Alabama Journal of Mathematics
Spring/Fall 2009

Multiple Linear Regression Analysis: A Matrix Approach with MATLAB

Scott H. Brown
Auburn University Montgomery

Linear regression is one of the fundamental models in statistics used to determine the relationship between dependent and independent variables. An extension of this model, namely multiple linear regression, is used to represent the relationship between a dependent variable and several independent variables. This article focuses on expressing the multiple linear regression model using matrix notation and analyzing the model using a script approach with MATLAB. This approach is designed to enable high school or university students to better understand matrix operations and the algorithm used to analyze multiple linear regression.

Multiple Linear Regression Model

A simple linear regression illustrates the relation between the dependent variable y and the independent variable x based on the regression equation

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

Using the least squares method, the best fitting line can be found by minimizing the sum of the squares of the vertical distance from each data point on the line. For further interesting 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

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

The simple linear regression model is used to find the straight line that best fits the data. On the other hand, the multiple linear regression model, for example with two independent variables, is used to find the *plane* that best fits the data. Models that involve more than two independent variables are more complex in structure but can still be analyzed using multiple linear regression techniques.

In multiple linear regression analysis, the method of least squares is used to estimate the regression coefficients in 2. The regression coefficients illustrate the unrelated contributions of each independent variable towards predicting the dependent variable. Unlike the simple linear regression, there must be inferences made about the degree of interaction or correlation between each of the independent variables. The computations used in finding the regression coefficients ($\beta_i, \quad i = 1, \ldots, k$), residual sum of squares (SSE), regression sum of squares (SSR), etc. are rather complex. To simplify the computation, the multiple regression model in terms of the observations can be written using matrix notation.

A Matrix Approach to Multiple Linear Regression Analysis

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

$$Y = X\beta + \epsilon \tag{3}$$

and when written in matrix notation we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{4}$$

Note that Y is an $n \times 1$ dimensional random vector consisting of the observations, X is an $n \times (k \times 1)$ matrix determined by the predictors, β is a $(k+1) \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ 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 $\hat{y}$ 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 $\hat{y}$ and y. Algebraically $\hat{\beta}$ can be expressed by using matrix notation. An important stipulation in multiple regression analysis is that the variables $x_1, x_2, \ldots, x_n$ be linearly independent. This implies that the correlation between each x_i 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

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

Now if the variables $x_1, x_2, \ldots, x_n$ are linearly independent, then the inverse of $X^T X$, namely $\left(X^T X\right)^{-1}$ will exist. Multiplying both sides of the normal equation 5 by $\left(X^T X\right)^{-1}$, 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 MathWorks (2006), Kohler and Kreuter (2005), and Research (2006). The reader will also find the more advanced Texas Instrument (TI) graphing calculator that will allow a student to perform multiple linear regression analysis by using the matrix approach. An application of the graphing calculator approach can be found in Wilson et al. (2004). We will focus on using MATLAB and the option to write a program with matrix commands. *The purpose of creating a program in this manner fosters a good understanding of matrix algebra and multiple linear regression analysis.*

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 either 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.

In 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
.19	.5	.4
.28	.8	.6
.30	.9	.7
.25	1.1	1.2
.29	1.3	1.4
.28	1.4	1.7

(7)

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

```
X=[1 .5 .4;1 .8 .6;1 .9 .7;1 1.1 1.2;
1 1.3 1.4;1 1.4 1.7];
X
Y=[.19;.28;.30;.25;.29;.28];
Y
A=XT*X;
A
K=(XT*X)^-1;
K
B=K*XT*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 $\hat{\beta} = (X^T X)^{-1} X^T Y$. The first step in the program computes the product of X^T and X as follows:

$$\begin{aligned} A &= X^T X \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ .5 & .8 & .9 & 1.1 & 1.3 & 1.4 \\ .4 & .6 & .7 & 1.2 & 1.4 & 1.7 \end{bmatrix} \begin{bmatrix} 1 & .5 & .4 \\ 1 & .8 & .6 \\ 1 & .9 & .7 \\ 1 & 1.1 & 1.2 \\ 1 & 1.3 & 1.4 \\ 1 & 1.4 & 1.7 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 6 & 6 \\ 6 & 6.56 & 6.83 \\ 6 & 6.8 & 7.3 \end{bmatrix} \end{aligned} \tag{8}$$

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,

$$K = (X^T X)^{-1} = \begin{bmatrix} 5.2818 & -12.0205 & 6.9054 \\ -12.0205 & 33.2481 & -21.2276 \\ 6.9054 & -21.2276 & 14.3223 \end{bmatrix} \tag{9}$$

We can now find the least squares estimators,

$$B = \hat{\beta} = KX^T Y = \begin{bmatrix} 0.0658 \\ 0.0453 \\ -0.2540 \end{bmatrix} \tag{10}$$

According to these values the corresponding fitted regression model is:

$$y = 0.0658 + (0.4532)x_1 + (-0.2540)x_2 \tag{11}$$

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:

$$E = \epsilon = \begin{bmatrix} -0.00008 \\ 0.0041 \\ 0.0041 \\ -0.0095 \\ -0.0094 \\ 0.0115 \end{bmatrix} \tag{12}$$

Based on these values one will find the maximum error to be 0.0115, which indicates the model accurately follows the data.

Conclusion

In this paper we introduced an alternative approach of combining MATLAB script and matrix algebra to analyze multiple linear regression. This approach is relatively simple and offers the students the opportunity to develop their conceptual understanding of matrix algebra and multiple linear regression model.

It has been my experience in analyzing a multiple linear regression model using the MATLAB script approach is that it better enables one to observe what is going on "behind the scenes" during computations. *Normally using a windows approach in SPSS or function 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.

Acknowledgements

The author would like to thank the anonymous referee for the suggestions and comments that improved this paper.

References

Dunn, P. (2000). glmlab-generalized linear models in MATLAB. accessed from (LINK) on December 21, 2006.

Gordon, S. and Gordon, F. (2004). Deriving the regression equations without calculus. *Mathematics and Computer Education*, 38(1):64–68.

Kohler, U. and Kreuter, F. (2005). *Data Analysis Using Stata*. Stata Press, College Station, TX.

MathWorks (2006). Multiple regression. accessed from (LINK) on December 21, 2006.

Montgomery, D. and Peck, E. (1992). *Introduction to Linear Regression Analysis*. John Wiley and Sons, Inc., New York, NY, 2nd edition.

Neter, J., Wasserman, W., and Kutner, M. (1990). *Applied Linear Statistical Models*. Richard D. Irwin, Inc., Homewood, CA, 3rd edition.

Research, W. (2006). How do I perform multivariate linear regression with *Mathematica*? accessed from (LINK) on December 21, 2006.

Scarianio, S. and Calzada, M. (2004). Three perspectives on teaching least squares. *Mathematics and Computer Education*, 38:255–264.

Wilson, W., Geiger, L., Madden, S., Mecklin, C. J., and Dong, A. (2004). Multiple linear regression using a graphing calculator. *Journal of Chemical Education*, 81(6):903–907.