R Project

Daniel Kakon

Introduction

We were asked to examine the relationship between a particular subject that is the explained variable and a number of explanatory variables.
We chose the subject of road accidents (the explained variable) and chose 5 explanatory variables:

1.Sex (1-male/2-female)
2.Age
3.Seniority of the driver
4.Year of Automotive Manufacturing
5.Number of hours of sleep on average of the driver (5-8)

Taken from a sample of 60 people.
We received the sample through a survey by Google.

Our goal in the project is to examine the linear relationship between the main subject and the explained variables, to perform an analysis of the data we received. Build a regression model and test the quality of the model using outputs of the R language in the software assistant "RStudio".

In fact, in our project, we will examine whether our explanatory variables (gender, age, seniority of driving, vehicle manufacturing year and average number of hours of sleep) and how they affect and whether they are related to the number of road accidents that the driver has committed.

Theorical statistics

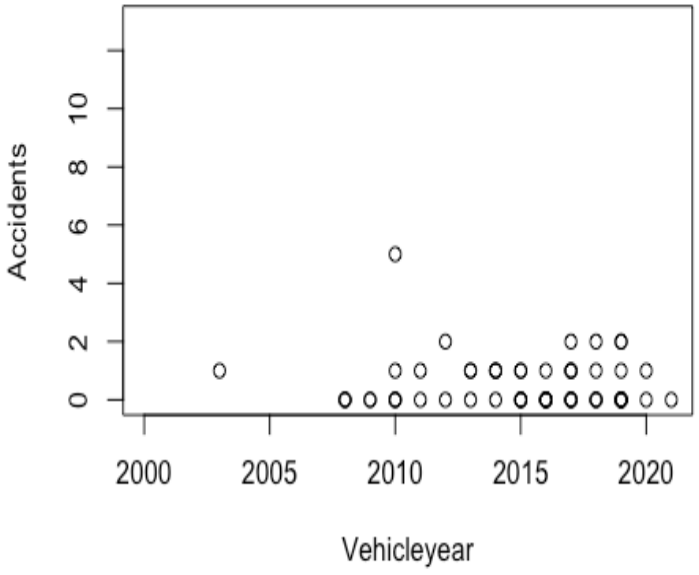
In the above table we will see the variables that we will use during the project, we will also be able to show each variable a number of parameters that are: the minimum value, the maximum value, the average and the median.

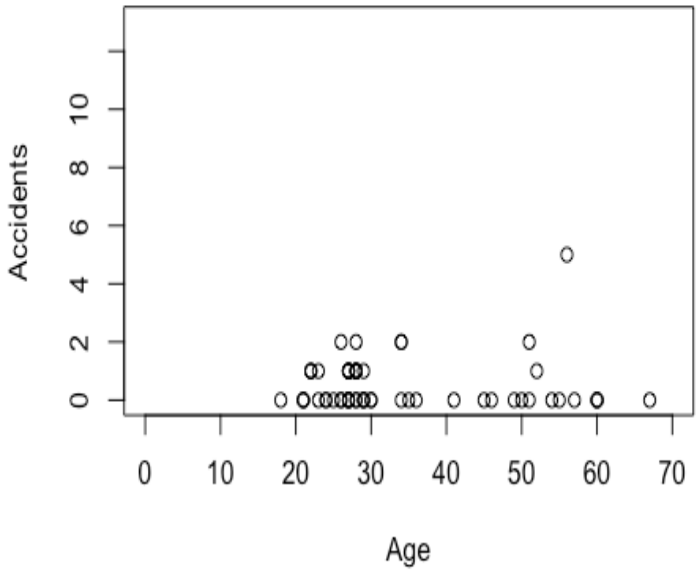
```
  MyData <- read.csv("/Users/danielkakon/Desktop/projectR/RRR2.csv", header =
TRUE , sep = "," )
summary(MyData)

##      Gender          Age           hourssleep    Drivingexperience  Vehicley
ear
##  Min.   :1.0   Min.   :18.00   Min.   :5.000   Min.   : 1.00     Min.   :2
003
##  1st Qu.:1.0   1st Qu.:26.75   1st Qu.:7.000   1st Qu.: 7.00     1st Qu.:2
013
##  Median :1.5   Median :28.00   Median :7.000   Median :10.00     Median :2
016
##  Mean   :1.5   Mean   :34.02   Mean   :6.983   Mean   :14.85     Mean   :2
015
##  3rd Qu.:2.0   3rd Qu.:42.00   3rd Qu.:7.250   3rd Qu.:24.25     3rd Qu.:2
018
##  Max.   :2.0   Max.   :67.00   Max.   :8.000   Max.   :45.00     Max.   :2
021
##    Accidents
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.5333
##  3rd Qu.:1.0000
##  Max.   :5.0000
```

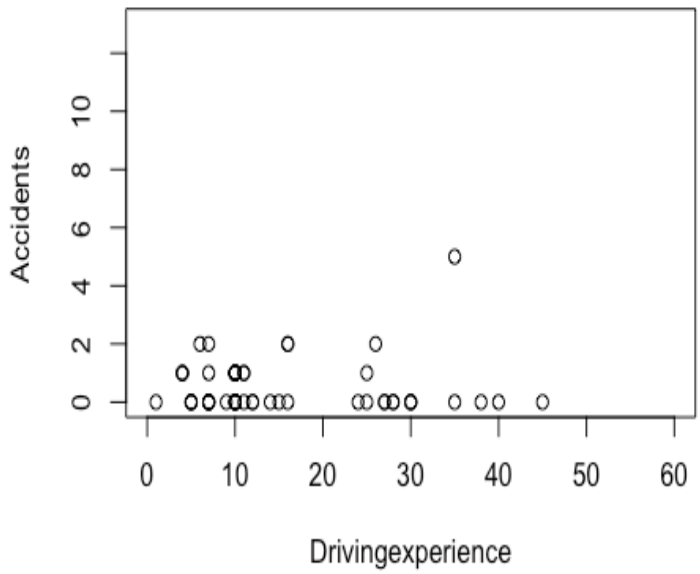
Plot charts

Plot charts to visually understand the data.

```
plot(MyData$Drivingexperience, MyData$Accidents ,
     xlab = 'Drivingexperience' , ylab = 'Accidents', xlim = c(0,60),ylim = c
(0,13))

  plot(MyData$Age, MyData$Accidents ,
       xlab = 'Age' , ylab = 'Accidents', xlim = c(0,70),ylim = c(0,13))

  plot(MyData$Vehicleyear, MyData$Accidents ,
       xlab = 'Vehicleyear' , ylab = 'Accidents', xlim = c(2000,2021),ylim =
c(0,13))
```







Dependence between variables

We will examine whether there is a dependence between the explanatory variables with the help of a matrix.
It seems that there is a dependence only between the seniority of driving and age, for the rest of the variables there does not seem to be a dependence between them.

Blue - marks a positive correlation.
Red - marks a negative correlation

```
corrplot(cor(MyData),method = "number")
```

Adjust the model

We will now see if there is a linear relationship between the explanatory variables and the explained variable.
We'll use the multiple regression model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

```
LinearModel <- lm(formula = MyData[['Accidents']] ~ MyData[['Gender']]+MyData
[['Age']]+MyData[['hourssleep']]+MyData[['Drivingexperience']]+MyData[['Vehic
leyear']])
summary(LinearModel)

##
## Call:
## lm(formula = MyData[["Accidents"]] ~ MyData[["Gender"]] + MyData[["Age"]]
+
##     MyData[["hourssleep"]] + MyData[["Drivingexperience"]] +
##     MyData[["Vehicleyear"]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0643 -0.5295 -0.2504  0.4518  3.9082
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   12.743726  63.552685   0.201   0.8418
## MyData[["Gender"]]            -0.505802   0.232781  -2.173   0.0342 *
## MyData[["Age"]]                0.028710   0.044030   0.652   0.5171
## MyData[["hourssleep"]]         0.263952   0.162045   1.629   0.1092
## MyData[["Drivingexperience"]] -0.031155   0.052807  -0.590   0.5577
## MyData[["Vehicleyear"]]       -0.006853   0.031429  -0.218   0.8282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8642 on 54 degrees of freedom
## Multiple R-squared:  0.1025, Adjusted R-squared:  0.01939
## F-statistic: 1.233 on 5 and 54 DF,  p-value: 0.3064
```

Model equation:

$$Y_i = 12.743726 - 0.505802 X_{1i} + 0.028710 X_{2i} + 0.263952 X_{3i} - 0.031155 X_{4i} - 0.006853 X_{5i}$$

According to the output we received, we will conclude that:

X1- We have defined the sex variable by 1 or 2 and therefore the number of accidents will be reduced by 0.505802 depending on the variable.

X2-If we increase the age variable, the number of accidents will increase by 0.02871

X3-If we raise the number of hours of sleep variable, the number of accidents will increase

by 0.263952.

x_4-If we raise the seniority of the driver, the number of accidents will decrease by 0.031155

x_5-If we increase the year of the vehicle manufacturing, the number of accidents will decrease by 0.006853

We used the analysis of the regression model; it seems that the dimension of the quality of the model adjustment is very low (R_{adj}).

(0.01939), it is evident that all the explanatory variables do not have a significant effect on the explanatory variable (number of accidents) except for a "gender" variable that has a significant effect on the explained variable.

ANOVA Test

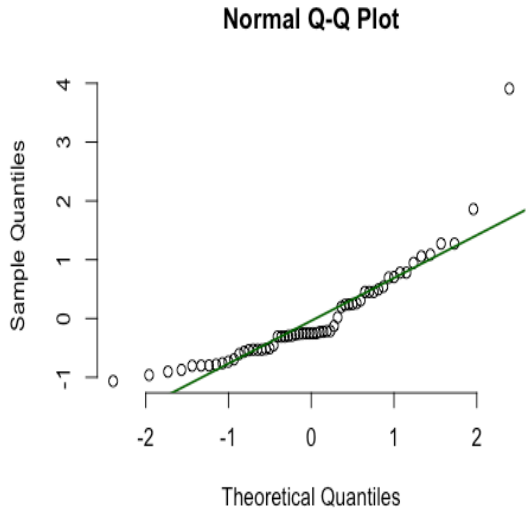
```
anova(LinearModel)

## Analysis of Variance Table
##
## Response: MyData[["Accidents"]]
##                                Df Sum Sq Mean Sq F value  Pr(>F)
## MyData[["Gender"]]              1  2.400 2.40000  3.2137 0.07863 .
## MyData[["Age"]]                 1  0.002 0.00250  0.0033 0.95410
## MyData[["hourssleep"]]          1  1.870 1.87021  2.5043 0.11938
## MyData[["Drivingexperience"]]   1  0.297 0.29723  0.3980 0.53079
## MyData[["Vehicleyear"]]         1  0.036 0.03551  0.0475 0.82821
## Residuals                      54 40.328 0.74681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can also see in the ANOVA test that for the F test carried out using differences analysis you can see that the value of the P-value is small compared to the statistics which shows the rejection of the H_0 that the observations have different differences between the groups.

Check the normality assumption of errors

```
qqnorm(LinearModel$residuals , pch = 1, frame = FALSE )
qqline(LinearModel$residuals, col = 'dark green', lwd = 1.7)
```

We will perform this test with the help of QQ-plot with the help of a KS test. We can see that some of the points are close to the line that represents the normal distribution, and some of the points are not around the line which can show a lack of normal distribution and therefore to verify this, we will use the Kolmgorov-Smirnov test.



KS test: Low P-value so we conclude that the model does not have a normal distribution and we will get H_0

```
 ks.test(x = LinearModel$residuals , y = "pnorm" , alternative =  "two.sided"
, exact = NULL)
```

```
## Warning in ks.test(x = LinearModel$residuals, y = "pnorm", alternative =
## "two.sided", : ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  LinearModel$residuals
## D = 0.18471, p-value = 0.03335
## alternative hypothesis: two-sided
```

Final model

In the final model we will insert only the distinct variables and we have one distinct variable that is the "gender" variable we will reconcile with R:

```
 linearmodel2 <- lm(formula = MyData[['Accidents']] ~ MyData[['Gender']])
 summary(linearmodel2)
```

```
##
## Call:
## lm(formula = MyData[["Accidents"]] ~ MyData[["Gender"]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7333 -0.7333 -0.3333  0.2667  4.2667
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.1333     0.3496   3.242  0.00197 **
## MyData[["Gender"]]  -0.4000     0.2211  -1.809  0.07562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8563 on 58 degrees of freedom
## Multiple R-squared:  0.05341,    Adjusted R-squared:  0.03709
## F-statistic: 3.273 on 1 and 58 DF,  p-value: 0.07562
```

```
 anova(linearmodel2)
```

```
## Analysis of Variance Table
##
## Response: MyData[["Accidents"]]
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## MyData[["Gender"]]  1  2.400 2.40000  3.2727 0.07562 .
## Residuals          58 42.533 0.73333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can see by the matching quality index and the values we received that there is no significance to the "gender" variable, and it seems that there is a linear connection to the cutter.

Model equation:

$$Y_i = 1.1333 + 0.3496X_{1i}$$

Discussion and conclusions

In the project, we wanted to examine whether there is a connection between the explanatory variables and the explained variable.

After examining the model several times, we saw that the explained variables we chose did not affect the number of accidents and therefore we lowered the explanatory variables that were not clear and were left with the variable "Gender" which in the end was also not clear in the last model.
And so we came to the conclusion that none of the variables we chose affected the number of accidents, perhaps if our sample was larger and with other explanatory variables we would have achieved other results.