JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

Formulating a Linear Model from a One-Way Classification Model

Badmus, Nofiu Idowu[1*] and Ogundeji, Rotimi Kayode[2]

[1&2]*Department of Statistics, Faculty of Science, University of Lagos, Akoka, Nigeria*

**Corresponding author: nibadmus@unilag.edu.ng*

Received 28 May 2024
Accepted 17 Oct 2024
Published

RESEARCH ARTICLE

Abstract

This study introduces a novel approach to formulating a linear regression model using a matrix method for Completely Randomized Design (CRD), a type of One-Way classification. In this approach, treatment is the sole classification, and the formulation utilizes response variables organized into rows and columns. The method yields the number of trials (n), slope, predictor, and regression parameters within the system. To ensure the normality of the response variable and select the appropriate error term distribution, we conducted normality tests (Shapiro-Wilk, Anderson-Darling, Cramér-von Mises, Lilliefors) and exploratory data analysis techniques (histogram, boxplot, QQ-plot). The formulation was validated through illustrations, and the results from the matrix method regression were compared to the ordinary least squares regression, yielding identical values for the regressors, and confirming the robustness of the proposed formulation. Furthermore, we evaluated the performance of machine learning linear regression model, which outperformed ordinary least squares regression in terms of mean absolute error, mean square error, and root mean square error, demonstrating the superior accuracy of the proposed approach.

Keywords: Anderson Darling, Boxplot, Machine Learning, Normality Test, Regression, Treatment

1. Introduction

A completely randomized design (CRD) is a type of experimental design where subjects are randomly assigned to different treatment groups. This study explores methods for formulating a linear model:

(a) One-way Analysis of variance (ANOVA)

Response variable: y (continuous), Predictor variable: treatment (categorical)

- Model is

$$y = \mu + \tau_i + \varepsilon$$

where μ, τ_i and ε are: the overall mean, the effect of the ith treatment and the error term.

(b) Multiple Regression

Response variable: y (continuous), Predictor variables: $X_1, X_2, \ldots, X_k$ (continuous or categorical)

- The Model

$$y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 +, \ldots, \gamma_k X_k + \varepsilon$$

where γ_0 is the intercept, $\gamma_1, \gamma_2, \ldots, \gamma_k$ are the regression coefficients, and ε is the random error.

(c) Analysis of Covariance (ANCOVA)

Response variable: y (continuous), Predictor variables: treatment (categorical), covariate (continuous)

- The Model

$$y = \mu + \tau_i + \gamma X + \varepsilon$$

where μ, τ_i, γ and ε are: the overall mean, the effect of the ith treatment, the regression coefficient for the covariate, and the error term.

(d) Regression with Interaction:

Response variable: y (continuous) and Predictor variables: x, x_2, (continuous or categorical)

- The Model

$$y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 +, \ldots, \gamma_3 X_1 X_2 + \varepsilon$$

where γ_0 is the intercept, γ_1, γ_2 and γ_3 are the regression coefficients, and ε is the random error (Ali & Younas 2021; Schober & Vetter 2021; Ratkovic 2023; Jenkins & Quintana-Ascencio 2020; Ahmad et al., 2023) and so on.

2. Materials and Methods

2.1 Model Formulation

2.1.1 One-Way Classification

The method of formulating a linear model is done from data obtaining through a completely randomized design (CRD) involving $k = 2$ treatments. (Wackerly et al., 2008). The study focuses on two treatments to provide a clear, interpretable framework for investigating treatment effects. While this limitation affects the model's general applicability, it allows for a detailed examination of the two-treatment comparison. The model is given by

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (1)$$

where, y_{ij} is the response variable, μ is the overall mean, τ_i is the treatment effect and ε_{ij} is the error term. Since y_{ij} is observed on the jth observation from treatment i, say i = 1, 2., using an indicator or coding (1, 0), variable x is defined as

$$x = \begin{cases} 1, & \text{if the observation comes from population 1} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

2.2 Linear Model

This model is often called simple linear equation and its model is given by

$$y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij} \tag{3}$$

where y_{ij}, $\beta's$, x and ε_{ij} are: dependent variable, intercept and regression parameter, independent variable and $\varepsilon_{ij} \sim N(0, \sigma^2)$ → (random error). (Rencher & Schaalje 2008; Flatt & Jacobs 2019; Knief & Forstmeier 2021; Taherdoost 2022).

$$\alpha_1 = E(y_{ij}) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

and

$$\alpha_2 = E(y_{ij}) = \beta_0 + \beta_1(0) = \beta_0$$

Its then follow that

$$\beta_1 = \alpha_1 - \alpha_2$$

and a test of hypothesis $\alpha_1 - \alpha_2$, this tantamount to test that $\alpha_1 = 0$. This may be written as

$$\bar{\alpha}_0 = \bar{Y}_2$$

and

$$\bar{\alpha}_1 = \bar{Y}_{1.} - \bar{Y}_{2.}$$

are good estimators of α_0 and α_1.

3. Illustration 1 (Matrix Method)

The data used for the illustration contains three different machines P, Q, and R, and a manufacturing company wishes to acquire one of the machines. Four experienced operations workers (as treatment) were assigned to work on each machine for equal periods. Each machine was given identical tasks to perform. The experiment lasted for a predetermined period, ensuring each worker-machine combination was tested for equal duration. The essence is to test whether there is a difference in the machines' performance.

3.1 Table 1. Number of Units Produced Per Machine

Table 1. Number of Units Produced Per Machine

Machine	P	Q	R
I	5	7	9
II	7	5	8
III	7	6	4
IV	6	6	2

Fit an appropriate linear model to the information above and test whether there is significant difference between the machines ($\alpha_1, \alpha_2, \alpha_3$ and α_4). It is a complete randomize design because it consists only treatment.

3.2 Normality Test of Machine Performance

Here, we investigate whether the observations come from the same population by testing their normality. To determine if the observations are normally distributed, we employed three statistical tests: the Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test and Lilliefors test. The results of these normality tests are presented in Table 2 below.

Table 2. Normality Test For the observations

Test	Value	*p*-value
Shapiro Wilk	W = 0.96054	0.7915
Anderson Darling	A = 0.28782	0.5550
Cramer von Mises	W = 0.04886	0.4984
Lillifors	D = 0.16667	0.4724

Table 2 presents the results of normality tests (Shapiro-Wilk, Anderson-Darling, Cramer-von Mises, and Lilliefors) for the observations. Since all p-values exceed 0.05, the data is assumed to follow a normal distribution, with no significant deviations from normality detected.

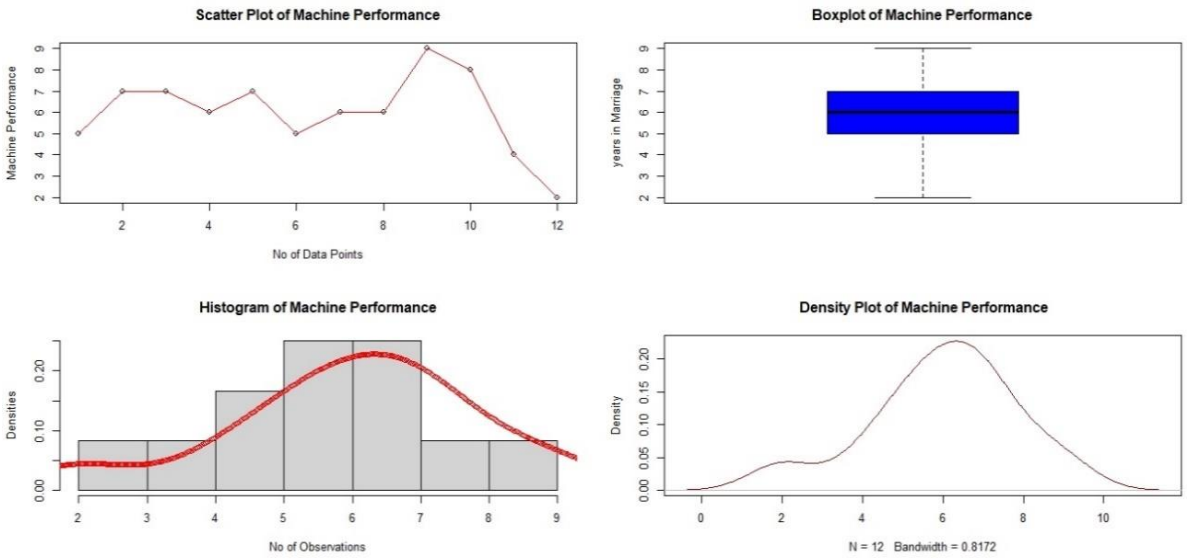

Figure 1. The Scatter plot, Boxplot, Histogram and Density Plot of Dataset

In Figure 1, the scatter plot shows the relationship between the variables, while the histogram and boxplot illustrate the data's distribution and outliers. Also, the density plot depicts the movement of the data and whether the data is skewed or not. But this suggests that the data likely follows normality.

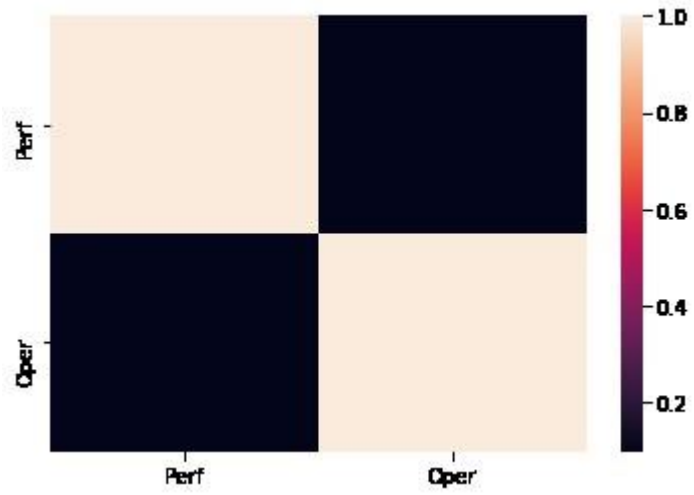

Figure 2. The Correlation Plot of Machine Performance (y) and Operation

Figure 2 depicts the correlation plot illustrating the positive relationship between operations and machine performance. The intensity of the colours represents the strength of the correlation, with darker colors indicating stronger correlations and lighter colors indicating weaker correlations. Notably, the plot reveals a significant positive correlation ($r = 0.8$), indicating that machine performance also improves as the number of operations increases.

To fit the linear model, we recall the equation in (3) as

$$y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij}$$

where

$$x = \begin{cases} 1, & \text{if the operation comes from machine 1} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The matrices used for the least-squares estimators are given by

$$y = \begin{pmatrix} 5 \\ 7 \\ 7 \\ 6 \\ 7 \\ 5 \\ 6 \\ 6 \\ 9 \\ 8 \\ 4 \\ 2 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 12 & 4 \\ 4 & 4 \end{pmatrix} . \frac{1}{det} . (X'X) = \begin{pmatrix} \frac{3}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} \end{pmatrix}$$

The least square estimates are given by

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{5}$$

where

$$(X'X)^{-1} = \begin{pmatrix} 0.125 & -0.125 \\ -0.125 & 0.375 \end{pmatrix}, (X'Y) = \begin{pmatrix} 72 \\ 25 \end{pmatrix}$$

The regression line equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{6}$$

Equation (5) can be written as

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 = & 5.875 \\ \hat{\beta}_1 = & 0.375 \end{pmatrix}$$

Meanwhile, equation (6) becomes

$$\hat{y} = 5.875 + 0.375 x_i \tag{7}$$

Equation 7 explains the relationship between treatment and machine performance. and x_i For every additional treatment (operation work), machine performance increases by 0.375 units. When no treatment is applied ($x = 0$), the baseline machine performance is 5.875. Therefore, the equation indicates a positive and linear relationship between the treatment and machine performance.
Recall that $\hat{\beta}_0 = 5.875 = \bar{Y}_2$ and $\hat{\beta}_1 = 0.375 = \bar{Y}_1 - \bar{Y}_2$.

Furthermore, we compute the sum of squares error using equation (8) below:

$$SSE = Y'Y - \hat{\beta}X'Y \tag{8}$$

where,

$$Y'Y = (5 \quad 7 \quad 7 \quad 6 \quad 7 \quad 5 \quad 6 \quad 6 \quad 9 \quad 8 \quad 4 \quad 2).\begin{pmatrix}5\\7\\7\\6\\7\\5\\6\\6\\9\\8\\4\\2\end{pmatrix} = 470$$

$$\hat{\beta}X'Y = (5.875 \quad 0.375).\begin{pmatrix}72\\25\end{pmatrix} = (432.375)$$

$$SSE = 37.625$$

Thus,

$$S^2 = \frac{SSE}{n-k} \tag{9}$$

where $n = 12$ (number of trials) and $k = 2$ (number of variables).

$$S^2 = 3.7625$$

and

$$S = \sqrt{3.7625} = 1.9397$$

To test H_0: $\beta_1 = 0$, we compute the $t - statistic$, we have

$$t = \frac{\widehat{\beta}_1 = 0}{S\sqrt{E_{22}}} \tag{10}$$

where, $E_{22} = \frac{3}{8}$. This implies that

$$t = 0.3150$$

4. Illustration 2 (Simple Regression Model)

This model often called "Simple Bivariate Regression model or least square regression model" due to its nature. It can only contains two variables such as the response/dependent variable and one predictor/independent variable. The model is

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i \tag{11}$$

where: y is the machine performance β_0 is the intercept, β is the regression coefficient, x is the operation and ε is the error term. However, we employed the method of least squares regression using R codes to estimate the intercept and parameter coefficient of the model. The results are presented in Table 3 below:

Table 3. Estimation of Coefficient, Standard Error, t-value and Probability

Coefficient	Estimate	Std Error	t - value	Pr
Intercept (β_0)	5.8750	0.6858	8.567	0.43e-06**
Operation (β_1)	0.3750	1.1878	0.316	0.759

Interpretation

$$\hat{y} = 5.875 + 0.375 operation(x_i) \quad (12)$$

Table 3 is derived from the analysis of data in Table 1 (the number of units produced per machine).

Table 4. Comparison Between Matrix and Least square Method

Coefficient	Matrix Method	Least Square Method
Intercept (β_0)	5.8750	0.3750
Operation (β_1)	5.8750	0.3750
T - value	0.3160	0.3160

The equivalence of equations (7) and (12) validates the formation of the linear model from the One-way classification model. Notably, the outcomes from both the matrix method and the least squares method are identical, as presented in Table 4, further confirming the accuracy of the linear model formulation.

4.3 Evaluation of Model Performance

A comparison of model performance between the machine learning linear regression model and the least squares regression model was conducted using various model selection criteria, including mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). The results, presented in Table 5 below, reveal that the machine learning model outperforms the least squares model, with smaller error values across all criteria. This suggests that the machine learning approach is preferable to the least squares method for this particular dataset. It implies that machine learning produces lower value in all the errors considered from a linear regression model implemented through a machine learning library. The reasons are: machine learning can capture non-linear relationships more effectively, reducing error, does not rely solely on assumptions, unlike OLS which assumes linearity, normality, and homoscedasticity. It can handle feature interactions and better handling of complex data patterns whereas ordinary least square often struggle with these aspects.

Table 5. Evaluation of Model Performance

Model	Intercept	β	MAE	MSE	RMSE
Machine Learning	5.3333	0.6667	1.3333	2.4999	0.5811
Least Squares	5.8750	0.3750	6.0000	36.0313	6.0026

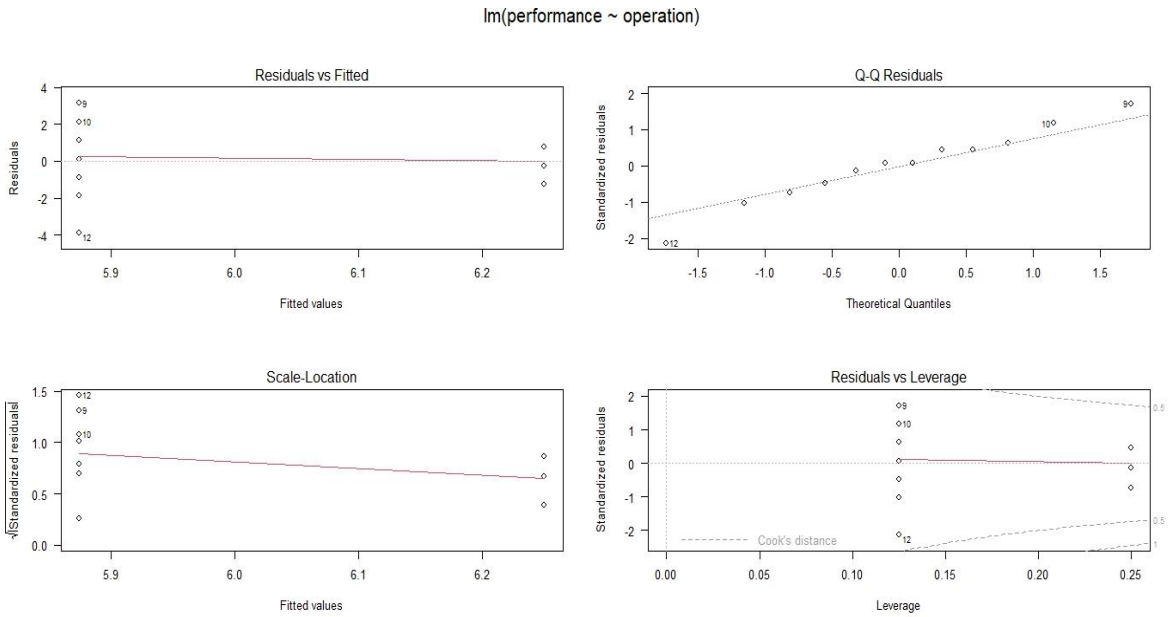

Figure 3. The Leverage Plot of Dataset

This plot is used to identify data points with high leverage (influential points) and data points with large residuals (outliers) and check for non-random trends in the residuals. However, there are no high-leverage data points that may influence the regression line, outliers to indicate unusual observations, and non-random trends that may suggest non-linear relationships.

5. Conclusion

We successfully formulated a linear model from a one-way classification model by employing a coding method, which led to the matrix method, to generate the predictor variable (x) since the response variable (y) is determined by the number of units produced per machine. Consequently, we obtained the linear regression line in equation (7) and the least squares estimates in equation (12), which yield identical outputs. Additionally, we calculated the sum of squared errors (SSE), variance S^2 standard deviation (S), and t-statistic due to the number of units per machine $(n = 12)$. Furthermore, we evaluated the performance of both the least squares model and the machine learning regression model, revealing that machine learning outperformed the least squares model. This is because machine learning does not assume a specific distribution of the error term; instead, it trains and tests the dataset, leading to improved accuracy. In addition, machine learning does not rely on strict assumptions (e.g., normality of error terms) and can better adapt to different data patterns, meanwhile, ordinary least squares do struggle with these features.

6. Future Study/Work

This study can be extended to a Two-Way classification model using a Complete Randomized Block Design (CRBD) to investigate interactions between treatments and blocks, accounting for more complex data structures.

7. Acknowledgements

We acknowledge those who have contributed to the success of this manuscript.

8. References

Ali, P. & Younas, A. (2021). Understanding and Interpreting Regression Analysis. *Evidence Based Nursing, 24*(4), 116 – 117.

Flatt, C & Jacobs, R. L. (2019). Principle Assumptions of Regression Analysis: Testing, Techniques, and Statistical reporting of Imperfect Data Sets. *Advances in Developing Human Resources, 21*(4), 484 – 502.

Jenkins, D. G & Quintana-Ascencio, P. F. (2020). A Solution to minimum sample size for regression. *PLOS ONE, 15*(2), 1-15.

Knief, U & Forstmeier, W. (2021). Violating the normality Assumptions may be the Lesser of Two Evils. *Behaviour Research Methods, 53*, 2576 - 2590.

Ratkovic, M. (2023). Relaxing Assumptions, Improving Inference: Learning and the Linear Regression. *American Political Science Review, 117*(3), 1053 - 1069.

Rencher, A & Schaalje, G. B. (2008). *Linear Models in Statistics (2nd Ed.).* A John Wiley & Sons, Inc. Publication.

Schober, P & Vetter, T. R. (2021). Linear Regression in Medical Research. *Anesthesia & Analgesia, 132*(1), 108-109.

Taherdoost, H. (2022). Different Types of Data Analysis, Data Analysis Methods and Techniques in Research Project. *International Journal of Academic Research in Management (IJARM), 9*(1), 1 - 9.

Yunus, R. M. & Khan, S. (2010). M-Test for Intercept After Pre-Testing on Slope. *Journal of Statistical Modelling and Analytics, 1*(1), 45 – 47.