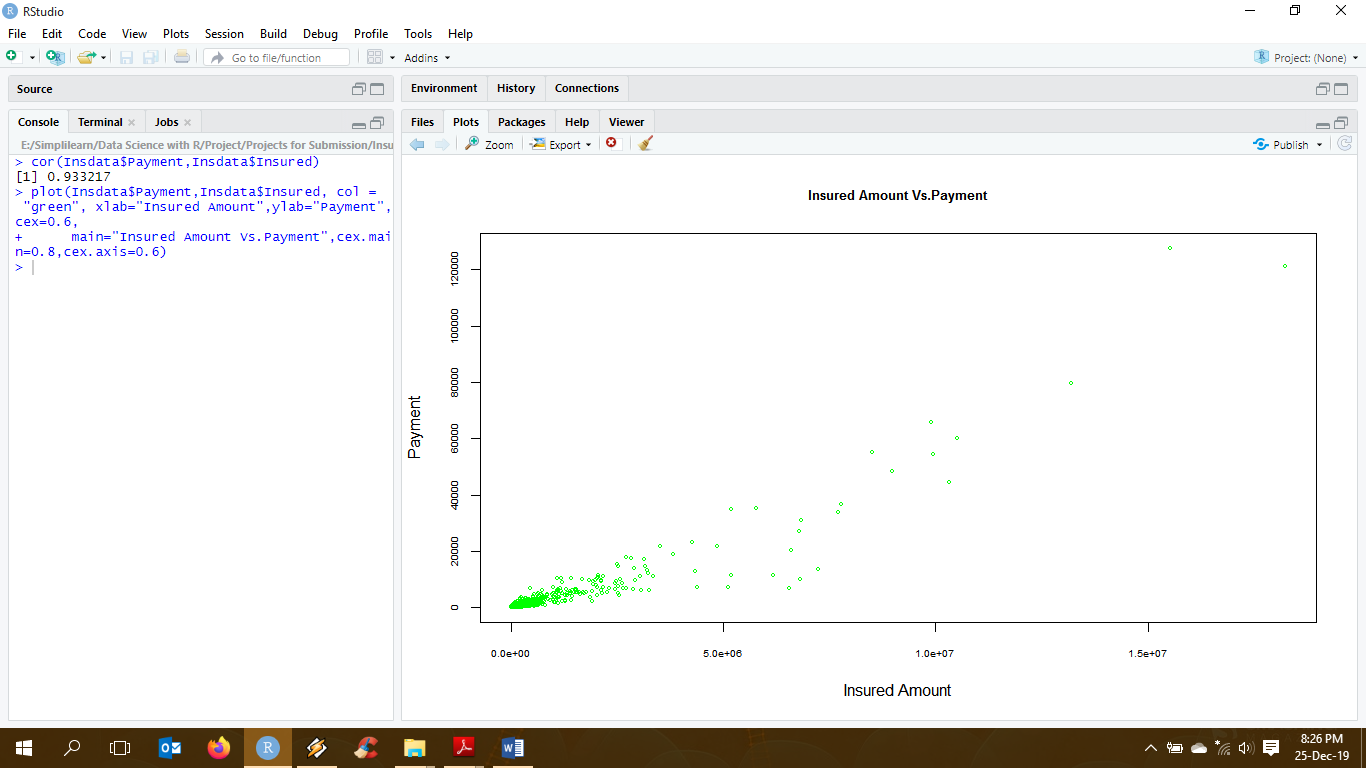
In order to check the dependencies of the variables to each other (payment with claims and payment with number of policy years) we can either create a linear regression model or check the correlation between the variables.



**Interpretation:**

Here, p-value for both claims and number of years insured is way less than the 0.05 we assume while creating any linear model signifying that payments is positively dependent on both the variables.



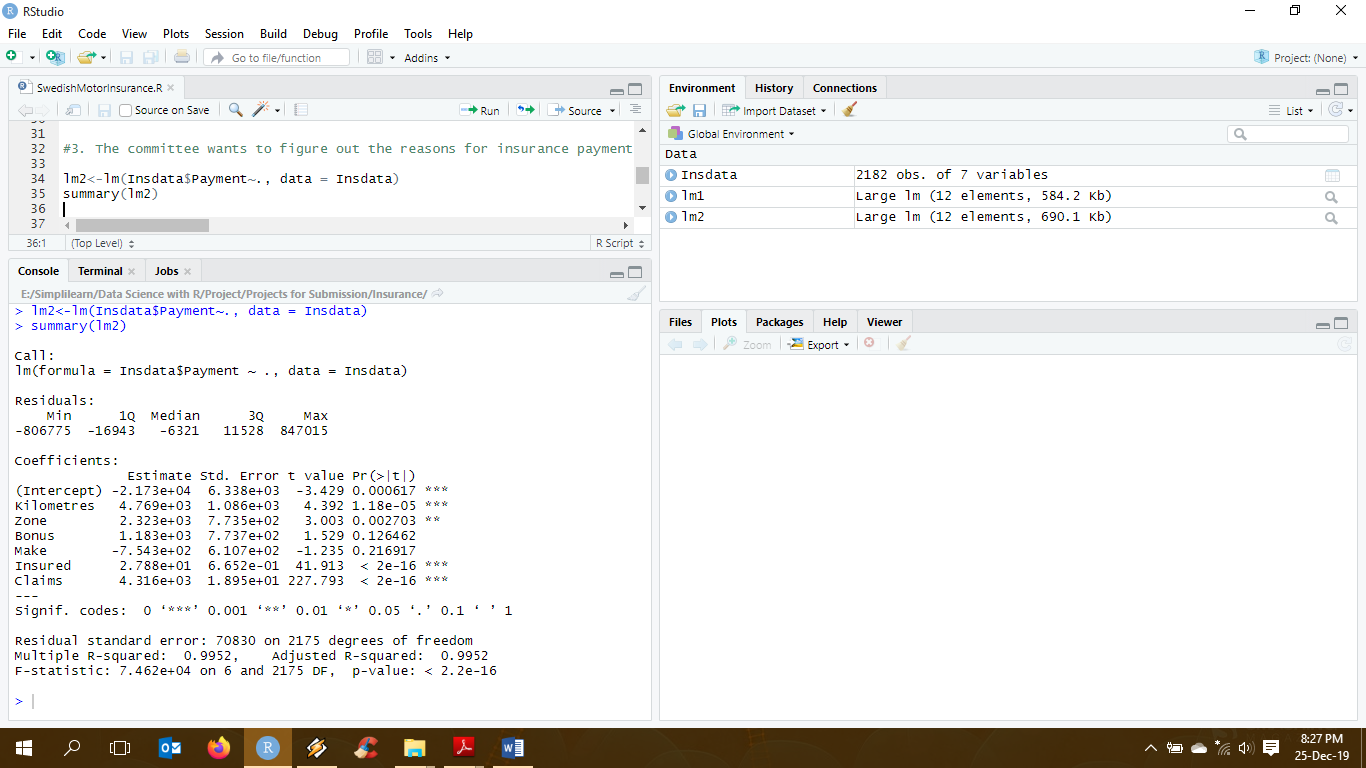


**Interpretation:**

As seen in the plots above, both variables, claims and number of years insured form an almost positive linear correlation curve with payment signifying a direct dependencies between the variables.

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.**

To check the dependencies of distance, location, bonus, make, insured amount and claims on payment we create a linear regression model.



**Interpretation:**

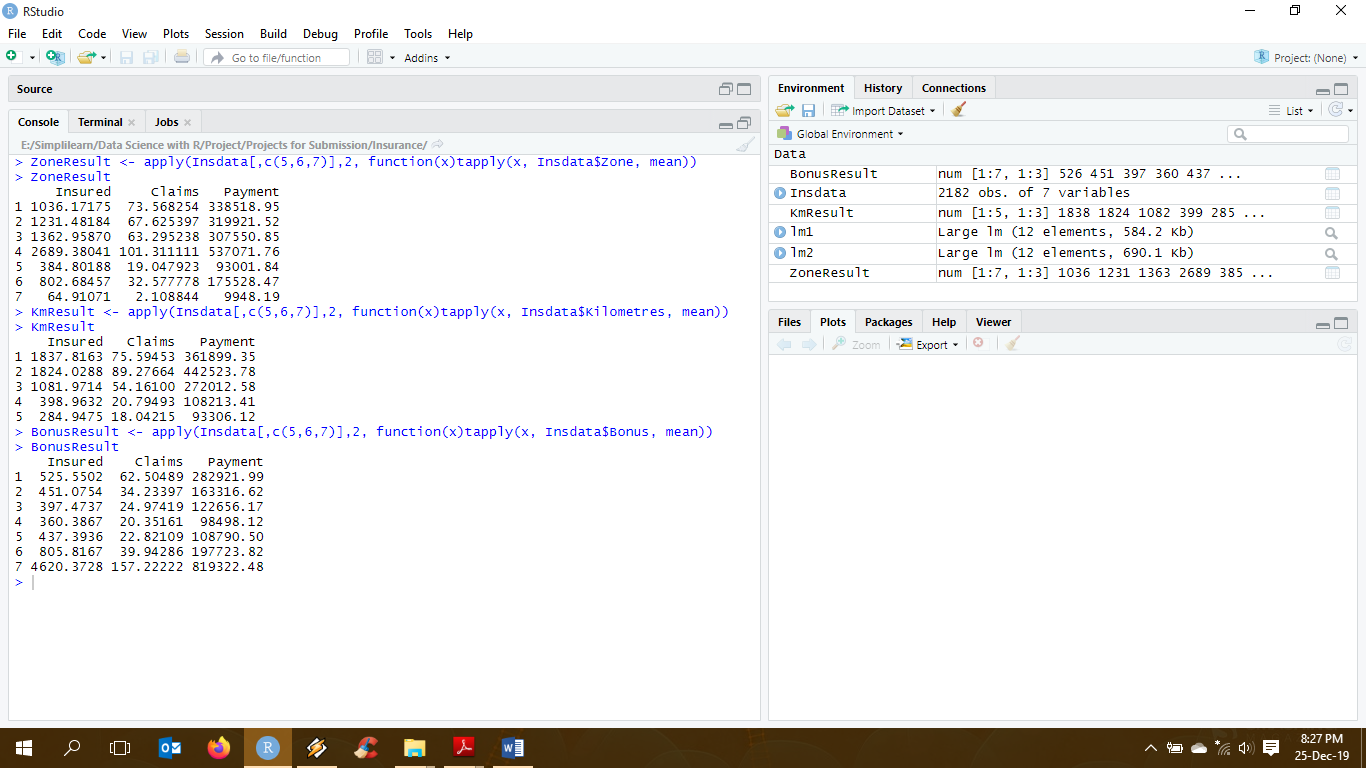
As we know, when p-value is less than 0.05, which we assume while creating a linear regression model, we consider that variable to be significant in the model.

Here, we can see that Bonus and Make have very high p-values of 0.13 and 0.22. Hence, we can say that they are not significant variables in this data. This means that these variables do not make much difference to the payment.

However, Kilometers (distance), Zone (region), number of years insured and claims have significantly less p-value and contribute or effect payment.

1. **The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased.**

To find at what level the claims and payment increases, we need to find the categorical values of location, kilometer and bonus.



**Interpretation:**

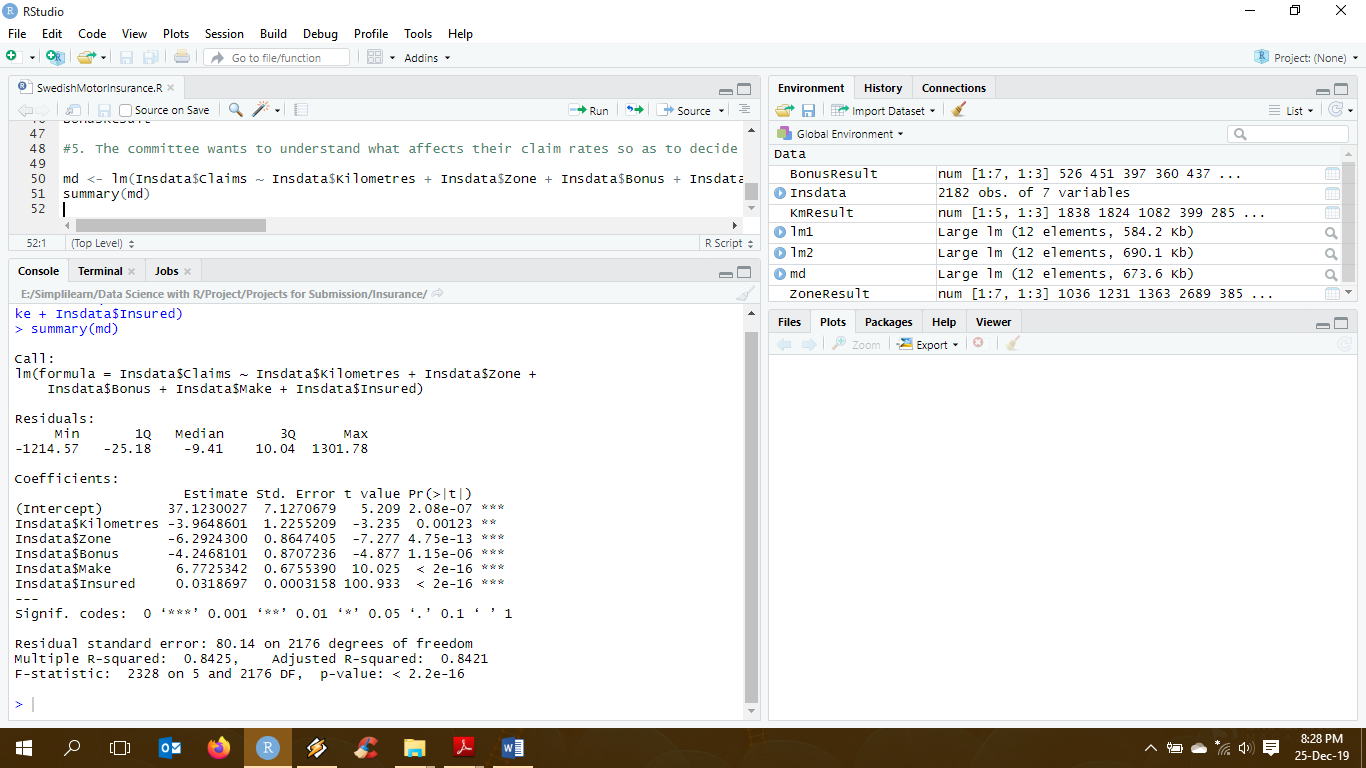
From the above, we can see that Zone 4 has the maximum number of insured years, claims and thus payment whereas Zone 7 has the least number of insured years and claims making payments in the zone the least. Zone 1-4 have the maximum number of insured years, claims and corresponding payments.

Kilometer group 5 has the least number of insured years, claims and payment while group 1 has the maximum number of insured years, claims and thus payments.

Bonus group 7 has a radically high number of insured years, claims and subsequent payments as compared to other groups.

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, kilo meter, bonus, or make affects the claim rates and to what extent.**

To understand what variable affects the claim rates and the extent of it we create a linear regression model.



**Interpretation:**

As we know, when p-value is less than 0.05, which we assume while creating a linear regression model, we consider that variable to be significant in the model.

Here, we can see that all independent variables have p-values less than 0.05 and are quite significant and thus, have a strong influence on claims.

**Programming Codes:**

#Reading Insurance Data

Insdata <- read.csv("SwedishMotorInsurance.csv")

head(Insdata)

#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.

summary(Insdata)

#2. 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.

#Approach 1 - liner regression model

lm1<-lm(Insdata$Payment~Insdata$Claims+Insdata$Insured)

summary(lm1)

#Approach 2 - Find correlation between variables

cor(Insdata$Payment,Insdata$Claims)

plot(Insdata$Payment,Insdata$Claims, col = "red",xlab="Claims",ylab="Payment",cex=0.6,

main="Claims Vs.Payment",cex.main=0.8,cex.axis=0.6 )

cor(Insdata$Payment,Insdata$Insured)

plot(Insdata$Payment,Insdata$Insured, col = "green", xlab="Insured Amount",ylab="Payment",cex=0.6,

main="Insured Amount Vs.Payment",cex.main=0.8,cex.axis=0.6)

#3. 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.

lm2<-lm(Insdata$Payment~., data = Insdata)

summary(lm2)

#4. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased. (Hint: Aggregate Dataset)

ZoneResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Zone, mean))

ZoneResult

KmResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Kilometres, mean))

KmResult

BonusResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Bonus, mean))

BonusResult

#5. 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, kilo meter, bonus, or make affects the claim rates and to what extent.

md <- lm(Insdata$Claims ~ Insdata$Kilometres + Insdata$Zone + Insdata$Bonus + Insdata$Make + Insdata$Insured)

summary(md)

-------------------------------------------------------------------The End-------------------------------------------------------------------