An econometric model for Linear Regression using Statistics

Renvil Dsa[1], Remston Dsa[2]

[1]Renvil Dsa Fr. Conceicao Rodrigues College of Engineering, Mumbai, India
[2]Remston Dsa New York University, New York, USA

---***---

Abstract - *This research paper discusses the econometric modeling approach of linear regression using statistics. Linear regression is a widely used statistical technique for modeling the relationship between a dependent variable and one or more independent variables. The paper begins by introducing the concept of linear regression and its basic assumptions.*

The univariate and multivariate linear regression models are discussed, and the coefficients of the regression models are derived using statistics. The matrix form of the Simple Linear Regression Model is presented, and the properties of the Ordinary Least Squares (OLS) estimators are proven. Hypothesis testing for multiple linear regression is also discussed in the matrix form.

The paper concludes by emphasizing the importance of understanding the econometric modeling approach of linear regression using statistics. Linear regression is a powerful tool for predicting the values of the dependent variable based on the values of the independent variables, and it can be applied in various fields, including economics, finance, and social sciences. The paper's findings contribute to the understanding of the linear regression model's practical application and highlight the need for rigorous statistical analysis to ensure the model's validity and reliability.

Key Words: **econometric model, linear regression, statistics, univariate regression, multivariate regression, OLS estimators, matrix form, hypothesis testing.**

1. INTRODUCTION

Linear regression is a powerful statistical modeling technique widely used to analyze the relationship between a dependent variable and one or more independent variables. In a simple linear regression model, the dependent variable is assumed to be a linear function of one independent variable, while in a multiple linear regression model, the dependent variable is a linear function of two or more independent variables.

The coefficients of the linear regression model are represented by beta (β) values, which are constants that determine the slope and intercept of the regression line. The beta values are estimated using the Ordinary Least Squares (OLS) method, which involves minimizing the sum of squared errors between the predicted values and the actual values of the dependent variable.

The matrix form of the linear regression model is a useful tool for understanding the relationship between the dependent and independent variables. The matrix form allows for a more efficient calculation of the beta values and the variance-covariance matrix of the beta values.

Partial derivatives are used to derive the expected values and variation of the beta values. The expected values of the beta values are equal to the true beta values, and the variation of the beta values can be used to derive confidence intervals and hypothesis tests for the beta values.

The relationship between the beta values and the normal distribution is important in understanding the properties of the OLS estimators. The beta values are normally distributed, and their variance-covariance matrix can be used to derive confidence intervals and hypothesis tests.

Hypothesis testing is an essential component of linear regression analysis. The null hypothesis is typically that the beta value is equal to zero, indicating that there is no relationship between the independent variable and the dependent variable. The t-test and F-test are commonly used to test hypotheses about the beta values.

The ANOVA matrix is a useful tool for decomposing the total variation in the dependent variable into the explained variation (sum of squared regression) and unexplained variation (sum of squared error). The sum of squared total is the sum of squared deviations of the dependent variable from its mean.

The relationship between the sum of squared regression, sum of squared error, and sum of squared total with the mean squared error and the degrees of freedom is important in interpreting the results of the linear regression analysis.

In summary, linear regression is a powerful statistical modeling technique that can be used to analyze the relationship between a dependent variable and one or more independent variables. The beta values, matrix form of the model, partial derivative concept, expected values and variation of beta values, relationship with normal distribution, hypothesis testing, ANOVA matrix, formulas for sum of squared regression, sum of squared error, sum of squared total, and relationship with mean squared error, t-test, and F-test with degrees of freedom are important concepts in understanding and interpreting the results of linear regression analysis.

2. METHODS

2.1 Deriving the coefficient of the regression model for the simple linear regression model

The important result that is needed to derive $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Suppose, $\bar{y}$ is the sample mean of y then $\bar{y}$ is defined as 1 over n times the summation from i = 1 to n of y_i

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

It also means that $n\bar{y}$ is equal to the summation from

i = 1 to n of y_i

$$n\bar{y} = \sum_{i=1}^{n} y_i \text{ ------------- (a)}$$

Similarly,

$\bar{x}$ is the sample mean of y and is defined as 1 over n times the summation from i = 1 to n of x_i

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It also means that $n\bar{x}$ is equal to the summation from

i = 1 to n of x_i.

$$n\bar{x} = \sum_{i=1}^{n} x_i \text{ ------------ (b)}$$

There are some important results that we need to know using the summation operator,

$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) =$	The written L.H.S. can also be written as: $\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ -------- (1) $\sum_{i=1}^{n} (x_i - \bar{x}) y_i$ -------- (2) $\sum_{i=1}^{n} (y_i - \bar{y}) x_i$ -------- (3)

Table 1: Different forms of $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ [1]

Derive: $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$

L.H.S. = $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$= \sum_{i=1}^{n} (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y})$

$= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$

From equation (a) and equation (b)

$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + \bar{y}n\bar{x} - \bar{y}n\bar{x}$

$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$

= R.H.S [Equation (1)]

Derive: $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$

L.H.S. = $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

$= \sum_{i=1}^{n} (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y})$

$= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$

From equation (b)

$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$

$= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i$

$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i$

= R.H.S [Equation (2)]

Derive: $\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y}) = \sum_{i=1}^{n} (y_i - \underline{y})\, x_i$

L.H.S. = $\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})$

= $\sum_{i=1}^{n} (x_i\, y_i - \underline{y}\, x_i - \underline{x}\, y_i + \underline{x}\, \underline{y})$

= $\sum_{i=1}^{n} x_i\, y_i - \underline{y} \sum_{i=1}^{n} x_i - \underline{x} \sum_{i=1}^{n} y_i + n\, \underline{x}\, \underline{y}$

From equation (a)

= $\sum_{i=1}^{n} x_i\, y_i - \underline{y} \sum_{i=1}^{n} x_i - n\, \underline{x}\, \underline{y} + n\, \underline{x}\, \underline{y}$

= $\sum_{i=1}^{n} x_i\, y_i - \underline{y} \sum_{i=1}^{n} x_i$

= $\sum_{i=1}^{n} (y_i - \underline{y})\, x_i$

= R.H.S. [Equation (3)]

Deriving the formula for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ [2]

1. S.S.E: The sum of squared error is equal to the summation from i = 1 to n of the square of the difference between y_i and $\widehat{y_i}$

2. y_i is the real observation and $\widehat{y_i}$ is the predicted observation

3. The goal is to minimize the distance between the real observation and the predicted observation.

4. The equation is a quadratic equation and the objective is to find the minimum point and not the maximum point.

S.S.E = $\sum_{i=1}^{n} \left(y_i - \widehat{y_i} \right)^2$

substituting $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$

S.S.E = $\sum_{i=1}^{n} \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i) \right)^2$

Applying Partial Derivative with respect to $\widehat{\beta_0}$

$\frac{\delta S.S.E}{\delta \widehat{\beta_0}} = 2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right) (-1)$

To find the maxima or minima,

Equivalently, the derivative equals 0.

$0 = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right)$

$0 = \sum_{i=1}^{n} y_i - n\, \widehat{\beta_0} - \widehat{\beta_1} \sum_{i=1}^{n} x_i$

from equation (a) and equation (b)

$0 = n\, \underline{y} - n\, \widehat{\beta_0} - n\, \widehat{\beta_1}\, \underline{x}$

$0 = \underline{y} - \widehat{\beta_0} - \widehat{\beta_1}\, \underline{x}$

$\widehat{\beta_0} = \underline{y} - \widehat{\beta_1}\, \underline{x}$

Applying Partial Derivative with respect to $\widehat{\beta_1}$

$\frac{\delta S.S.E}{\delta \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right) (x_i)$

To find the maxima or minima,

Equivalently, the derivative equals 0.

$0 = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right) (x_i)$

$0 = \sum_{i=1}^{n} x_i\, y_i - \widehat{\beta_0} \sum_{i=1}^{n} x_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i^2$

From equation (b)

$0 = \sum_{i=1}^{n} x_i\, y_i - \widehat{\beta_0}\, n\, \underline{x} - \widehat{\beta_1} \sum_{i=1}^{n} x_i^2$

As $\widehat{\beta_0} = \underline{y} - \widehat{\beta_1}\, \underline{x}$

$0 = \sum_{i=1}^{n} x_i\, y_i - [\underline{y} - \widehat{\beta_1}\, \underline{x}]\, n\, \underline{x} - \widehat{\beta_1} \sum_{i=1}^{n} x_i^2$

$0 = \sum_{i=1}^{n} x_i\, y_i - n\, \underline{x}\, \underline{y} + \widehat{\beta_1}\, n\, \underline{x}^2 - \widehat{\beta_1} \sum_{i=1}^{n} x_i^2$

From equation (1)

$0 = \sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y}) + \widehat{\beta_1} (n\, \underline{x}^2 - \sum_{i=1}^{n} x_i^2)$

$-\widehat{\beta_1} (n\, \underline{x}^2 - \sum_{i=1}^{n} x_i^2) = \sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})$

$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})}{(\sum_{i=1}^{n} x_i^2 - n\, \underline{x}^2)}$

As $\sum_{i=1}^{n} (x_i - \underline{x})^2 = (\sum_{i=1}^{n} x_i^2 - n\, \underline{x}^2)$

$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})}{\sum_{i=1}^{n} (x_i - \underline{x})^2}$

2.2 Matrix form of the Simple Linear Regression Model

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x & x & . & . & . & x_n \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 . & . & . & . & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Here, $X^T X$ is a 2 x 2 symmetric matrix

$\det(X^T X) = n\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i * \sum_{i=1}^n x_i$

$= n\sum_{i=1}^n x_i^2 - [\sum_{i=1}^n x_i]^2$

$= n\sum_{i=1}^n x_i^2 - n^2 \underline{x}^2$

$= n(\sum_{i=1}^n x_i^2 - n \underline{x}^2)$

$S_{xx} = (\sum_{i=1}^n x_i - n \underline{x}^2) = \sum_{i=1}^n (x_i - \underline{x})^2$

$\det(X^T X) = n\sum_{i=1}^n (x_i - \underline{x})^2 = nS_{xx}$

$$(X^T X)^{-1} = \frac{1}{det(X^T X)} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x & x & . & . & . & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_1 \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

$$Y = X\widehat{\beta}$$

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_1 \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Scalar form of Y is

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

For Y, β_0 the intercept, β_1 is the slope, and ϵ_i is the error term epsilon. This is known as the scalar form. Suppose there are n data points, then i ranges from 1 to n

The Matrix Form of Y is

$Y = X\beta + \epsilon$[3]

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & x_{31} & x_{32} & . & . & x_{3p} \\ . & x_{41} & x_{42} & . & . & x_{4p} \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_0 \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

2.3 SLR Matrix Form Proof of Properties of OLS(Ordinary Least Squares) Estimators

The first assumption is that the expected value of epsilon is equal to zero for all i and these i are in the range of i = 1, 2,..., n, where n is the number of data points.

The mathematical representation is stated below.

$E[\varepsilon_i] = 0 \;\forall$ i, where i = 1, 2, 3,..., n.

The anticipated value of vector epsilon is going to be zero vector or in rows, so it's essentially epsilon, the expected value of epsilon is going to be simply a vector of zero, where there will be n rows, So this is the first assumption that will be evaluated when working with the simple linear regression model in matrix form.

Expected value and Variance of ε [4]

$E[\varepsilon_i] = 0_n$

$$E[\varepsilon] = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 0_n \end{bmatrix}$$

The second assumption is about the variance of the epsilon

Var[] i. The epsilon vector's variance matrix is equal to sigma squared times the identity matrix, with dimensions of n by n.

Var[ε_i] = $\sigma^2 I_{n\times n}$

So the variance of epsilon is equal to sigma squared on all diagonal terms up to n rows, and all these entries should be zero other than diagonal terms.

Var[ε_i] = σ^2I

$$Var[\varepsilon] = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

The third assumption is that the epsilon values are uncorrelated. If they are uncorrelated then that means that for any Epsilon values ε_1 and ε_2, the variance of ε_1 and ε_2 is equal to zero. This also means that the covariances of ε_1 and ε_2 are equal to zero. Thus, $\forall i, j$ where i $\neq$ j, the covariance of ε_i and ε_j is equal to zero.

As ε_i's are uncorrelated

The mathematical representation of the above third assumption is:

Var($\varepsilon_1, \varepsilon_2$) = 0

Cov($\varepsilon_1, \varepsilon_2$) = 0

Cov($\varepsilon_i, \varepsilon_j$) = 0 $\forall i, j$ i $\neq j$

As the epsilon of i's(ε_i's) are uncorrelated with each other, and based on these three assumptions, when combined with the fourth assumption, allow for a very good powerful construction of the simple linear regression model, from which various results can be derived.

The fourth and last assumption is that the error terms are normally distributed.

The epsilon i(ε_i) follows the normal distribution with some mean(μ) and some variance(σ^2) .

$\varepsilon_i \sim$ N(μ, σ^2) $\forall i$ [5]

From Assumptions 1 and 2, we know that μ = E[ε_i] = 0_n and Var[ε_i] = $\sigma^2 I_{n\times n}$ where σ is the standard deviation. Also, from assumption 3 where $\forall$ i, where i = 1, 2, 3,………, n

Cov($\varepsilon_i, \varepsilon_j$) = 0 $\forall i, j$ i $\neq j$ which means these epsilon are independent and are identically distributed.

Hence,

$\varepsilon \sim$ N($0_n, \sigma^2 I_{n\times n}$)

Hence, $\hat{\varepsilon}$ = Y - $\hat{Y}$

$\hat{\varepsilon}$ = Y - $X\hat{\beta}$

Derive and show that E[$\hat{\beta}$] = β

As $\hat{\beta}$ = $(X^T X)^{-1} X^T Y$

E[$\hat{\beta}$] = E[$(X^T X)^{-1} X^T Y$]

Note: $(X^T X)^{-1} X^T$ is not stochastic, it's not random at all, it is all just constant known values that are multiplied by each other.

E[$\hat{\beta}$] = $(X^T X)^{-1} X^T$E[Y]

As, Y = Xβ + ε

E[Y] = E[Xβ + ε]

= E[Xβ] + E[ε]

= E[Xβ] ----(From Assumption 1, E[ε] = 0)

= Xβ

Hence,

E[$\hat{\beta}$] = $(X^T X)^{-1} X^T$E[Y]

= $(X^T X)^{-1}(X^T$X)β

As $(X^T X)^{-1}(X^T$X) = I_n where X^TX is symmetric matrix

E[$\hat{\beta}$] = $I_n \beta$

E[$\hat{\beta}$] = β

Derive and show that Var[β] = $\hat{\sigma}^2 (X^T X)^{-1}$

Var[$\hat{\beta}$] = Var[$(X^T X)^{-1} X^T Y$]

Let B = $(X^T X)^{-1} X^T$

Var[B Y]

= B Var[Y] B^T

= B Var[Xβ + ε] B^T = B Var[$X\beta$] B^T + B Var[ε] B^T

Here, $X\beta$ in Var[$X\beta$] is not random and continuous hence we can assume that Var[$X\beta$] = 0

From Assumption 2 where Var[ε_i] = σ^2I

Var[$\hat{\beta}$] = Bσ^2IB^T

= $(X^TX)^{-1}X^T\sigma^2\mathrm{I}[(X^TX)^{-1}X^T]^T$

As X^TX is symmetric, $[(X^TX)^{-1}X^T]^T$= X $[(X^TX)^{-1}]^T$

= X $[(X^TX)^{-1}]$

Var[$\hat{\beta}$] = $(X^TX)^{-1}X^T\sigma^2\mathrm{I}[(X^TX)^{-1}X^T]^T$

= $(X^TX)^{-1}X^T\sigma^2$I X $[(X^TX)^{-1}]$

= $\sigma^2I(X^TX)^{-1}X^TX(X^TX)^{-1}$

As $(X^TX)^{-1}(X^T$X) = I_n where X^TX is symmetric matrix

= $\sigma^2\mathrm{I}(X^TX)^{-1}$

= $\sigma^2(X^TX)^{-1}$

As are unobservable, σ^2 cannot be computed hence we assume $\widehat{\sigma^2}$is an estimator of σ^2

Hence, Var[$\hat{\beta}$] = $\hat{\sigma}^2(X^TX)^{-1}$

2.4 Hypothesis testing for Multiple Linear Regression- Matrix form

Y = Xβ + ε

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & x_{31} & x_{32} & . & . & x_{3p} \\ . & x_{41} & x_{42} & . & . & x_{4p} \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_0 \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

here, $\hat{\varepsilon}$ = Y - $\hat{Y}$ where Y above in the equation is the actual observation and $\hat{Y}$ is the fitted observation.

The following equation gives the sum of squared errors(SSE):

SSE = $\hat{\varepsilon}^T\ \hat{\varepsilon}$

Therefore, the above function can be written as:

SSE = $(Y - X\hat{\beta})^T(Y - X\hat{\beta})$

To fit the regression model, the goal should be to minimize the SSE,

minimize(SSE) = minimize($\hat{\varepsilon}^T\ \hat{\varepsilon}$)

= minimize$[(Y - X\hat{\beta})^T(Y - X\hat{\beta})]$

Here, $\hat{\beta}$ = $(X^TX)^{-1}X^TY$

where

$$X^TX = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & . & . & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & . & . & \sum x_{i1}x_{ip} \\ . & \sum x_{i1}x_{i2} & \sum x_{i1}^2 & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sum x_{ip} & . & . & . & . & \sum x_{ip}^2 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ . \\ . \\ . \\ \sum x_{ip}y_i \end{bmatrix}$$

E[$\hat{\beta}$] = β

Var[$\hat{\beta}$]=$\hat{\sigma}^2(X^TX)^{-1}$

$\widehat{Var[\hat{\beta}]}$=$\hat{\sigma}^2(X^TX)^{-1}$

σ^2 = MSE = $\frac{SSE}{n-p-1}$ = $\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-p-1}$

Here, p regressors of interest 1 is for β_0 and n is the number of samples in the population

Let Se be the Standard Error

Hence,

Se($\hat{\beta}$) = $\sqrt{\widehat{Var[\hat{\beta}]}}$ = $\sqrt{\hat{\sigma}^2(X^TX)^{-1}}$

Here, Se($\hat{\beta}_1$) = $\sqrt{\hat{\sigma}^2{(X^TX)^{-1}}_{11}}$

Hypothesis to be tested: H_0: $\widehat{\beta}_j = 0$ vs H_1: $\widehat{\beta}_j \neq 0$ **[6]**

Test Statistic $= \frac{\widehat{\beta}_j - \widehat{\beta}_{jo}}{Se(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1 - 0}{Se(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2 (X^T X)^{-1}{}_{11}}}$

Hypothesis testing on a linear combination of regression parameters in the matrix form of the linear regression problem.

Below is the general formula of the regression model that is in matrix form:

$Y = X\beta + \varepsilon$

Scalar representation of the regression parameters:

$Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots\ldots\ldots + \beta_p x_{ip}$.

Linear combination:

Let $\widehat{L} = c_0\beta_0 + c_1\beta_1 + c_2\beta_2 + \ldots\ldots\ldots\ldots + c_p\beta_p$.

$$C = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ . \\ . \\ . \\ c_p \end{bmatrix}$$

The $c_0, c_1, c_2, \ldots, c_p$ in the above equation represent the constants.

$$\widehat{L} = c^T \widehat{\beta}$$

$$\hat{L} = \begin{bmatrix} c_0 & c_1 & . & . & . & c_p \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ . \\ \hat{\beta}_p \end{bmatrix}$$

$$\hat{L} = \begin{bmatrix} c_0\beta_0 & c_1\beta_1 & . & . & . & c_{p\beta_p} \end{bmatrix}$$

Hypothesis to be tested: H_0: $\hat{L} = L_0$ vs H_1: $\hat{L} \neq L_0$

Expected value of $\widehat{L}$:

$E[\widehat{L}] = E[\widehat{L}] = [c^T \widehat{\beta}] = c^T E[\widehat{\beta}] = c^T \beta = L$

c^T is the row vector of the constants.

Variance of $\widehat{L}$:

$Var[\widehat{L}] = Var[c^T \beta] = c^T Var[\beta] c = c^T \widehat{\sigma}^2 (X^T X)^{-1} c$

βFormulation of the test statistics is a t-test where the parameter of interest ($\widehat{L}$) follows the normal distribution with mean of L and variance of $c^T \widehat{\sigma}^2 (X^T X)^{-1}$c.

$\widehat{L} \sim N(L, c^T \widehat{\sigma}^2 (X^T X)^{-1} c)$

Test statistics T:

$T = \frac{\widehat{L} - L_0}{Se(\widehat{L})}$

L_0 : It is the value of the linear combination under the null hypothesis

where $Se(\widehat{L}) = \sqrt{\widehat{\sigma}^2 c^T (X^T X)^{-1} c}$

Then test statistics becomes,

$T = \frac{c^T \beta - L_0}{\sqrt{\widehat{\sigma}^2 c^T (X^T X)^{-1} c}} \sim t_{n-p-1}$

ANOVA Matrix Form [7]:

SSR : Sum of squares Regression : $\sum_{i=1}^{n} \left(\widehat{y_i} - \underline{y_i}\right)^2$

SSE : Sum of squares Errors : $\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$

SST : Sum of Squared Total : $\sum_{i=1}^{n} \left(y_i - \underline{y_i}\right)^2$

Source of variation	Sum of Squares	degrees of freedom	Mean Square	F statistic
Regresion	SSR	p	MSE = $\frac{SSR}{P}$	F = $\frac{MSR}{MSE}$
Errors	SSE	n-p-1	MSE = $\frac{SSR}{n-p-1} = \widehat{\sigma}^2$	

Total	SST = SSR + SSE	n-1	MST = $\frac{SST}{n-1}$	

Table 2: ANOVA matrix form [8]

F = $\frac{MSR}{MSE} \sim F_{p,n-p-1}$

Hypothesis to be tested: H_0:F = β_0=β_1=β_2=...=β_p

vs H_1: F≠ H_0 **[9]**

f F > $F_{\alpha,p,n-p-1}$ then reject H_0 else accept H_0

3. CONCLUSIONS

In conclusion, this research paper has explored the econometric modeling approach of linear regression using statistics. The paper has presented the fundamental concepts of linear regression, such as the basic assumptions and the definition of the dependent and independent variables.

The paper has then discussed the univariate and multivariate linear regression models, and the coefficients of the regression models have been derived using statistical methods. The matrix form of the Simple Linear Regression Model has been presented, and the properties of the Ordinary Least Squares (OLS) estimators have been proven.

Finally, the paper has discussed hypothesis testing for multiple linear regression in the matrix form. The importance of understanding the econometric modeling approach of linear regression using statistics has been emphasized, and the paper's findings contribute to the understanding of the practical application of the linear regression model.

Overall, linear regression is a powerful tool for predicting the values of the dependent variable based on the values of the independent variables. It can be applied in various fields, including economics, finance, and social sciences. Rigorous statistical analysis is necessary to ensure the model's validity and reliability, and this paper has provided insights into the statistical methods used in linear regression modeling.

REFERENCES

[1] Introduction to Linear Regression Analysis, 5th Edition" by Montgomery, Peck, and Vining

[2] Applied Linear Regression, 3rd Edition" by Weisberg

[3] Matrix Approach to Simple Linear Regression" by Kato and Hayakawa

[4] Linear Regression Analysis: Theory and Computing" by Liu and Wu

[5] Linear Models with R, 2nd Edition" by Faraway

[6] Applied Regression Analysis and Generalized Linear Models, 3rd Edition" by Fox

[7] Statistical Models: Theory and Practice" by Freedman, Pisani, and Purves

[8] Statistical Inference" by Casella and Berger

[9] Probability and Statistics for Engineering and the Sciences, 9th Edition" by Devore