Insurance factors identification

By Ifalore Simeon

DESCRIPTION

Background and Objective:

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.

Domain: Insurance

Dataset Description:
The insurance dataset holds 7 variables and the description of these variables are given below:

Attribute	Description
Kilometers	Kilometers travelled per year 1: < 1000 2: 1000-15000 3: 15000-20000 4: 20000-25000 5: > 25000
Zone	Geographical zone 1: Stockholm, Göteborg, 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.

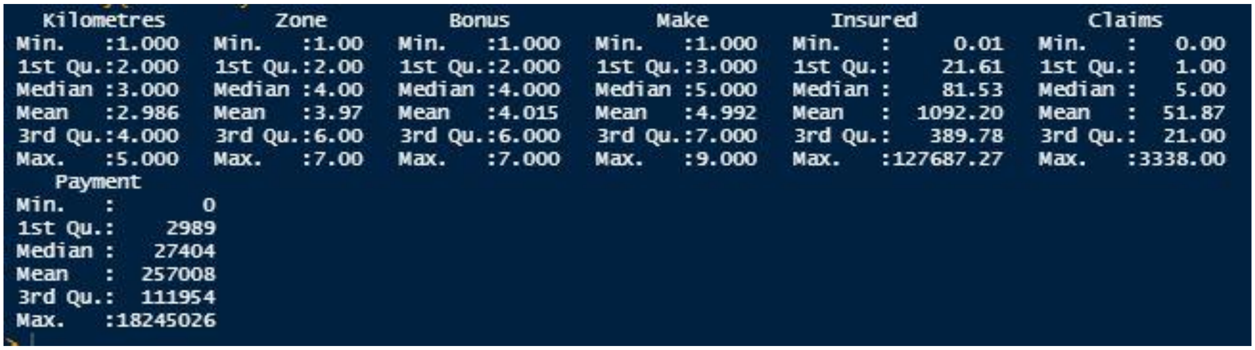
Make	1-8 represents eight different common car models. All other models are combined in class 9.
Insured	The number of insured in policy-years.
Claims	Number of claims
Payment	The total value of payments in Skr (Swedish Krona)

Analysis Tasks: After understanding the data, you need to help the committee with the following by the use of the R tool:

- 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.
- 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 the number of claims and the number of insured policy years. They also want to visualize the results for better understanding.
- 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.
- The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometre, and bonus level their insured amount, claims, and payment gets increased.
- 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, kilometre, bonus, or make affects the claim rates and to what extent.

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.

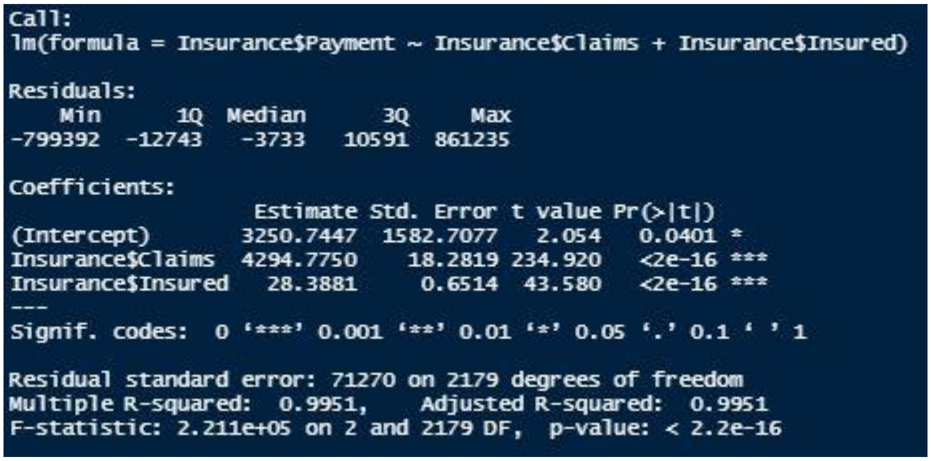
```
library(readr)

Insurance<- read_csv("C:/Users/Simeon/Desktop/Insurance data.csv")

summary(Insurance)
```



```
  Kilometres        Zone           Bonus            Make          Insured              Claims
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :     0.01   Min.   :   0.00
 1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:    21.61   1st Qu.:   1.00
 Median :3.000   Median :4.00   Median :4.000   Median :5.000   Median :    81.53   Median :   5.00
 Mean   :2.986   Mean   :3.97   Mean   :4.015   Mean   :4.992   Mean   :  1092.20   Mean   :  51.87
 3rd Qu.:4.000   3rd Qu.:6.00   3rd Qu.:6.000   3rd Qu.:7.000   3rd Qu.:   389.78   3rd Qu.:  21.00
 Max.   :5.000   Max.   :7.00   Max.   :7.000   Max.   :9.000   Max.   :127687.27   Max.   :3338.00
    Payment
 Min.   :       0
 1st Qu.:    2989
 Median :   27404
 Mean   :  257008
 3rd Qu.:  111954
 Max.   :18245026
```

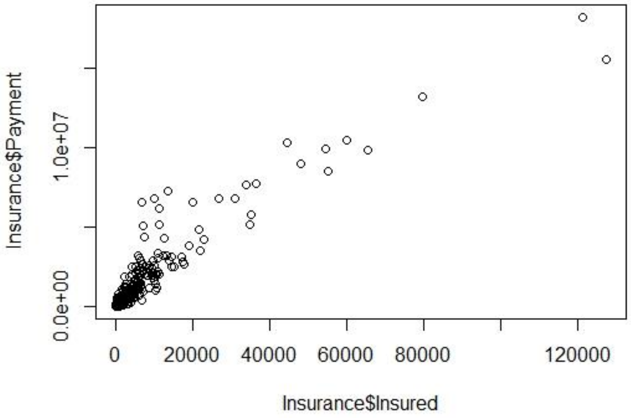
2. **The total value of payment by an insurance company is an important factor to be monitored. So, the committee has decided to find whether this payment is related to the number of claims and the number of insured policy years. They also want to visualize the results for better understanding.**

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

summary(lm1)
```

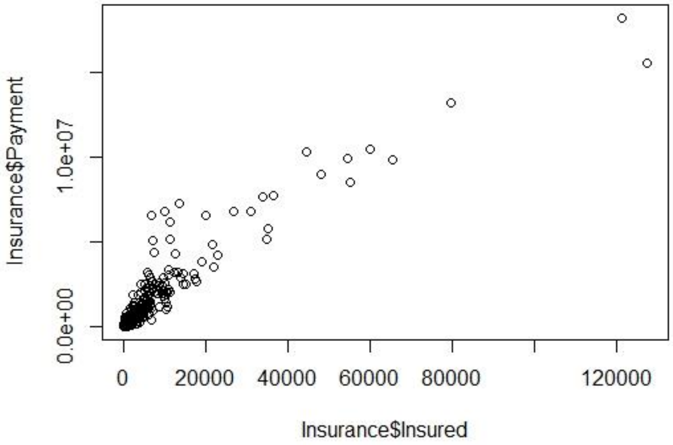


```
Call:
lm(formula = Insurance$Payment ~ Insurance$Claims + Insurance$Insured)

Residuals:
    Min      1Q  Median      3Q     Max 
-799392  -12743   -3733   10591  861235 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3250.7447  1582.7077   2.054   0.0401 *  
Insurance$Claims  4294.7750    18.2819 234.920   <2e-16 ***
Insurance$Insured   28.3881     0.6514  43.580   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71270 on 2179 degrees of freedom
Multiple R-squared:  0.9951,	Adjusted R-squared:  0.9951 
F-statistic: 2.211e+05 on 2 and 2179 DF,  p-value: < 2.2e-16
```

```
plot(Insurance$Claims,Insurance$Payment)
```

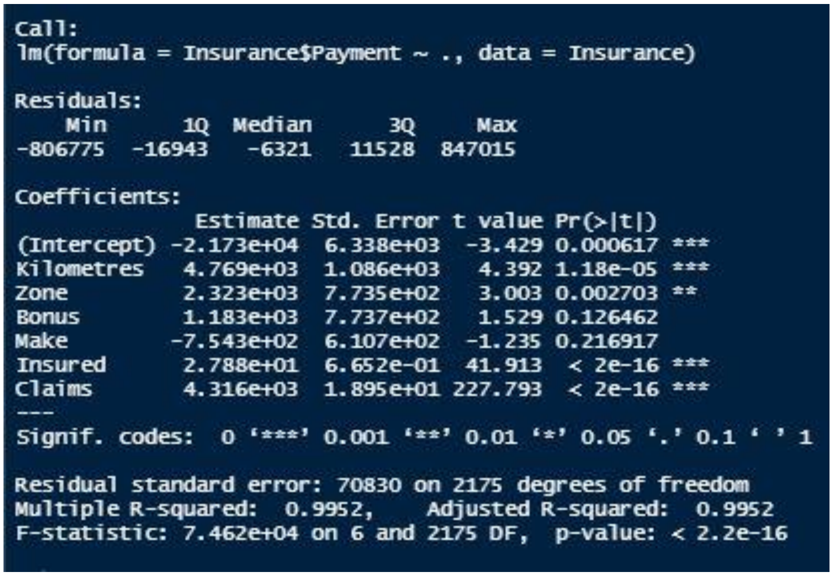


```
plot(Insurance$Insured,Insurance$Payment)
```



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(Insurance$Payment~.,data=Insurance)
summary(lm2)
```

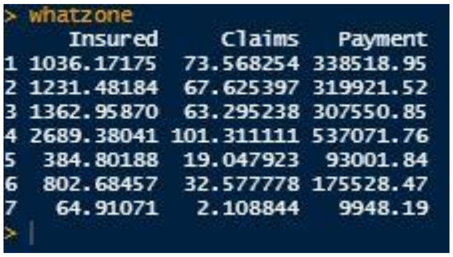


```
Call:
lm(formula = Insurance$Payment ~ ., data = Insurance)

Residuals:
    Min      1Q  Median      3Q     Max 
-806775  -16943   -6321   11528  847015 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.173e+04  6.338e+03  -3.429 0.000617 ***
Kilometres   4.769e+03  1.086e+03   4.392 1.18e-05 ***
Zone         2.323e+03  7.735e+02   3.003 0.002703 ** 
Bonus        1.183e+03  7.737e+02   1.529 0.126462    
Make        -7.543e+02  6.107e+02  -1.235 0.216917    
Insured      2.788e+01  6.652e-01  41.913  < 2e-16 ***
Claims       4.316e+03  1.895e+01 227.793  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70830 on 2175 degrees of freedom
Multiple R-squared:  0.9952,	Adjusted R-squared:  0.9952 
F-statistic: 7.462e+04 on 6 and 2175 DF,  p-value: < 2.2e-16
```

** We can see that all factors except Bonus and Make are affecting the payment significantly

4. **The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometre, and bonus level their insured amount, claims, and payment gets increased.**

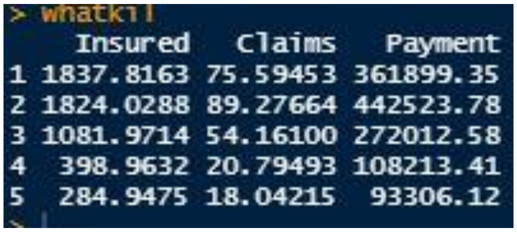
```
whatzone<-apply(Insurance[,c(5,6,7)], 2, function(x) tapply(x, Insurance$Zone, mean))

whatzone
```



```
> whatzone
     Insured     Claims   Payment
1 1036.17175  73.568254 338518.95
2 1231.48184  67.625397 319921.52
3 1362.95870  63.295238 307550.85
4 2689.38041 101.311111 537071.76
5  384.80188  19.047923  93001.84
6  802.68457  32.577778 175528.47
7   64.91071   2.108844   9948.19
> |
```

***Zone 4 has the highest number of claims, and thus payment as well. Zones 1-4 have more insured years, claims, and payments.*

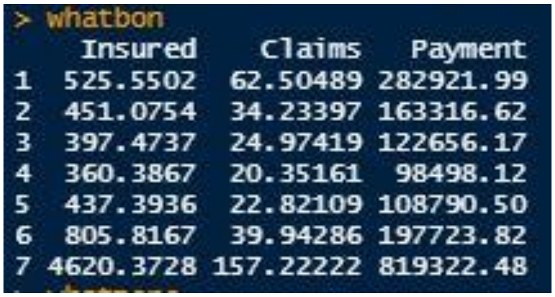
```
whatkil<-apply(Insurance[,c(5,6,7)],2,function(x)tapply(x,Insurance$Kilometres,mean))

whatkil
```



```
> whatkil
     Insured   Claims   Payment
1 1837.8163 75.59453 361899.35
2 1824.0288 89.27664 442523.78
3 1081.9714 54.16100 272012.58
4  398.9632 20.79493 108213.41
5  284.9475 18.04215  93306.12
```

*** Kilometer group 2 has the maximum payments. Though the insured number of years is lesser than kilometre 1, the claims and payments are higher for group 2*

```
whatbon<-apply(Insurance[,c(5,6,7)],2,function(x)tapply(x,Insurance$Bonus,mean))

whatbon
```

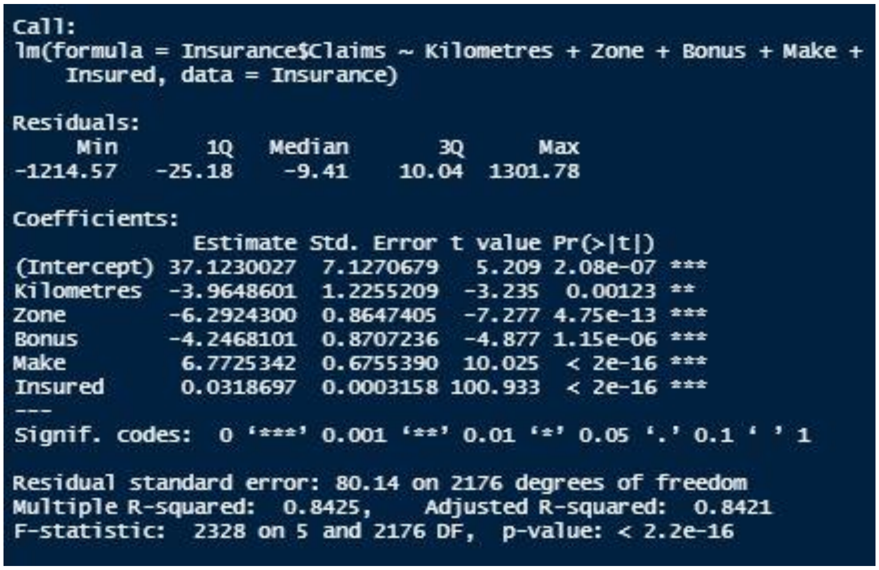


```
> whatbon
    Insured    Claims   Payment
1  525.5502  62.50489 282921.99
2  451.0754  34.23397 163316.62
3  397.4737  24.97419 122656.17
4  360.3867  20.35161  98498.12
5  437.3936  22.82109 108790.50
6  805.8167  39.94286 197723.82
7 4620.3728 157.22222 819322.48
```

*** Bonus group 7 has the maximum payments, Insured number and Claims. This is followed by Bonus group 1 which although this group has a lower Insured when compared to Bonus group 6*

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

```
ClaimMod<-lm(Insurance$Claims~Kilometres+Zone+Bonus+Make+Insured,data=Insurance)

summary(ClaimMod)
```



```
Call:
lm(formula = Insurance$Claims ~ Kilometres + Zone + Bonus + Make +
    Insured, data = Insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-1214.57   -25.18    -9.41    10.04  1301.78

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.1230027  7.1270679   5.209 2.08e-07 ***
Kilometres  -3.9648601  1.2255209  -3.235  0.00123 **
Zone        -6.2924300  0.8647405  -7.277 4.75e-13 ***
Bonus       -4.2468101  0.8707236  -4.877 1.15e-06 ***
Make         6.7725342  0.6755390  10.025  < 2e-16 ***
Insured      0.0318697  0.0003158 100.933  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.14 on 2176 degrees of freedom
Multiple R-squared:  0.8425,    Adjusted R-squared:  0.8421
F-statistic:  2328 on 5 and 2176 DF,  p-value: < 2.2e-16
```

***The results shows that all the p values of independent variables, such as kilometres, zone, bonus, make, and insured are highly significant and are making an impact on the claims.*