A STUDY ON MULTIPLE LINEAR REGRESSION USING MATRIX CALCULUS

B. MAHABOOb1, J. PETER PRAVEEN, B. V. APPA RAO, Y. HARNATH, C. NARAYANA, AND G. BALAJI PRAKASH

ABSTRACT. In spite of accessibility of a large number of imaginative tools in Applied Mathematics the most important tool of the Mathematician is always the linear model as it involves uncomplicated and apparently most restrictive properties namely linearity, constancy of variance, normality and independence. Linear models and the methods connected to it are remarkable, flexible and powerful. Since almost all advanced statistical tools are generalizations of linear models proficiency in linear models is a prerequisite to study advanced statistical tools. This research article primarily focuses on the specific forms of Simple Linear Regression Model, Multiple Linear Regression Model, LSE of its parameters and the properties of LSE. Furthermore an innovative proof of Gauss-Markov theorem has been proposed by means of Principles of Matrix Calculus. In addition to these the concept of BLUE has been depicted.

1. INTRODUCTION

A model with theoretical framework helps in better understanding of a given phenomenon and it is a mathematical construction which generates the given observations. The linear model is primary to the training of any theoretical or applied mathematician. Scientific method is often used as a guided approach for

1corresponding author

2010 Mathematics Subject Classification. 62G08.

Key words and phrases. Response and Predictor variables, LSE (Least Square Estimator), MLE (Maximum Likelihood Estimator), Regression Coefficients, BLUE (Best Linear Unbiased Estimator).
learning process. Linear statistical models are extensively used as part of these processes and they are useful in the biological, physical, and social sciences as well as in business and engineering. Though these constructed mathematical models may be oversimplification of the real world complex problems they can give useful approximations of the relationships among the observations. Good estimators of parameters are essential for better performance in prediction. The scientists and engineers employ estimated models for describing or summarizing the observed data. B. Mahaboob et al. in 2017 in their paper estimated the parameters of CES production functional model using the principles of Matrix Calculus. C. Narayana et al., in 2018 in their paper studied the misspecification and predictive accuracy of stochastic linear regression models. In 2017 B. Mahaboob et al., in their research article proposed some computational techniques for least squares estimator and maximum likelihood estimator by means of Matrix Calculus. In 2019 B. Mahaboob et al., in their research paper depicted the estimation methods of Cobb-Douglas production functional model. In 2018 Kushbukumari et al., in their paper explains the basic concepts of Linear Regression Analysis and explains how we can do linear regression calculations in SPSS and excel. In 2018, Shrikant I. Bangdiwala in his article Regression: simple linear explained the methods of fitting simple linear regression models. W. Superta et al., in 2017, in their article developed a numerical model with the use of nonlinear model to estimate the thunder storm activity.

For further references, see [1]-[28].

2. Simple Linear Regression Model

In simple linear regression model an attempt can be made to model the relationship between the two variables by the form

\[ Y = \alpha_0 + \alpha_1 X + \epsilon. \]  

where

\( Y = \) Response Variable (Dependent Variable)  
\( X = \) Predictor Variable (Independent Variable)  
\( \epsilon = \) Random Variable (Error Term)
The linearity appearing in (2.1) is just an assumption. Some more assumptions namely distribution of independence of \( Y \) can be added. The \( \alpha_0 \) and \( \alpha_1 \) can be estimated by making use of observed values of \( X \) and \( Y \) and some inferences namely confidence intervals and testing of hypothesis for \( \beta_0 \) and \( \beta_1 \) can be made. The estimated model can be used to forecast the value of \( Y \) for a particular value of \( X \).

3. Multiple Linear Regression Model

The dependent variable \( Y \) is sometimes affected by more than one independent variable. A linear model making association between \( Y \) and several predictors can be framed as

\[
Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_i X_i + \epsilon.
\]

The arbitrary constants \( \alpha \)'s are called regression coefficients. The random variation in \( Y \) which cannot be explained by predictors is provided by the error term \( \epsilon \). The model in (3.1) is linear in \( \beta \)'s and not necessary in predictors. To estimate \( \beta \)'s in (3.1) one can use a sample of 'n' observations on \( Y \) and the associated \( X \). For \( k^{th} \) observation the model is

\[
Y_k = \alpha_0 + \alpha_1 X_{k1} + \alpha_2 X_{k2} + \ldots + \alpha_i X_{ki} + \epsilon_k.
\]

The assumptions for \( \epsilon_k \) are as follows:

(i) \( E(\epsilon_k) = 0 \) for \( k = 1, 2, \ldots, m \), that is \( E(Y_k) = \alpha_0 + \alpha_1 X_{k1} + \alpha_2 X_{k2} + \ldots + \alpha_i X_{ki} \).

(ii) \( Var(\epsilon_k) = \sigma^2 \) for \( k = 1, 2, \ldots, m \), that is \( Var(Y_k) = \sigma^2 \).

(iii) \( Cov(\epsilon_k, \epsilon_l) = 0 \) for all \( k \neq l \), that is \( Cov(Y_k, Y_l) = 0 \).

(i) tells us that model is correct i.e., it is indeed linear, and
(ii) implies that variance of \( Y \) is constant.

If all the above assumptions hold good the poor estimator will be achieved. If the normality assumption is inserted one can obtain MLE which can have excellent
properties. For each of `n` observations (3.2) can be depicted as

\[
Y_1 = \alpha_0 + \alpha_1 X_{11} + \alpha_2 X_{12} + \ldots + \alpha_i X_{1i} + \epsilon_1 \\
Y_2 = \alpha_0 + \alpha_1 X_{21} + \alpha_2 X_{22} + \ldots + \alpha_i X_{2i} + \epsilon_2 \\
\vdots \\
Y_m = \alpha_0 + \alpha_1 X_{mi} + \alpha_2 X_{m2} + \ldots + \alpha_i X_{mi} + \epsilon_m .
\]

Using matrices the above can be put as

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_m
\end{bmatrix} =
\begin{bmatrix}
1 & X_{11} & X_{12} & \ldots & X_{1i} \\
1 & X_{21} & X_{22} & \ldots & X_{2i} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{mi} & X_{m2} & \ldots & X_{mi}
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_i
\end{bmatrix}
+ 
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_m
\end{bmatrix}.
\]

(3.3)

\[
\bar{Y} = X\bar{\alpha} + \bar{\epsilon}.
\]

hence

\[
E(\bar{\epsilon}) = 0 \text{ or } E(\bar{y}) = \bar{X}\bar{\alpha} \\
Cov(\bar{\epsilon}) = \sigma^2\bar{I} \text{ or } Cov(\bar{Y}) = \sigma^2\bar{I} .
\]

\(X\) is a matrix of order \(m \times (i + 1)\) and assume that \(m > i + 1\) and \(\rho(X) = i + 1\). The \(\alpha\) parameters in (2.1) or (3.3) are called regression coefficients. The partial
derivative of

\[
E(Y) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots \alpha_i X_i ,
\]

with respect to \(X_2\) for example is \(\alpha_2\). Hence regression coefficients are sometimes
referred to as partial regression coefficients. Thus \(\alpha_2\) indicates the change in
\(E(Y)\) with a unit change in \(X_2\) when \(x_1, X_2, X_3 \ldots X_i\) are fixed. \(\alpha_2\) depicts the
influence of \(X_2\) of \(E(Y)\) in the presence of other \(X\)'s. This effect is different from
the effect of \(X_2\) on \(E(Y)\) if the other \(X\)'s are not appearing in the model. For
instance \(\alpha_0\) and \(\alpha_1\)

\[
Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon
\]
differ from \(\alpha_0^*\) and \(\alpha_1^*\) in

\[
Y = \alpha_0^* + \alpha_1^* X_1 + \epsilon^* .
\]

Roughly speaking if one is \(X\) is deleted then the parameters will be changed.
4. Least Square Estimator

For the parameters \( \alpha_0, \alpha_1, \ldots, \alpha_i \) one can seek the estimates that minimize

\[
\sum_{k=1}^{m} \hat{\epsilon}_k^2 = \sum_{k=1}^{m} (Y_k - \hat{Y}_k)^2 = \sum_{k=1}^{m} (Y_k - \hat{\alpha}_0 - \hat{\alpha}_1 X_{k1} - \hat{\alpha}_2 X_{k2} - \ldots - \hat{\alpha}_i X_{ki})^2.
\]

To find the values of \( \hat{\alpha}_i \) that minimize (4.1) one can differentiate \( \sum_{k=1}^{m} \hat{\epsilon}_k^2 \) with respect to each \( \hat{\alpha}_i \) and equate the results to 0 to get \((i+1)\) equations that are solved for \( \hat{\alpha}_i \)’s.

This process in compact form with matrix notation is explained below. (4.1) is rewritten as

\[
\hat{\epsilon}^T \hat{\epsilon} = (\bar{Y} - \bar{X} \hat{\alpha})^T (\bar{Y} - \bar{X} \hat{\alpha})
\]

\[
\hat{\epsilon}^T \hat{\epsilon} = \bar{Y}^T \bar{Y} - 2 \bar{Y}^T \bar{X} \hat{\alpha} + \hat{\alpha}^T \bar{X}^T \bar{X} \hat{\alpha}
\]

\[
\frac{\partial \hat{\epsilon}^T \hat{\epsilon}}{\partial \hat{\alpha}} = 0 \Rightarrow \bar{0} - \bar{x}^T \bar{Y} + 2 \hat{\alpha}^T \bar{Y} + 2 \bar{X}^T \bar{X} \hat{\alpha} = 0.
\]

These give normal equations as

\[
\bar{X}^T \bar{X} \hat{\alpha} = \bar{X}^T \bar{Y}.
\]

\( \hat{\alpha} \) is called the Least Square Estimator (LSE). Let \( \bar{b} \) be another estimator in the place of \( \hat{\alpha} \). Then

\[
\hat{\epsilon}^T \hat{\epsilon} = (\bar{Y} - \bar{X} \bar{b})^T (\bar{Y} - \bar{X} \bar{b})
\]

\[
= (\bar{Y} - \bar{X} \hat{\alpha} + \bar{X} \bar{b} - \bar{X} \bar{b})^T (\bar{Y} - \bar{X} \hat{\alpha} + \bar{X} \bar{b} - \bar{X} \bar{b})
\]

\[
= (\bar{Y} - \bar{X} \hat{\alpha})(\bar{Y} - \bar{X} \hat{\alpha}) + (\bar{\alpha} - \bar{b})^T \bar{X}^T \bar{X} (\bar{\alpha} - \bar{b}) + 2(\bar{\alpha} - \bar{b})^T \bar{X}^T \bar{Y} - \bar{X}^T \bar{X} \hat{\alpha}.
\]

From (4.2) the third term becomes 0. Second expression is quadratic form with positive definiteness, so \( \hat{\epsilon}^T \hat{\epsilon} \) is minimized when \( \bar{b} = \hat{\alpha} \).

\( \bar{X}^T \bar{X} \) is square matrix of order \( i+1 \) and it is got as products of columns of \( \bar{X} \)
\( \bar{X}^T \bar{Y} \) is possessing products of columns of \( \bar{X} \) and \( \bar{Y} \)

\[
\bar{X}^T \bar{X} = \begin{bmatrix}
\sum_{k=1}^{m} X_{k1} & \cdots & \sum_{k=1}^{m} X_{ki} \\
\sum_{k=1}^{m} X_{K1} & \cdots & \sum_{k=1}^{m} X_{KL}, X_{ki} \\
\vdots & \ddots & \vdots \\
\sum_{k=1}^{m} X_{kl} & \cdots & \sum_{k=1}^{m} X_{kl}, X_{ki} \\
\sum_{k=1}^{m} X_{ki} & \cdots & \sum_{k=1}^{m} X_{ki}^2
\end{bmatrix}
\]
\[
\hat{X}^T \bar{Y} = \begin{bmatrix}
\sum_k Y_k \\
\sum_k X_{ki} Y_k \\
\vdots \\
\sum_k X_{kl} Y_k
\end{bmatrix}
\]

if \( \hat{\alpha} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T \bar{Y} \) then \( \hat{\epsilon} = \bar{Y} - \hat{X} \hat{\alpha} = \bar{Y} - Y \) is the residual vector
\( \hat{\epsilon}_1 = Y_1 - \hat{Y}_1, \hat{\epsilon}_2 = Y_2 - \hat{Y}_2, \ldots, \hat{\epsilon}_m = Y_m - \hat{Y}_m \).

5. Properties of LSE

(i) If \( E(\bar{Y}) = \hat{X} \hat{\alpha} \) then \( \hat{\alpha} \) is an unbiased estimator for \( \hat{\alpha} \) (shown below)
\[
E(\hat{\alpha}) = E(\hat{X}^T \hat{X})^{-1} \hat{X}^T \bar{Y}
\]
\[
= (\hat{X}^T \hat{X})^{-1} \hat{X}^T E(\bar{Y})(\hat{X}^T \hat{X})^{-1} \hat{X} \hat{\alpha} = \hat{\alpha}
\]

(ii) If \( \text{Cov}(\bar{Y}) = \sigma^2 I \) the covariance matrix for \( \hat{\alpha} \) is \( \sigma^2 (\hat{X}^T \hat{X})^{-1} \) (shown below)
\[
\text{Cov}(\hat{\alpha}) = \text{Cov}( (\hat{X}^T \hat{X})^{-1} \hat{X}^T \bar{Y} )
\]
\[
= (\hat{X}^T \hat{X})^{-1} \hat{X}^T \text{Cov}(\bar{Y})[\hat{X}^T \hat{X}]^T
\]
\[
= (\hat{X}^T \hat{X})^{-1} \hat{X}^T (\sigma^2 I) \hat{X} (\hat{X}^T \hat{X})^{-1}
\]
\[
= \sigma^2 (\hat{X}^T \hat{X})^{-1} \hat{X} \hat{X} (\hat{X}^T \hat{X})^{-1}
\]
\[
= \sigma^2 (\hat{X}^T \hat{X})^{-1}.
\]

6. Principle due to Gauss-Markov

if \( E(\bar{Y}) = \hat{X} \hat{\alpha} \) and \( \text{Cov}(\bar{Y}) = \sigma^2 I \) the LSE \( \hat{\alpha}_l, l = 0, 1, 2 \ldots i \) will possess least variance among all unbiased estimators. Choose a linear estimator \( \bar{X} \bar{Y} \) of \( \hat{\alpha} \) and seek the matrix \( \bar{A} \) for which \( \bar{A} \bar{Y} \) is a least variance unbiased estimator of \( \hat{\alpha} \). For \( \bar{A} \bar{Y} \) to an unbiased estimator of \( \hat{\alpha} \) one should have \( E(\bar{A} \bar{Y}) = \hat{\alpha} \). Since \( E(\bar{Y}) = \hat{X} \hat{\alpha} \) this is written as
\[
E(\bar{A} \bar{Y}) = \bar{A} E(\bar{Y})
\]
\[
= \bar{A} \hat{X} \hat{\alpha} = \hat{\alpha}.
\]
This gives the unbiasedness condition \( \bar{A} \bar{X} = \bar{I} \) since the relationship \( \bar{A} \bar{X} \bar{\alpha} = \bar{\alpha} \) should hold for any possible value of \( \bar{\alpha} \). The covariance matrix for the estimation \( \bar{A} \bar{A} \) is obtained by

\[
\text{Cov}(\bar{A} \bar{Y}) = \bar{A} (\sigma^2 \bar{I}) \bar{A}^{-1} = \sigma^2 \bar{A} \bar{A}^{-1}.
\]

The variance of the \( \alpha_i \)'s are on the diagonal of \( \sigma^2 \bar{A} \bar{A}^T \). So one have to choose \( \bar{A} \) subject to \( \bar{A} \bar{X} = \bar{I} \) and hence the diagonal elements of \( \bar{A} \bar{A}^T \) are minimized. To relate \( \bar{A} \bar{Y} \) to \( \hat{\bar{\alpha}} = (\bar{X}^T \bar{X})^{-1} \bar{X} \bar{Y} \) one can add and subtract \( (\bar{X}^T \bar{X})^{-1} \bar{X}^T \) to get

\[
\bar{A} \bar{A}^T = [\bar{A} - (\bar{X}^T \bar{X})^{-1} \bar{X}^T + (\bar{X}^T \bar{X})^{-1} \bar{X}^T][\bar{A} - (\bar{X}^T \bar{X})^{-1} \bar{X}^T + (\bar{X}^T \bar{X})^{-1} \bar{X}^T]^T.
\]

Expanding this in terms of \( \bar{A} - (\bar{X}^T \bar{X})^{-1} \bar{X}^T \) and \( (\bar{X}^T \bar{X})^{-1} \bar{X}^T \) one can obtain 4 terms in which two become 0 according to \( \bar{A} \bar{X} = \bar{I} \). Thus the output is

\[
(6.1) \quad \bar{A} \bar{A}^T = [\bar{A} - \bar{X}^T \bar{X}]^{-1}[\bar{A} - \bar{X}^T \bar{X}]^{-1} \bar{X}^T \bar{Y}.
\]

The matrix on the RHS of (6.1) is positive semi definite and its diagonal elements are 0. There elements can be made equal to 0 by taking \( \bar{A} \) as \( \bar{A} = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \). The final minimum variance estimation of \( \bar{\alpha} \) is

\[
\bar{A} \bar{Y} = (\bar{X}^T \bar{X})^{-1} \bar{X} \bar{Y}.
\]

This is nothing but the LSE of \( \bar{\alpha} \).

7. Conclusions and Future Research

In this research article simple and multiple linear regression models are specified and the compact form of multiple linear regression model with the help of Matrix Algebra is proposed. Least square estimation of regression coefficients has been described by means of Matrix Calculus. Moreover properties of LSE are established and an innovative proof of Gauss-Markov Theorem has been depicted. In the context of future research the above ideas can be implemented to simple linear regression models also some more concepts namely Geometry Least Squares, Maximum Likelihood Estimation (MLE) and Generalized Least Squares can be established.

References

A STUDY ON MULTIPLE LINEAR REGRESSION...


DEPARTMENT OF MATHEMATICS
KONERU LAKHMAIAH EDUCATION FOUNDATION
VADDESWARAM, GUNTUR, A.P-522502
E-mail address: mahaboob@kluniversity.in

DEPARTMENT OF MATHEMATICS
KONERU LAKSHMAIAH EDUCATION FOUNDATION
VADDESWARAM, GUNTUR, A.P-522502
E-mail address: jppraveen17@kluniversity.in

DEPARTMENT OF MATHEMATICS
KONERU LAKSHMAIAH EDUCATION FOUNDATION
VADDESWARAM, GUNTUR, A.P-522502
E-mail address: bvardr2010@kluniversity.in

DEPARTMENT OF MATHEMATICS
AUDISHANAKRA COLLEGE OF ENGINEERING AND TECHNOLOGY (AUTONOMOUS)
GUDUR, SPSR NELLORE,A.P, -524101
E-mail address: harnath.yeddala@gmail.com

DEPARTMENT OF MATHEMATICS
SRI HARSHA INSTITUTE OF P.G STUDIES
NELLORE, A.P
E-mail address: nareva.nlr@gmail.com

DEPARTMENT OF MATHEMATICS
KONERU LAKSHMAIAH EDUCATION FOUNDATION
VADDESWARAM, GUNTUR, A.P-522502
E-mail address: balajiprakash@kluniversity.in