**European Journal of Molecular & Clinical Medicine**
**ISSN 2515-8260 Volume 07, Issue 3, 2020**

# METHODS AND APPLICATIONS OF LINEAR REGRESSION MODELS

**B.Mahaboob[1], C.Narayana[2], P. Sreehari Reddy[3], Ch. Suresh[4],G.Balaji Prakash[5], Y. Hari Krishna[6]**

[1]Department of Mathematics, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P. India.
[2]Department of Mathematics, Sri Harsha Institute of P.G Studies, Nellore, A.P.
[3]Department of Mathematics, NBKR S&A COLLEGE, Vidyanagar, Nellore, A.P.
[4]Department of Mathematics,Audisankara college of Engineering &Technology(Autonomous) , Gudur, Spsr Nellore Dt,AP,India
[5]Department of Mathematics, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P. India.
[6]Department of Mathematics, ANURAG Engineering College, Anathagiri (v), Kodad, Suryapet, Telangana-508206, India.

**ABSTRACT:**
Linear regression analysis is a statistical phenomenon in order to evaluate the association between the variables. Multiple linear regression models are the one in which there is one dependent variable and more than one independent variables. Regression analysis is an important tool to identify and characterize the relationshipsof multiple factors. The goal of this article is to introduce some methods and applications of linear regression models. The central concepts in linear regression analysis namely estimation theory, maximum likelihood, and linear hypothesis are comprehensively discussed. Moreover an innovative proof of Gauss –Markov theorem in full rank case has been proposed here.

**Keywords:** Linear Regression, Estimation Theory, Maximum Likelihood, Linear Hypothesis, Testing of hypothesis,

## Introduction:

In 1894, Sir Francis Galton introduced the concept of linear regression. Linear regression analysis is a statistical tool applied to the given set of data. In order to trace the quantifying relationship between variables. In 2018, khushbukumari et al. , in their article explained the fundamental properties of linear regression and the methods of performing its calculations in SPSS and excel. Gulden kaya Uyanik et al. in 2013, in their paper,examined the assumptions of multi linear regression analysis –normality,linearity, no extreme values and missing value analysis. RoddyTheobold, in 2017, in their research paper,described an effective frame work of multiple linear regression models. FatemahJalayer et al, in 2015, in their research article,explained Bayesian cloud analysis using linear regression. Gibbs Y. Kanyongo, in 2006, in their research article,applied linear regression analysis in framing the association between home and reading environments.

The specific form of linear hypothesis is described by

$$\bar{z} = \alpha_1 \bar{y}_1 + ..... + \alpha_l \bar{y}_l + \bar{\varepsilon} \qquad (1)$$

Here $\bar{y}_1,.....\bar{y}_l$ are given vectors of constants

$\bar{\varepsilon}$ follows $N_m\left(\bar{0}, \sigma^2 I_m\right)$

The unknown parameters are $\alpha_1....\alpha_l$.

(1) is also known as multiple regression model.

This model includes a large number regression models namely analysis of covariance, one-way analysis of variance , two –way analysis of variance, higher order analysis of variance, simple linear regression. $\bar{z}, y_1,...y_l$ usually take their values on the inner product space $R_m$. An $m \times l$ matrix $\bar{y} = \left(\bar{y}_1,....\bar{y}_l\right)$ and the column vector $\bar{\alpha} = \left(\alpha_1,....\alpha_l\right)^T$ change (1) as $\bar{z} = \bar{y}\bar{\alpha} + \bar{\varepsilon}$.

For instance

i) For the ordered pairs $\left(y_j, z_j\right)$; j=1,2,…m
If we assume $z_j = \alpha y_j + \varepsilon_j$ then it takes
the vector form $\bar{z} = \alpha\bar{y} + \bar{\varepsilon}$ and it is known as regression through origin. If $\bar{y}$ is vector of all ones and $\alpha$ as $\mu$ then $z_i$ follows normal distribution with mean $\mu$ and variance $\sigma^2$.

ii) Let $z_{jk}$= yield of wheat under condition i on $j^{th}$ plot
$y_{jk}$= Fertility of plot k for condition j
k=1,….$m_j$ , j=1,2

$\psi$ = Set of vectors $\begin{bmatrix} z_{11} & z_{21} \\ . & \\ . & \\ z_{1m_1} & z_{2m_2} \end{bmatrix}$

$\bar{z} \in \psi$

$\bar{y}$ = corresponding vector of $y_{jk}$'s

Indicator of first column= $\bar{u}_1$

Indicator of second column= $\bar{u}_2$

Then the model $\bar{z} = \alpha_1\bar{u}_1 + \alpha_2\bar{u}_2 + \alpha_3\bar{y} + \bar{\varepsilon}$ is to used.

iii) If the pairs $\left(y_j, z_j\right), j = 1,2...m$ are observed then the model is $\bar{z}_j = \alpha_0 + \alpha_1 y_j + \alpha_2 y_j^{\,2} + \alpha_3 y_j^{\,3} + \varepsilon_j$

$\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ are independent with the distribution $N(0, \sigma^2)$

Renaming 1 by $u_{oj}$,,$y_j$ by $u_{1j}$, $y^2_j$ by $u_{2j}$ and $y^3_j$ by $u_{3j}$ the above becomes

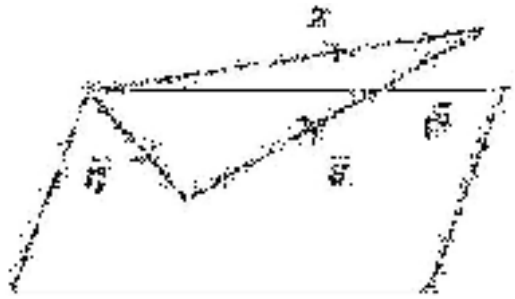
$$z_i = \sum_{k=0}^{3} \alpha_k u_{kj}$$

In vectors it is denoted by $\bar{z} = \sum_{k=0}^{3} \alpha_k u_k + \varepsilon$

**2. Estimation Theory**:

The linear hypothesis can be put as

$\bar{z} = \bar{\eta} + \bar{\varepsilon}$ and described by the following Fig.



$\bar{\eta} \in \bar{V} = L(y_1, \ldots . y_l)$
$\bar{\varepsilon} \sim N(\bar{0}, \sigma^2 I_n)$

In some cases it sufficesto compute $\bar{\eta}$ and its representation as $\sum_{k=1}^{l} \alpha_k y_k$ is not that much important. But the regression coefficients $\alpha_1, \ldots . \alpha_l$ have much importance. Hence matrix notation is implemented. Each and every vector is treated as a column. Put $\bar{y} = (y_1, \ldots y_l)$ and call by design matrix. Now $\eta = \bar{y}\alpha$ . If $y_1, \ldots . . y_l$ are LI then one can have $(\bar{y}^T \bar{y})^{-1} \bar{y}^T \bar{n} = \alpha$

The least squares phenomenon estimates $\eta$ and let

It be $\hat{\eta}$ which minimizes $\|\bar{z} - \eta\|^2 = W(\eta)$

Here $\eta \in \bar{V}$ and $\hat{\eta} = P(\bar{z}/\bar{v}) = \bar{z}$

If $\bar{y}$ has full column rank then

$$\hat{\eta} = \bar{y}(\bar{y}^T \bar{y})^{-1} \bar{y}^T \bar{z}$$

$$\hat{\alpha} = \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \bar{z}$$

Consequently $\hat{\alpha} = \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \left(\bar{y}\beta + \varepsilon\right)$

$$= \bar{\beta} + \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \varepsilon$$

Here $\left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T$ is the Moore-Penrose inverse $\bar{y}^+$ of $\bar{y}$ and it is called the coefficient matrix.

If the column vector of $\bar{y}$ are orthogonal then

$$\bar{z} = \sum p\left(\bar{z}/y_k\right) = \sum \hat{\alpha}_k y_k$$

Where $\hat{\alpha}_k = \left(\bar{z}, y_k\right)/\left\|y_k\right\|^2$

$$= \alpha_k + \left(\varepsilon, y_k\right)/\left\|y_k\right\|^2$$

Moreover $E\left(\hat{\bar{\alpha}}\right) = \bar{\alpha} + \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T E\left(\varepsilon\right) = \bar{\alpha}$

$$D(\hat{\bar{\alpha}}) = D\left[\left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \varepsilon\right]$$

$$= \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \left(\sigma^2 I_m\right)\left[\left(\bar{y}^T \bar{y}\right)\bar{y}^T\right]^T$$

$$= \left(\bar{y}^T \bar{y}\right)^{-1} \sigma^2$$

$\hat{\bar{\alpha}}$ Follows a multivariate normal distribution.

**3. Maximum Likelihood:**

The likelihood function is

$$L\left(\eta, \sigma^2, z\right) = \left(2\pi\right)^{\frac{m}{2}} \sigma^{-m} e^{-0.5\left\|z-\eta\right\|^2 \sigma^{-2}}$$

For each observed $\bar{z} = z$ and $\eta \in \bar{V} = L\left(y_1, \ldots . y_l\right)$

$\sigma^2$ is always +ve.

The phenomenon of maximum likelihood gives the estimates of the pair $\left(\eta, \sigma^2\right)$ which optimises L for each $\bar{z} = z$ . In other words it optimizes

$$\log L = -m(0.5)\log 2\pi - m(0.5)2\log\sigma - (0.5)\left\|z - \hat{\eta}\right\|^2 \sigma^{-2}$$

By choosing $\eta = p\left(\eta/V\right) = \hat{\eta}$, log L is optimized for each fixed $\sigma^2$.

Moreover for this choice of $\eta$ one can see

$$\log L = -m(0.5)\log 2\pi - m(0.5)2\log\sigma - (0.5)\|z-\hat{\eta}\|^2\sigma^{-2}$$

Replacing $\sigma^2$ by u and choosing the differential coefficients of u, one can obtain

$$\frac{d}{du}(\log L) = -m(0.5)u^{-1} + (0.5)\|z-\hat{\eta}\|^2 u^{-2}$$

This becomes 0 for $u = \hat{\sigma}^2 = \|z-\hat{\eta}\|^2 m^{-1}$

It can be easily seen that second derivative is –ve.

Hence $\hat{\sigma}^2 = \|z-\eta\|^2 m^{-1}$ optimizes log Lfor each $\hat{\eta}$. Consequently the pair $(\hat{\eta},\hat{\sigma}^2)$optimizes L .This pair is the MLE of $(\eta,\sigma^2)$.

4. **Estimation of $\sigma^2$**

MLE of $\sigma^2$ **is** $\hat{\sigma}^2 = \|\bar{z}-\hat{\eta}\|^2 m^{-1}$

$$\begin{aligned} E(\hat{\sigma}^2) &= \sigma^2 m^{-1}\dim(V^{\perp}) \\ &= \sigma^2 m^{-1}(m-\dim V) \end{aligned}$$

$\hat{\sigma}^2$ is a biased estimates of $\sigma^2$

Hence the commonly used estimates of $\sigma^2$ **is**

$$R^2 = \|\bar{z}-\hat{\eta}\|^2[m-\dim V]^{-1}$$

If $\varepsilon$ has a MND then $\|\bar{z}-\hat{\eta}\|^2\sigma^{-2} \sim \chi^2_{m-\dim V}$.

As the central $\chi^2$ distribution with n degrees of freedom has a variance 2n, one can obtain

$$\begin{aligned} Var(R^2) &= 2\sigma^4[m-\dim V][m-\dim V]^{-2} \\ &= 2\sigma^4[m-\dim V]^{-1} \end{aligned}$$

**5. Properties of $\hat{\eta}$ and $R^2$**

By facts that $\hat{\eta} = P\left(\bar{z}/V\right), \bar{z}-\hat{\eta} = P(\bar{z}|V^{\perp})$

$V$ and $V^{\perp}$ are orthogonal spaces, one can see that $\hat{\eta}$ and $\bar{z}-\hat{\eta}$ are uncorrelated random vectors, which are independent under normality.

Hence $\hat{\eta}$ and $R^2=\left\|\bar{z}-\hat{\eta}\right\|^2\left[m-\dim V\right]^{-1}$ are independent in the case that the columns of $\bar{y}$ are a basis for V.

If $\bar{z}$ is a multivariate normal random variable $\hat{\alpha}=\left(\bar{y}^T\bar{y}\right)^{-1}\bar{y}^T\hat{\eta}=\left(\bar{y}^T\bar{y}\right)^{-1}\bar{y}^T\bar{z}$ and the residual vector $\bar{e}=\bar{z}-\hat{\eta}$ are uncorrelated independent random vectors.

In order to summerize all the results under the model $\bar{z}=\hat{\eta}+\hat{\varepsilon}$ for $\hat{\eta}\in V$, $\varepsilon\sim N_m\left(\bar{0},\sigma^2 I_m\right)$

One can obtain the following

i) $\hat{\eta}\sim N_m\left(\eta,P_v\sigma^2\right)$

ii) $e=\bar{z}-\hat{\eta}\sim N_m\left(\bar{0},\left(I_m-P_V\right)\sigma^2\right)$

iii) $\hat{\eta}$ and $\bar{z}-\hat{\eta}$ are independent random vectors

iv) $\left\|\bar{z}-\hat{\eta}\right\|^2\sigma^2\sim\chi^2_{m-\dim V}$

v) Hence $R^2=\left\|\bar{z}-\hat{\eta}\right\|^2\left(m-\dim V\right)^{-1}$ is an unbiased estimator of $\sigma^2$

vi) If the columns of $\bar{y}$ serve as a basis of V and $\bar{\eta}=\bar{y}\bar{\alpha}$ then $\bar{\alpha}=\left(\bar{y}^T\bar{y}\right)^{-1}\bar{y}^T\bar{z}$ and $R^2$ are independent provided $\hat{\bar{\alpha}}=N_l\left(\bar{\alpha},\left(\left(\bar{y}^T\bar{y}\right)^{-1}\right)\sigma^2\right)$.Besides the columns of $\bar{y}$ are mutual orthogonal the estimators $\hat{\alpha}$ are not correlated and hence they become independent.

**6. Confidence intervals and Tests on** $\theta=a_1\alpha_1+.....+a_l\alpha_l$

One is generally interested in a linear combination $\theta=\left(\bar{a},\bar{\alpha}\right)=a_1\alpha_1+.....+a_l\alpha_l$. $\hat{\theta}$ is an unbiased estimator of $\theta$ by the linearity of expectation. Its variance is given by

$$Var\left(\hat{\theta}\right)=\bar{a}^1\bar{N}^{-1}\bar{a}\sigma^2$$
$$=d\sigma^2$$

Here $\bar{N}$ is the inner product matrix.

The corresponding estimator of $Var\left(\hat{\theta}\right)$ is $S_\theta{}^2=dS^2$

Particularly when $\theta=\alpha_k$ , a is the $k^{th}$ unit vector and d is the kk term of $N^{-1.}$

$$\hat{\theta} \sim N\left(\theta, d\sigma^2\right)$$

Hence $\dfrac{\hat{\theta}-\theta}{\sqrt{d\sigma^2}} \sim N(0,1)$

$$\frac{\hat{\theta}-\theta}{\sqrt{dS^2}} \sim t_{m-2}$$

Therefore for $t = t_{m-l,1-\alpha}$ , $1-\alpha = P\left(-t \le \dfrac{\hat{\theta}-\theta}{\sqrt{dS^2}} \sim t\right)$

$$P\left(\hat{\theta}\right) - t\sqrt{ds^2} \le \theta \le \hat{\theta} + t\sqrt{ds^2}$$

Hence $\left[\hat{\theta} + t_{m-l,1-\alpha}\sqrt{ds^2}\right]$ is a $100(1-\alpha)\%$ confidence interval on $\theta$.

**7. Tests of hypothesis on $\theta = \sum a_k \alpha_k$**

Let one want to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$

Here $\theta_0$ is a known constant, generally it is 0.

Since $t = \dfrac{\hat{\theta}-\theta_0}{\sqrt{dS^2}} \sim t_{m-l}\left(\dfrac{\hat{\theta}-\theta_0}{\sqrt{d\sigma^2}}\right)$

And this becomes central t where $\theta = \theta_0$.

The tests which reject $H_0$ for $t = \dfrac{\hat{\theta}-\theta_0}{\sqrt{S^2 d}} > t_{m-l,1-\alpha}$ is an $\alpha$ - level test. The two sided hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$ is rejected for $|t| \ge t_{m-l,1-\alpha(0.5)}$

**8. The Gauss-Markov theorem :( Full rank case)**

Suppose $\bar{z} = \sum_{k=1}^{l} \beta_k y_k + \varepsilon$ where $y_1 \ldots . y_l$ are L.I.

here $E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2 I_n$.

Let $\theta = \sum a_k \alpha_k$ and $\theta^*$ be any linear unbiased estimator of $\theta$.

Then $Var(\theta^*) \ge Var(\hat{\theta})$ with equality only if $\theta^* = \hat{\theta}\ \forall\ \bar{V}$

Proof: $\alpha = \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \eta$ and $\theta = \bar{a}^T \bar{\alpha} = \bar{a}^T \left(\bar{y}^T \bar{y}\right)^{-1} \bar{y}^T \eta = \left(\bar{b}, \eta\right)$

Where $\bar{b} = \bar{y}\left(\bar{y}^T \bar{y}\right)^{-1} \bar{a}$.

Take any linear estimator $\theta^* = (\bar{d}, \bar{z})$of $\theta$. Then

$E(\theta^*) = (\bar{d}, \bar{\eta})$

$\theta^*$ is unbiased for $\theta$ if $(\bar{d}, \bar{\eta}) = (\bar{b}, \bar{\eta}) \forall \eta \in \bar{V}$

That is if $(\bar{d} - \bar{b}, \bar{\eta}) = 0$ for all $\eta \in \bar{V}$

In other wordsif $(\bar{d} - \bar{b}) \perp V$ then

$$\theta^* = (\bar{d}, \bar{z}) = (\bar{b}, \bar{z}) + (\bar{d} - \bar{b}, \bar{z}) = \hat{\theta} + (\bar{d} - \bar{b}, \eta + \varepsilon)$$
$$= \hat{\theta} + (\bar{d} - \bar{b}, \varepsilon)$$

As $\bar{b} \perp (\bar{d} - \bar{b})$, the random variables $\hat{\theta}$ and $(\bar{d} - \bar{b}, \bar{a})$ are uncorrelated. Hence one can see $Var(\theta^*) = Var(\hat{\theta}) + \|\bar{d} - \bar{b}\|^2 \sigma^2$

Consequently $Var(\theta^*) \geq Var(\hat{\theta})$ withequality only if $\bar{d} = \bar{b}$ i.e $\theta^* = \hat{\theta}$ for all V.

**9. Conclusions and Future Research:**

The above talk mainly explores on most important concepts of linear regression analysis namely estimation theory, maximum likelihood, specific form of linear hypothesis ,testing of hypothesis and an innovative proof of Gauss-Markov theorem for full rank case. In the context of future research one can extend these ideas to Gauss-Markov theorem for the general case, interpretation of regression coefficients, multiple correlation coefficient and partial correlation coefficient.

**References:**

[1]. Kushbukumari, SunitiYadav (2018), Linear Regression Analysis Study, Journal of the Practice of Cardiovascular Sciences, 2018.volume 4, Issue 1, Pages 33-36

[2]. Gulden Kaya Uyenik,NeseGuler (2013), " A study on multiple linear regression analysis',Procedia- Social and Behavioral Sciences,106:234-240.

[3]. RoddyTheobold, Scott Freeman (2017), "Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research', CBE-Life Sciences Education,Vol.13,No.1

[4]. FatemahJalayer, Raffale De Risi, Gaetano Manfredi (2015), Bayesian cloud Analysis: efficient structural fragility assessment using linear regression ', Bull Earth quakeEng (2015) 13:1183-1203.

[5]. Gibbs Y.Kanyongo,Janine Certo,Brown I Launcelot (2006), "Using regression analysis to establish the relationship between home environment and reading achievement : A case of Zimbabwe" International Education Journal,2006,7(5),632-641.

[6]. B.Mahaboob.,et.al.,(2019)On Cobb-Douglas Production Function Model,AIP Conference Proceedings Recent Trends in Pure and Applied Mathematics2019AIP Conf. Proc. 2177020040-1– 020040-4; https://doi.org/10.1063/1.5135215 Dec 4,2019

[7].B.Mahaboob.,et.al.,(2019)Criteria for Selection of Stochastic Linear Model Selection,AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019,AIP Conf. Proc. 2177, 020041- 1–020041-5; https://doi.org/10.1063/1.5135216,Dec 4,2019

[8].B.Venkateswarlu, et.al.,(2020)Application of DEA in Super Efficiency Estimation, International Journal of Scientific Technology and Research,IJSTR 2020Volume 9,Issue02,February2020,ISSN 2277-8616,Pages:4496-4499

[9]. B.Venkateswarlu.et.al. (2020),Evaluating Different Types of Efficiency Stability Regions and theiInfeasibility in DEA,International Journal of Scientific Technology and Research, IJSTR 2020 Volume 9, Issue02, February2020, ISSN 2277-8616, Pages: 3944-3949

[8]. B.Venkateswarlu. et.al (2020),Multi-Criteria Optimization, Techniques in DEA: Methods& ApplicationsInternationalJournalofScientificTechnologyandResearch, IJSTR2020Volume9, Issue02, February2020, ISSN 2277-8616, Pages: 509-515

[9]. B.Mahaboob. et.al.,(2019),An Evaluation in Generalized LSE of LinearizedStochasticStatistical Modelwith Non-Spherical Errors,AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019AIP Conf. Proc. 2177, 020038-1–020038-5; https://doi.org/10.1063/1.5135213, Dec 4, 2019

[10]. B.Mahaboob. et.al.,(2019),On Misspecification Tests for Stochastic Linear, Regression Model, AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics, 2019,AIP Conf. Proc. 2177, 020039-1–020039-5; https://doi.org/10.1063/1.5135214,Dec 4,2019

[11]. J.PeterPraveen.,et.al.,(2019)A Glance on the Estimation of Cobb-Douglas, Production Functional Model, AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019, AIP Conf.Proc. 2177, 020067-1–020067-4; https://doi.org/10.1063/1.5135242,Dec 4,2019

[12].J.PeterPraveen. et.al.,(2019)On Stochastic Linear Regression Model Selection, AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019,AIP Conf. Proc.2177, 020068-1– 020068-5; https://doi.org/10.1063/1.5135243,Dec 4,2019

[13].D.Ranadheer Reddy, et.al. (2019)Estimation Methods of Nonlinear Regression Models, AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019,AIP Conf. Proc. 2177, 020081-1–020081-5; https://doi.org/10.1063/1.5135256,Dec 4,2019

[14].D.RanadheerReddy., et.al.,(2019)Numerical Techniques of Nonlinear Regression ModelEstimation, AIP Conference Proceedings, Recent Trends in Pure and Applied Mathematics,2019,AIP Conf. Proc. 2177, 020082-1–020082-6; https://doi.org/10.1063/1.5135257,Dec 4,2019

[15].B.Mahaboob.et.al. (2019),On OLS Estimation of Stochastic Linear Regression ModelInternational Journal of Engineering and Advanced Technology (IJEAT),2019ISSN: 2249 – 8958, Volume-8 Issue-6, August, 2019