Multiple Regression Analysis in Teaching Statistics by R Software*

Krasimira Kostadinova

Multiple Regression Analysis in Teaching Statistics by R Software: *This paper presents the necessity to introduce the students majoring in 'Computer Systems and Technologies' or 'Informatics' in the applications of the software R to Statistics. This programming model develops student's memory and helps them to better understand the theory.*

Key words: *Regression Analysis, Statistics, R software.*

INTRODUCTION

Students taking a Statistics class are required to understand, apply, and use a large volume of statistical formulas and methods. Aquiring such large volume of quantitative knowledge will not be efficient enough without illustrating the appropriate applications. It requires appropriate software.

Different programming models, for different operation systems (OS), are necessary when teaching the numerical procedures in statistics. Such sorts of software models are MATLAB, SPSS, MS EXCEL, etc. The software R accompanies their functionality entirely and it is absolutely free. Also it can be installed on the most frequently used OS - Windows, Linux and Mac OS. This allows its easier introduction in teaching.

In particular, in this paper we are going to consider Multiple Regression Analysis, realized by the R software.

MAIN RESULTS

First, let us recall the definition of Regression Analysis (RA).

Definition: RA is a method for modeling the functional relationship between a dependent variable, Y and one or more predictor variables (independent variables) X_i. RA is also used to understand which among the predictor variables are related to the dependent variable, and to explore the forms of these relationships. More specifically, RA is a method for investigating the functional relationship among variables or RA can be used to determine relationships between Y and X_i.

It is common for more than one factor to influence an outcome Y. Fitting regression models to data involving two or more predictors X_i is one of the most widely used statistical procedures.

On linear and nonlinear RA by R software - brief description.

Example 1. The 10 statistical data points are observed (Table 1). Let X and Y be economical indicators. We are going to call Y dependent variable and X – predictor variable. At this data we want to develop a regression equation to model the relationship between Y and X and find a 95% prediction interval for Y when the value of X is fixed, for instance $X = 72$.

Table 1

Dependent variable Y	75	115	90	105	90	144	120	220	145	160
Predictor variable X	60	91	70	85	72	87	78	120	90	150

First, we begin by considering the simple linear regression model.

The equation of linear regression has the following form:

$$\hat{y} = a_0 + a_1 x \qquad (1)$$

where

$\hat{y}$ is the theoretical (fitting) value of dependent variable Y;

x is a value of the predictor variable X;

* This work is supported by grant RD-05-333/2010 (RD-07-1028) of Shumen University, Bulgaria

a_i, i = 1, 2 are the coefficients, chosen by the method of the least square, i.e., such that the sum of squared residuals of the linear model (1) is minimal.

We define two vectors

```
> y<-c(75,115,90,105,90,144,120,220,145,160)
> x<-c(60,91,70,85,72,87,78,120,90,150)
```

Next, we draft a plot of the relationship between *Y* and *X* (Figure 1) and a plot of dependence between the standardized residuals from model (1) and the predictor *X* (Figure 2):

```
> plot(x,y)
```

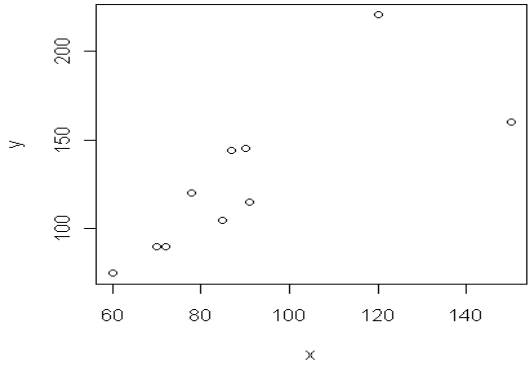

Figure 1

```
> fm1<-lm(y~x); StanRes1<-rstandard(fm1)
> plot(x,StanRes1,ylab="Standartized Residuals")
```

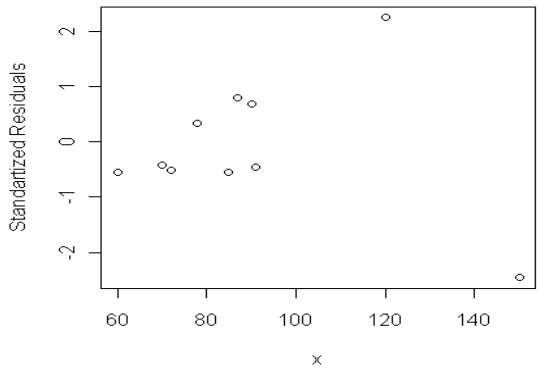

Figure 2

A curved pattern resembling a quadratic shape is clearly evident in Figure 2. Then we begin by considering the nonlinear regression model:

$$\hat{y} = a_0 + a_1 x + a_2 x^2 . \qquad (2)$$

Following the same steps, as above we get:

```
> fm2<-lm(y~I(x^2)+x)
> xnew<-seq(20,110,len=10); dummy<-data.frame(x=xnew)
> lines(x,predict(fm2,newdata=dummy))
> StanRes2<-rstandard(fm2); plot(x,StanRes2,ylab="Standartized Residuals")
```

The random pattern indicates that model (2) is a valid model for the *Y* data.

Figure 3 shows a plot of the dependence between the predictor *X* and the leverage from model (2). It is evident from Figure 3 that the smallest and the largest *x*-values are leverage points:

```
> leverage2<-hatvalues(fm2); plot(x,leverage2); abline(h=0.4,lty=2)
```

Here the height h=0.4 of the horizontal dashed line is calculated by the formula:

$$h = \frac{2(\text{number predictors} + 1)}{\text{number observed variables}}$$

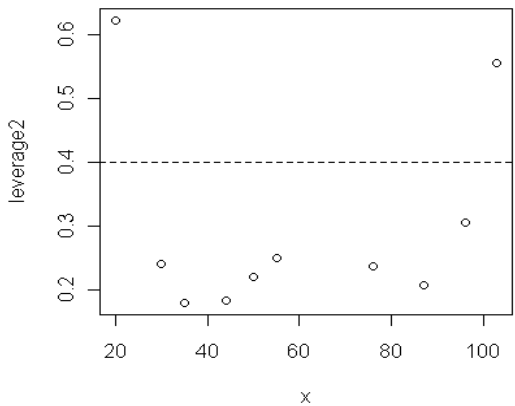

Figure 3

Furthermore, we can see four other summary diagnostic plots, produced by the following function of R:

```
> par(mfrow=c(2,2)); plot(fm2)
```

Once we verify that (2) is a valid regression model, we can use the following command to summarize the results:

```
> summary(fm2)
```

In particular, we get the corresponding equation: $\hat{y} = -244.71 - 0.03x^2 + 6.55x$.

We can also make a prediction for *X*, say *X* = *72*, in the following way:

```
> predict(fm2,newdata=data.frame(x=c(72)),interval="prediction",level=0.95 )
       fit      lwr      upr
1 96.79107 37.82577 155.7564
```

It gives the respective fitting value of the dependent variable $\hat{y}$ and a 95% prediction interval. In our case, for *X* = *72*, the 95% prediction interval is (37.83, 155.76).

We are going to show an advantage of R over to MS Excel. Namely, we change the settings of an example from [3] by adding a nonlinear regression. In the new settings MS Excel can not be used, while R is very efficient.

Example 2 (see [3]). The 10 statistical data points are given in Table 2. Let X_i, *i = 1, 2* and *Y* be economical indicators. We call *Y* dependent variable and X_i, *i = 1, 2* – predictor

variables. At this data we want to determine the parameters in the equation of the linear and nonlinear (polynomial) regression and to make a regression analysis.

Table 2

№ of the observed variable	Dependent variable Y	Predictor X_1	Predictor X_2
1	150	138	400
...	...	...	...
10	63	51	165

In this example, the equation of linear regression has the following form:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2, \quad (3)$$

and the equation of nonlinear (polynomial) regression:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + a_3x_1^2 + a_4x_2^2, \quad (4)$$

Here

$\hat{y}$ is the theoretical (fitting) value of dependent variable Y;

x_i, i = 1, 2 are values of the predictor variables X_i;

a_i, i = 1, 2, 3, 4 are the coefficients, determined by the method of the least square.

Now, we need to introduce our data from Table 2 to determine the coefficients of the corresponding linear and nonlinear (polynomial) regression.

```
> y<-c(150,180,75,123,28,200,136,55,101,63)
> x1<-c(138,170,61,99,20,147,125,40,90,51)
> x2<-c(400,320,153,245,60,550,302,120,205,165)
```

After we introduced the statistical data, we apply the regression models by executing the following sequences:

the simple linear regression model	the nonlinear (polynomial) regression model
> fm1<-lm(y~x1+x2) > summary(fm1)	> fm2<-lm(y~I(x1^2)+I(x2^2)+x1+x2) > summary(fm2)

The results of the simple linear regression model realized by R and MS Excel (see [3]) are analogous. In this case the equation has the following form:

$$\hat{y} = 6.08 + 0.73x_1 + 0.15x_2. \quad (5)$$

From this results (as well as in MS Excel) it follows that the free coefficient a_0 is not statistically significant, and the coefficients a_i, $i = 1, 2$ are statistically significant. Then using the command

```
> fm1<-lm(y~x1+x2+0); summary(fm1)
```

and we obtain the following equation

$$\hat{y} = 0.77x_1 + 0.15x_2, \quad (6)$$

which is statistically significant.

The result of the nonlinear (polynomial) regression model is:

```
Call:
lm(formula = y ~ I(x1^2) + I(x2^2) + x1 + x2)
Residuals:
     1      2      3      4      5      6      7      8      9     10
-8.589  3.431 -0.529  9.171 -4.586  3.081 -3.881  4.052 -5.186  3.035
Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 15.5739159  10.8664907    1.433    0.2112
I(x1^2)     -0.0032035   0.0023399   -1.369    0.2293
I(x2^2)      0.0005434   0.0003259    1.667    0.1563
x1           1.7494796   0.6813738    2.568    0.0502 .
```

```
x2          -0.3108696   0.2762495   -1.125  0.3116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.322 on 5 degrees of freedom
Multiple R-squared: 0.9907,     Adjusted R-squared: 0.9832
F-statistic: 132.6 on 4 and 5 DF,  p-value: 2.931e-05
```

The equation of the regression is the following:

$$\hat{y} = 15.57 - 0.003x_1^2 + 0.0005x_2^2 + 1.75x_1 - 0.31x_2. \quad (7)$$

Note however, that only the coefficient in front of x_1 is statistically significant (see column Pr(>|t|)). Hence the linear model is better than the polynomial. Let us point out again that the analysis of the polynomial regression model (7) can not be done with MS Excel.

Once we verify that the model (6) is acceptable, we can plot several graphs: a plot of the relationship between the predictors x_i, $i = 1, 2$ and the dependent variable y. So we introduce:

```
> plot(x1,y); abline(lsfit(x1,y)); plot(x2,y); abline(lsfit(x2,y))
```

The plot between the predictors x_i, $i = 1, 2$ and the standardized residuals from the regression model (6). We introduce:

```
> par(mfrow=c(2,2)); plot(x1,fm1$residuals,ylab="Residuals",main="Grafika")
> abline(lsfit(x1,fm1$residuals))
> plot(x2,fm1$residuals,ylab="Residuals",main="Grafika")
> abline(lsfit(x2,fm1$residuals))
```

We can also find a 95% prediction interval for the coefficients in the regression equation through:

```
> round(confint(fm1,level=0.95),3)
             2.5 % 97.5 %
x1           0.541  0.993
x2           0.067  0.233
```

as well as a 95% prediction interval for the dependent variable y and for its fitting value $\hat{y}$ respectively:

```
> predict(fm1,newdata=dummy,interval="prediction",level=0.95)
> predict(fm1,newdata=dummy,interval="confidence",level=0.95)
```

CONCLUSIONS

The above examples illustrate that the R software is very useful for learning the theoretical material of RA.

REFERENCES

[1] Kabacoff, R., R in Action. Early Access Edition, 2009.
[2] Sheather, S., A Modern Approach to Regression with R. Springer, 2009.
[3] Костадинова, Кр.. Използване на MS Excel в обучението по статистика. Сборник Научни трудове на Русенски университет, том 48, серия 6.1, Русе, 2009, стр. 50-54 вкл.

ABOUT THE AUTHOR

Assist. Krasimira Yankova Kostadinova, Department of Economics and Modeling, University of Shumen, Phone: +359 54 830 495, e-mail: kostadinova_kr@abv.bg.

Докладът е рецензиран.