International Journal of Computational and Applied Mathematics.
ISSN 1819-4966 Volume 12, Number 3 (2017), pp. 923-934

http://www.ripublication.com

Applications of Computational Statistics with Multiple Regressions

S MD Riyaz Ahmed

Research Scholar, Department of Mathematics,
Rayalaseema University, A.P, India

Abstract

Applications of Computational Statistics with Multiple Regression through the use of Excel, MATLAB and SPSS. It has been widely used for a long time in various fields, such as operations research and analysis, automatic control, probability statistics, statistics, digital signal processing, dynamic system mutilation, etc. The regress function command in MATLAB toolbox provides a helpful tool for multiple linear regression analysis and model checking. Multiple regression analysis is additionally extremely helpful in experimental scenario wherever the experimenter will management the predictor variables.

Keywords - Multiple Linear Regression, Excel, MATLAB, SPSS, least square, Matrix notations

1. INTRODUCTION

Multiple linear regressions are one of the foremost wide used of all applied mathematics strategies. Multiple regression analysis is additionally extremely helpful in experimental scenario wherever the experimenter will management the predictor variables [1]. A single variable quantity within the model would have provided an inadequate description since a number of key variables have an effect on the response variable in necessary and distinctive ways that [2]. It makes an attempt to model the link between two or more variables and a response variable by fitting a equation to determined information. Each value of the independent variable x is related to a value of the dependent variable y. The population regression line for p informative variables

$x_1, x_2, x_3 ..... x_n$ is defined to be:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p$$

This line describes how the mean response μ changes with the explanatory variables. The observed values for y vary about their means μ and are assumed to have the same standard deviation σ. The fitted values $b_0, b_1, ..., b_p$ estimate the parameters $\beta_0, \beta_1, \beta_2, ...\beta_p$ of the population regression line [2]. β_0 is the mean of y when all x's are 0.Meanwhile, β_i is the change in the mean of Y associated with a unit increase in x_i, holding the values of all the other x's fixed. Coefficient estimated via least squares. A simple linear regression illustrates the relation between the dependent variable y and therefore the independent variable x based on the regression equation

$$y_i = \beta_0 + \beta_1 x_i + e_i \ , \quad i = 1, 2, 3, ..., n \tag{1}$$

Using the least squares technique, the simplest fitting line will be found by minimizing the total of the squares of the vertical distance from every information on the road. For additional attention-grabbing discussion on this subject see Gordon and Gordon (2004) and Scarianio and Calzada (2004). According to the multiple linear regression model the dependent variable is related to two or more independent variables. The general model for k variables is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \ , \quad i = 1, 2, 3, ..., n \tag{2}$$

Using matrices allows for a more compact framework in terms of vectors representing the observations, levels of regression variables, regression coefficients, and random errors. The model is in the form

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \tag{3}$$

and when written in matrix notation we have

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{1k} \\ 1 & x_{21} & x_{1k} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_1 \\ \vdots \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{4}$$

Note that Y is an nx1 dimensional random vector consisting of the observations, X is an n x (k + 1) matrix determined by the predictors, β is a (k+1)x1 vector of unknown parameters, and ε is an n x 1 vector of random errors. The first step in multiple linear regression analysis is to determine the vector of least squares estimators, $\hat{\beta}$, which gives the linear combination ˆy that minimizes the length of the error vector. Basically the estimator $\hat{\beta}$ provides the least possible value to sum of the squares difference

between ˆy and y, algebraically $\hat{\beta}$ can be expressed by using matrix notation. An important stipulation in multiple regression analysis is that the variables $x_1, x_2, ..., x_n$ be linearly independent. This implies that the correlation between each xi is small. Now, since the objective of multiple regression is to minimize the sum of the squared errors, the regression coefficients that meet this condition are determined by solving the least squares normal equation

$$X^T \times \hat{\beta} = X^T Y \tag{5}$$

Now if the variables $x_1, x_2, ..., x_n$ are linearly independent, then the inverse of $X^T X$, namely $(X^T X)^{-1}$

will exist. Multiplying both sides of the normal equation 5 by $(X^T X)^{-1}$, we obtain

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

Several mathematical software packages such as Mathematica, Stata and MATLAB provide matrix commands to determine the solution to the normal equation as shown in Math Works (2006), Kohler and Kreuter (2005), and Research (2006). The aim of making a program during this manner fosters a decent understanding of algebra and multiple linear regression analysis.

2. A MATLAB APPROACH

There are several options in MATLAB to perform multiple linear regression analysis. One option is Generalized Linear Models in MATLAB (glmlab) which is available in Windows, Macintosh, or UNIX. Variables and data can be loaded through the main glmlab window screen. For further details see Dunn (2000) about the capabilities of glmlab. Another option is the Statistical Toolbox, which allows the user to program with functions. MATLAB programs can also be written with m-files. These files are text files created with either functions or script. A function requires an input or output argument. While the function method simplifies writing a program, using script better illustrates the process of obtaining the least squares estimator using matrix commands. In our example we will use script to write our program. The following example we are measuring the quantity y (dependent variable) for several values of x_1 and x_2 (independent variables). We will use the following tables of values:

y	x_1	x_2
0.21	0.7	0.6
0.31	0.9	0.8
0.32	1.1	0.9
0.27	1.3	1.4
0.31	1.5	1.6
0.30	1.6	1.9

(7)

The least squares estimators of $\hat{\beta}$ are found by writing the following MATLAB program in script form using matrix notation:

```
X=[1 .7 .6;1 1.1 .8;1 1.1 .9;1 1.3 1.5;1 1.5 1.6;1 1.6 1.9];  X
Y=[.21;.31;.32;.27;.31;.30]; Y
A=transpose(X)*X; A
K=(transpose(X)*X)^-1; K
B=K*transpose(X)*Y; B
M=X*B; M
E=Y-M; E
MaxErr=max(abs(Y-M))
```

The importance of these steps in the program is to illustrate the use of matrix algebra to find the least square estimators. Recall the least squares estimators $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$

.

The first step in the program computes the product of X^T and X as follows: $A = X^T X$

$$\begin{vmatrix} 6 & 7.30 & 7.30 \\ 7.30 & 9.41 & 9.68 \\ 7.30 & 9.68 & 10.23 \end{vmatrix} \qquad (8)$$

In this next step, the instructor can reinforce the concept of the inverse existing only if

the columns of X are linearly independent. In our case the inverse does exist as, $K = \left(X^T X \right)^{-1}$

$$\begin{vmatrix} 5.69 & -8.91 & 4.37 \\ -8.91 & 17.96 & -10.63 \\ 4.37 & -10.63 & 7.04 \end{vmatrix} \quad (9)$$

We can now find the least squares estimators, $B = \hat{\beta} = KX^T Y$

$$\begin{vmatrix} 0.0768 \\ 0.3496 \\ -0.1771 \end{vmatrix} \quad (10)$$

According to these values the corresponding fitted regression model is:

$$y = 0.0768 + (0.3496)x1 + (-0.1771)x2 \quad (11)$$

One additional step is to validate the regression model for the data by computing the maximum error e. In our example we note the error matrix is as follows:

$$\begin{vmatrix} -0.0052 \\ -0.0097 \\ 0.0180 \\ 0.0044 \\ -0.0078 \\ 0.0003 \end{vmatrix} \quad (12)$$

Based on these values one will find the maximum error to be 0.0180, which indicates the model accurately follows the data.

3. SUGGESTION OF THE PROBLEM

In this problem we introduced an alternative approach of combining MATLAB script and matrix algebra to analyze multiple linear regressions. This approach is comparatively easy and offers the students the chance to develop their abstract understanding of algebra and multiple linear regression models. It's been my expertise in analyzing a multiple linear regression model exploitation the MATLAB script approach is that it higher permits one to look at what's happening “behind the scenes” throughout computations. Usually employing a windows approach in SPSS or perform approach in MATLAB involves inputting values and blindly applying the technology without understanding the relationship between the algorithm and the results. As with any software package, MATLAB has limitations with the script

approach to analyze more advanced statistical techniques. For this reason it is recommended the reader review the various software packages to determine which is best suited for their instructional needs.

4. MODELING WITH MULTIPLE REGRESSION USING SPSS

It is a useful method to generate the mathematical model where several variables are involved. A computer program can be written to evaluate the coefficients in the equation

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \ldots\ldots\ldots\ldots + a_n x_n \quad (n \neq 0)$$

Where $x_1, x_2, \ldots\ldots\ldots, x_n$ the independent variables and y are is the dependent variable. This model can applied in various applications such as:

a) Assessing the weights of 6 shipments.

b) Assessing the IQ of 5 students, number of hours they studied for the Examinations and the marks.

a) Assessing the weights of 6 shipments.

Applications:

Consider the data given below on the weights of 6 shipments, the distances covered by them and the damage incurred by the company.

Weight in 1000 Kg – X_1	Distance in 1000 Km – X_2	Damage in Rupees - Y
4.0	1.5	160
3.0	2.2	112
1.6	1.0	69
1.2	2.0	90
3.4	0.8	123
4.8	1.6	186

Assuming that the regression is linear we obtain the regression equation. $Y = a_0 + a_1 x_1 + a_2 x_2$

b) Assessing the IQ of 5 students, number of hours they studied for the Examinations and the marks.

The following data consist of the IQ of 5 students, number of hours they studied for the Examinations and the marks they obtained in an examination

IQ X_1	No. of hours they studied X_2	Score Y
112	5	79
126	13	97
100	3	51
114	7	65
112	11	82

Assuming that the regression is linear, we estimate the coefficients a_0, a_1 and a_2

RESULTS AND DISCUSSION:

In Application (a) : The regression equation calculated by using SPSS package is

Y = 23.025 +27.167.(weight) +10.703.(distance)

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT Damage_Rupees_Y
 /METHOD=ENTER Weight_1000kg_X1 Distance_1000km_X2
 /RESIDUALS HIST(ZRESID).
```

REGRESSION

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Distance_1000km_X2, Weight_1000kg_X1[a]	.	Enter

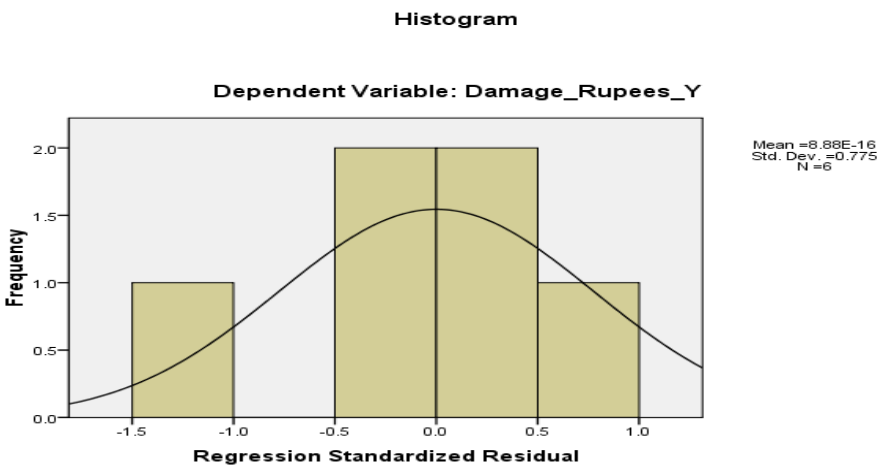
a. All requested variables entered.

b. Dependent Variable: Damage_Rupees_Y

Coefficients[a]

Unstandardized Coefficients		Standardized Coefficients		
B	Std. Error	Beta	t	Sig.
14.562	26.905		.541	.626
30.109	5.153	.959	5.843	.010
12.161	13.087	.153	.929	.421

a. Dependent Variable: Damage_Rupees_Y



And it can be concluded that in order to predict the damage only weight seems to be significant.

In Application (b) : The regression equation. Calculated by using SPSS package is

$$Y = -62.785 + 1.114 X_1 + 1.531 X_2$$

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
```

/NOORIGIN

/DEPENDENT Score_Y

/METHOD=ENTER IQ_X1 Number_of_hours_X2

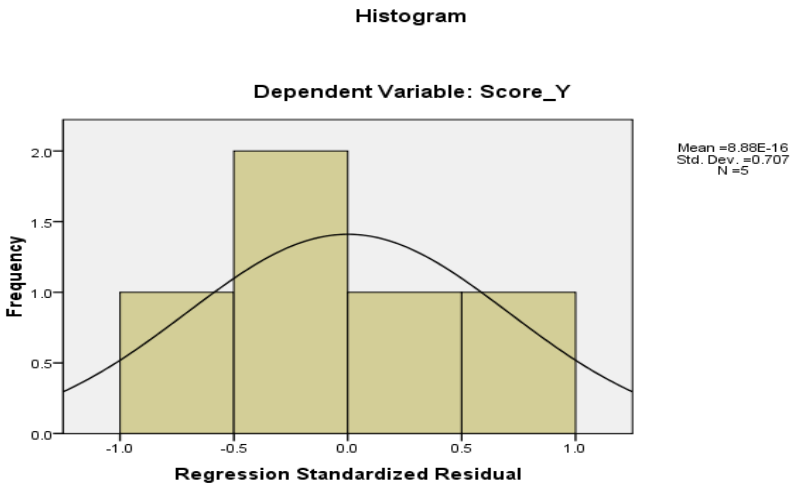
/RESIDUALS HIST(ZRESID).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-62.785	99.255		-.633	.592
	IQ_X1	1.114	1.005	.588	1.108	.383
	Number_of_hours_X2	1.531	2.237	.363	.684	.564

a. Dependent Variable: Score_Y

Histogram



Here no variable seems to be significant.

5. LEAST SQUARES NORMAL EQUATIONS BY USING MATRIX ALGEBRA

The normal equations in the matrix format:

$$\begin{bmatrix} n & \sum x_2 & \sum x_3 \\ \sum x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix}$$

with the following shorthand notation $\mathbf{Xb} = \mathbf{c}$: where,

$$\mathbf{X} = \begin{bmatrix} n & \sum x_2 & \sum x_3 \\ \sum x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix} \sum y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix}$$

Solutions for b_j are obtained by finding the product of the inverse matrix of **X,** $\mathbf{X^{-1}}$**,** times **c**. $\mathbf{b} = \mathbf{X^{-1}c}$

A regression of $monthly\ sales\ (y)$, in \$1,000's, on $price\ (x_2)$, in dollars, and $advertising\ (x_3)$, in 1,000's dollars, of a fast food restaurant. Use *Excel* to compute the values for the elements of the matrix **X** and **k**:

$$\begin{bmatrix} n & \sum x_2 & \sum x_3 \\ \sum x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} = \begin{bmatrix} 75 & 426 & 138 \\ 426 & 2446 & 787 \\ 138.3 & 787 & 306 \end{bmatrix}$$

$$\begin{bmatrix} \sum y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix} = \begin{bmatrix} 5803.1 \\ 32847.7 \\ 10789.6 \end{bmatrix}$$

Thus, normal equations can be written as:

$$\begin{bmatrix} 75 & 426.5 & 138.3 \\ 426.5 & 2445.707 & 787.381 \\ 138.3 & 787.381 & 306.21 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 5803.1 \\ 32847.7 \\ 10789.6 \end{bmatrix}$$

Since $b = X^{-1}c$, we need to find X^{-1}. Now that we know what the inverse of a matrix means and how to find it, we have *Excel* to do the hard work for us. In *Excel* use the array type function **=MINVERSE**(). We must focus a block of 3×3 cells, then call up the function =MINVERSE(). When it's "array", we must "lasso" the

block of cells comprising the elements of the X matrix and then press Ctrl+Shift+Enter keys composed. The result is the following:

1.689828	-0.28462	-0.03135
-0.28462	0.050314	-0.00083
-0.03135	-0.00083	0.019551

The matrix X^{-1}. Then pre multiplying k by X^{-1}, we have

b_1		1.689828	-0.28462	-0.03135		5803.1	118.9136
b_2	=	-0.28462	0.050314	-0.00083	×	32847.7	= -7.90785
b_3		-0.03135	-0.00083	0.019551		10789.6	1.862584

$\hat{y} = 118.9136 - 7.90785x_2 + 1.862584x_3$

$\widehat{SALES} = 118.914 - 7.908PRICE + 1.863ADVERT$

$b_2 = -7.908$ Indicates that for every $ growth in price monthly sales would decrease by 7,908dollar. Or, a 10¢ growth in price would outcome in a fall in monthly sales of 790.8 dollar. $b_3 = \$1.863$ Implies that for every extra $1,000 of marketing sales would growth by 1,863 dollar. The following is the *Excel* regression result shows the expected coefficients.

OUTPUT

Regression Statistics	
Multiple R	0.669521
R Square	0.448258
Adjusted R Square	0.432932
Standard Error	4.886124
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1396.5389	698.26946	29.247859	5.04086E-0
Residual	72	1718.9429	23.874207		
Total	74	3115.4819			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	118.91361	6.35164	18.72172	0.00000	106.25185	131.57537
PRICE	-7.90785	1.09599	-7.21524	0.00000	-10.09268	-5.72303
ADVERT	1.86258	0.68320	2.72628	0.00804	0.50066	3.22451

REFERENCES

[1] M Ozaki, Iznibinti Mustafar and RadzuanRazali.(2011).A Study on Prediction of Output in Oilfield Using Multiple Linear Regression.

[2] Kutner, Natchsheim and Neter. 2004. Applied Linear Regression Model,(4), New York, McGraw Hill.

[3] Z. Zhiyong: *Course of MATLAB* (Beijing University of Aeronautics and Astronautics Press, Beijing, (2009).

[4] H. Jianyong: *Operations Research* (Tsinghua University Press, Beijing, 2000) and M. Zhenhua: *Probability Theory and Random Process* (Tsinghua University Press, Beijing, (2000).

[5] B.Muthen, T. Asparouhov: Multilevel regression mixture analysis, Journal of the Royal Statistical Society Series a-Statistics in Society. Vol.172 (2009), p. 639-657.

[6] Chatterjee, S., and A. S. Hadi. "Influential Observations, High Leverage Points, and Outliers in Linear Regression." *Statistical Science*. Vol. 1, 1986, pp. 379–416.

[7] Belsley, D.A., R.E. Welsch and E. Kuh, 1980. Regression Diagnostics, John Wiley & Sons, Inc. (New York: New York).

[8] Chib, Siddhartha. 1993. \Bayes regression with autoregressive error: A Gibbs sampling approach," Journal of Econometrics, Vol. 58, pp.275-294.

[9] Geweke, John. 1993. "Bayesian Treatment of the Independent Student t Linear Model", Journal of Applied Econometrics, Vol. 8, s19-s40.