UDC: 368.03:303.724
Professional paper

APPLICATION OF REGRESSION ANALYSIS IN R

Lazim KAMBERI[1], Merita BAJRAMI[2], Mirlinda SELIMI[3], Rushadije RAMANI-HALILI[4]

[1*] *Department of Mathematics, Faculty of Natural Science and Mathematics, NMK*
[2] *Department of Mathematics, Faculty of Natural Science and Mathematics, N MK*
[3] *Department of Mathematics, Faculty of Natural Science and Mathematics, N MK*
[4] *Department of Mathematics, Faculty of Natural Science and Mathematics, N MK*
[*]*Corresponding Author: e-mail: lazim.kamberi@unite.edu.mk*

Abstract

Statistics as an independent scientific discipline has become an applicable tool in various fields of science, society, and business such as insurance, etc. Since the activities of an insurance company can be considered as a business process where data management and analysis play a key role, it is the idea why a basic understanding of data-related issues such as quality, analysis, application of linear models, and accurate interpretation of data is essential to the company's insurance business. In this paper, we study the application of statistical models, where the relationship between different categorical variables is analyzed through the regression method. This study, in addition to examining the theoretical part of regression models, is carried out using the programming language R, the use of which is widespread in the field of statistics. The study is based on real data from insurance companies, specifically on the claims of insured persons.

Keywords: Statistical models, Regression method, Variables, Relationship

1. Introduction

Statistics is a science that studies the organization, analysis, and interpretation of quantitative changes in the development of society, economy, culture, etc. collecting numerical data for them, which are grouped and processed with special methods. The knowledge gained from the subject of Statistics and Probability can be applied in practice and at the same time highlight the great importance of the field of Statistics and Probability in solving and approximating problems in almost every scientific and social field.
Developments in computer technology such as programming languages enable us to apply such methods when we have a large number of data. The aim of the paper is also the application of these statistical methods through adequate programming languages, such as the R programming language. So, at the same time, in addition to the scientific theoretical elaboration of the statistical methods, the programming code of these practical methods in the R programming language is also elaborated and studied.

Application of regression analysis in R

Regression analysis is a statistical technique used to model the relationship between a set of data or independent variables with one or more dependent variables. The simplest form of regression analysis is linear regression which is the statistical modeling between a variable X (independent variable) and a variable Y (dependent variable). We will explore this further using the R language.
First, we see how we can observe a connection between two variables without starting with the regression model.

1.1 Correlation: The correlation metric is a measure of the relationship between variables. Two variables may be strongly correlated, but not cause/effect. A metric correlation is called the correlation coefficient, which is usually denoted by r. The most commonly used metric correlations are Pearson's coefficient, Spearman's Rho,

and Kendall's Tau. The latter two are nonparametric correlation coefficients and are based on the range of the data.
Pearson's coefficient is determined through the mathematical formula:

$$r = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

where n is the number of data, S_x and S_y are the standard deviation of x and y respectively, and $\bar{x}$ and $\bar{y}$ are their means. The three metrics correlations mentioned above take values from -1 to 1, where 1 indicates that there is a positive correlation, -1 indicates that there is a negative correlation, and 0 that there is no correlation at all. So, these three correlations measure the strength of the linear relationship between two variables. In the case of positive correlation, the points $(x,)$ and y lie exactly on a line with a positive slope. This is usually not the case with real data. As an example, we are taking real data on the age of people who have reported damages during a year. After grouping the claim value data by age of the persons, we will analyze the correlation between age and value of damages. For better demonstrating this example we are limiting the data by taking it from the age of 30 years and above. We will elaborate on the correlation using R. First, we get the graph of these data, which we have stored in an Excel file.

Table 1. 1. Data according to age and damages

Age	Value

	A	B
1	mosha	vlera
2	30	98654.23
3	31	80935.02
4	32	95413.79
5	33	71585.86
6	34	64290.72
7	35	71495.2
8	36	73900.46
9	37	47594.51
10	38	80091.59
11	39	49765.43
12	40	61033.19
13	41	40711.08
14	42	46254.26
15	43	44428.07
16	44	40219.07
17	45	34069.26
18	46	38819.28

By importing this table into R we have:

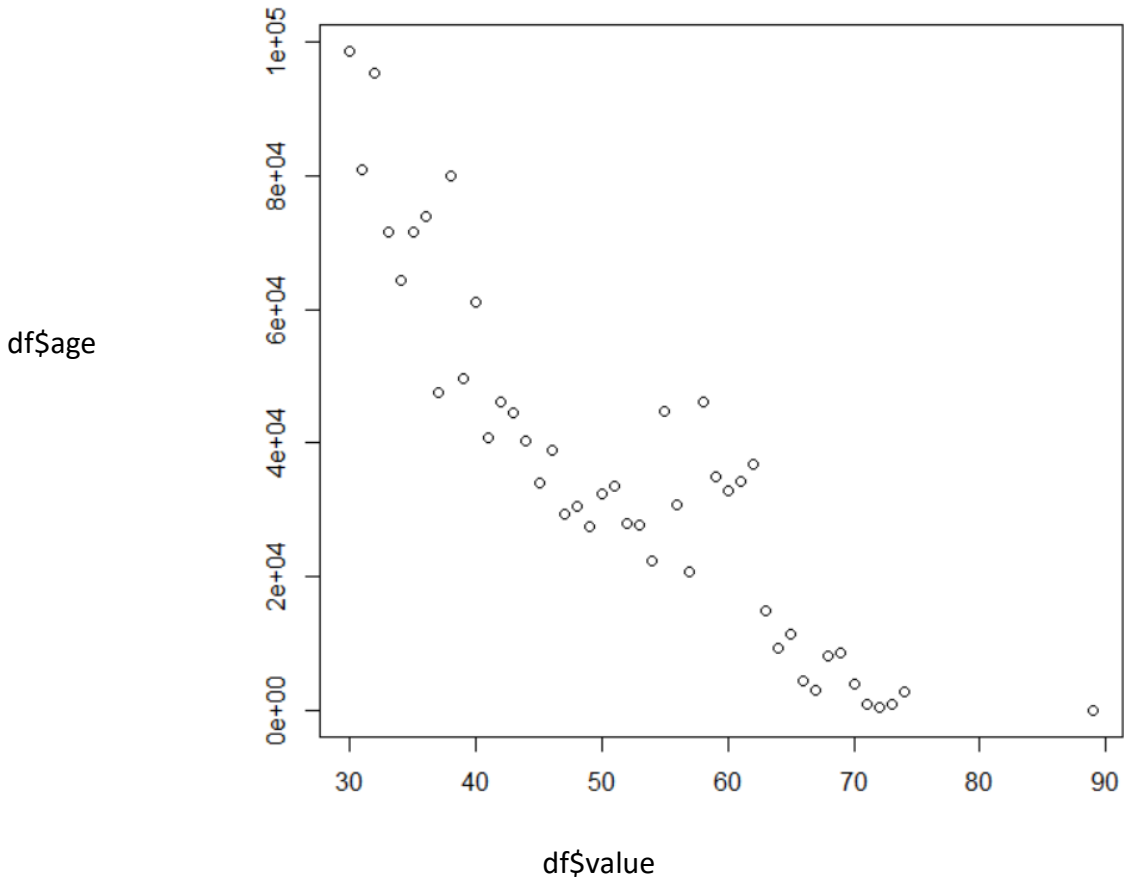
```
> df<-read.csv("C:/Users/vmorina/OneDrive - Skupina Sava Re/Desktop/TM_gusht 2021/lin1.csv")

> mean(df$value)
```

[1] 35221.08

> plot(dfage,dfvalue)

Chart 1.1 Data by age group



Graph 1.1 visually demonstrates that there is a relationship between the variables 'age' and 'value', but does not quantify this. Let us now measure the strength of the linear relationship between the variables 'age' and 'value' using the correlation coefficient. To do this we use the function 'cor. test' in R. The test automatically calculates the Pearson correlation coefficient, but other coefficients can be specified using the optional method parameter.

> cor.test(df$mosha,df$vlera)

Pearson's product-moment correlation

data: df$age and df$value

t = -13.506, df = 44, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9423433 -0.8212967

sample estimates:

cor

-0.8975905

> cor.test(dfage,dfvalue,method="spearman")

Spearman's rank correlation rho

data: df$age and df$value

S = 30926, p-value < 2.2e-16

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

-0.9072464

This correlation -0.90 or 0.91 depending on which measurement we use is quite strong and shows that between these two variables 'age' (variable x) and 'value' (variable y) there is a strong relationship, but from the minus sign we understand that we have a negative correlation between variables. What we miss here is that we cannot predict the new value of the variable Y given the new value of X. To do this we need to use simple linear regression.

1.2. Regression model: Simple linear regression is a statistical model between the variables X and Y in the real world. To study the relationship between X and Y, the simplest of them is the straight line, compared to more complex relationships such as the polynomial. Placing X with Y as in graph 1.1. we take an appropriate step to see if the linear model is appropriate.
To fit a regression model, several assumptions must be made:

1. For any given value of X, the true mean value of Y depends on X, which we denote by $\mu_{y|x}$. In egression, the line represents the mean values of Y and not individual values.

2. Each observation Y_i is independent from other observations, so $Y_j \neq Y_i$.
3. We assume linearity between X and the mean of Y. The mean of Y is defined by the straight line ratio which can be written as:$\mu_{y|x} = \beta_0 + \beta_1 x$, where betas are the free coefficient and the slope coefficient of the line.
4. The variance of $Y_i$is constant.
5. Estimates of $Y_i$have normal distribution with constant variance.

After creating the regression model, we are often interested in doing some correlation tests and model analysis to see if the model fits well or if we need to make changes to it. In R, the 'lm' function is simply a function that does not give more details, even though many calculations are done behind it. The 'summary' function in R displays many details that are used to diagnose the fit of the regression. Where do we get this result from:
> summary(geneLM)
Call:
lm(formula = df$value ~ df$age)
Residuals:
Min 1Q Median 3Q Max
-15527 -8299 -3068 8652 25491
Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 123299.7     6744.5   18.28   <2e-16 ***
df$mosha     -1668.0      123.5  -13.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 11670 on 44 degrees of freedom
Multiple R-squared:  0.8057,    Adjusted R-squared:  0.8013
F-statistic: 182.4 on 1 and 44 DF,  p-value: < 2.2e-16
```

In the first part, the model is first presented. The second part is a summary of the residuals. Usually in regression we look at the distribution of the residuals and check the assumption about their normality. The distribution should be zero-centered, so the median value should be close to zero, which in our case is not so.
In many situations, we have more than one variable and we want to study the dependence of one of them on the others. We call these "other" explanatory variables (predictor, regressor), while we call the first variable the explained ("dependent") variable. For this, we use the model of general linear regression, or multiple linear regression.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

In our concrete case, we will model the number of health damages and analyze some variables of individuals who have reported these damages in group insurance of employees of an organization. These variables are their age, gender, and status whether they are employees or their family members.
So our model is of general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Where Y represents the number of claims reported by each insured person
X_1- represents his age
X_2- represents his gender (with 0 we marked it for men and 1 for women)
X_3 - represents the status of whether he is an employee or a family member (1 for an employee and 0 for a family member)
Table 1.2 Model data

Age n.damagesgender gstatus s

	A	B	C	D	E
		ve	gjinia_g	statusi_s	
2	59	4	0	0	
3	34	1	1	0	
4	40	1	1	1	
5	44	2	1	1	
6	43	2	0	1	
7	12	2	1	1	
8	2	1	0	1	
9	55	3	0	0	
10	27	1	1	1	
11	1	1	1	1	
12	32	1	0	0	
13	29	6	1	1	
14	19	2	0	1	
15	13	2	1	1	
16	36	4	0	0	
17	33	3	0	0	
18	28	5	1	1	

The table continues.
Usually, the Poisson function is used to project the number of damages, which we will apply in the regression through R.
How to write the code for this model is shown below:

```
> gam_Analysis<-glm(df$nrdemeve~df$mosha+df$gjinia_g+df$statusi_s,family=poisson)
> summary(gam_Analysis)

Call:
Formula = df$n.damages ~ df$age ~ df$gender  g + df$status  s,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9595  -0.9501  -0.1858   0.5038   3.6895

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.180409   0.159622  -1.130  0.25838
df$mosha      0.021335   0.002878   7.414 1.23e-13 ***
df$gjinia_g   0.360987   0.113293   3.186  0.00144 **
df$statusi_s -0.105560   0.108386  -0.974  0.33009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 289.96  on 196  degrees of freedom
Residual deviance: 213.49  on 193  degrees of freedom
AIC: 656.21

Number of Fisher Scoring iterations: 5
```

From the result we see that the age and gender variables are significant and have an impact on the model and the empirical regression equation has the form:

$$Y = -0.18 + 0.02X_1 + 0.36X_2 - 0.1X_3$$

From this model we can conclude that:

1. Keeping the effect of gender and status (family or employee) constant, for an increase of 1 year in the age of persons, damages increase by 0.02 on average
2. Keeping the effect of age and status (family or employee) constant, women increase the value of damages by 0.36 euros on average
3. Keeping the effect of age and gender constant, employees reduce the value of damages by 0.1 euro on average.

Conclusion

The knowledge gained from the subject of Statistics and Probability can be applied in practice and at the same time highlight the great importance of the field of Statistics and Probability in solving and approximating problems in almost every scientific and social field.

The essence of this paper is to apply known statistical methods in practice. So from the real data of an insurance company, specifically the data on the damages of the insured persons, by applying known statistical methods such as different statistical distributions as well as linear regression and general regression methods, we conclude their average, variances as well and dependence between variables. Developments in computer technology such as programming languages enable us to apply such methods when we have a large number of data. The aim of the study is also the application of these statistical methods through adequate programming languages for this purpose, such as the R programming language. So, at the same time, in addition to the theoretical scientific elaboration of the statistical methods, the programming code of these practical methods in the R programming language is also elaborated and studied.

The form of the distributions and the construction of their graph have been treated, to visualize the results as best as possible. We have seen the linear relationship between the variables of health damages and the age of the insured. From these data and the model, we found that there is a strong relationship between these two variables, where the increase in age affects the increase in the value of damages. It was found in this model that age and gender have an influence, where the older age has an impact on the increase of damages and the female gender has an impact on the increase of damages.

References

[1]. Kaas, R., Goovaerts, M., & Dhaene, J. (2001). M. Denuit, “Modern Actuarial Risk Theory”.

[2]. Berenson, M., Levine, D., Watson, J., Jayne, N., & O'Brien, M. (2012). “Business Statistics: Concepts and Applications”.

[3]. Sahoo, P. (2013). “Probability and mathematical statistics. University of Louisville”.

[4]. Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). “Loss models: from data to decisions” (Vol. 715). John Wiley & Sons.

[5]. JourioTuinala, Dario Greco, “ Basic Statistics Using R”

[6]. Vitoo Ricci, “ An Introduction to R: “Example for Actuaries”

[7]. Maindonald, J. H. (2008). “Using Rfor Data Analysis and Graphics”.