Tikrit Journal of Administration and Economics Sciences, (31/3/2022);Vol. 18, No. 57, Part (2): 505-519
Doi: www.doi.org/10.25130/tjaes.18.57.2.30

Tikrit Journal of Administration and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
ISSN: 1813-1719 (Print)

Regression Techniques for Analysis of Variance with Application to the Reduction on Serum Cholesterol Level

Lecturer Dr. Hanaw Ahmed Amin
College of Science
University of Sulaimanyah
hanaw.ammin@univsul.edu.iq

Abstract:

In this study both of linear regression model and analysis of variance (ANOVA) for completely randomized design (CRD) with equal replications n_i were used. In fact, it can describe ANOVA method as a special case of the regression model and access to same results with dummy variables which are encoded as an indicator of (0, and 1) and then used directly for the solution of the linear system described by a linear regression. To prove and validate this fact regression model and one-way analysis of variance (ANOVA) were applied to the field of biological experiment diets effect to the reduction on serum cholesterol level (in mg/100 ml) for 32 normal men to test whether the means of all diets are significantly different. Concluding that multiple regression analysis is similar to analysis of variance according to the equivalency of the results achieved from the two studied models. Since, the null hypothesis was accepted, there was insufficient evidence to show that the means are not equal. It can be said that the diets studied has no an effect on the serum cholesterol level in normal males.

Keywords: Regression Treatment one-way ANOVA, Complete Randomized Design, Shapiro-Wilk Test, Levene Statistic, Serum Cholesterol Level.

أسلوب الانحدار الخطي لتحليل التباين مع التطبيق على انخفاض نسبة الكوليسترول في الدم

م.د. هه ناو أحمد أمين
كلية العلوم
جامعة السليمانية

المستخلص:

تم في هذه الدراسة استخدام كل من نموذج الانحدار الخطي وتحليل التباين للتصميم العشوائي الكامل في حالة تساوي التكرارات، في الواقع يمكن التعامل مع نموذج تحليل التباين كحالة خاصة لنموذج الانحدار المتعدد والتوصل إلى نفس نتائج المطابقة في كلا الطريقتين. وتم اثبات ذلك بالتطبيق على تجربة بيولوجية اختيرت عينة عشوائية من الذكور بحجم (32) مشاهدة وتوزيعهم في أربع مجاميع بتكرارات متساوية، وتم قياس نسبة الكولسترول لديهم ولم يكونوا يعانوا

من مشكلة الكولسترول، مع تحديد برنامج غذائي لكل مجموعة، لمقارنة متوسطات المجاميع باستخدام التصميم العشوائي الكامل وتحليل الانحدار لتحديد مدى تأثير البرنامج الغذائي على انخفاض نسبة الكولسترول لديهم. نظرًا لقبول فرضية العدم، لا توجد أدلة كافية لإثبات أن متوسط الكولسترول غير متساوٍ. وهذا يعني انه يمكن القول بأن النظام الغذائي للمجاميع الاربعة لم يكن لها تأثير على انخفاض مستوى الكوليسترول لديهم، لأنه في الواقع لم يكن لديهم أية مشكلة صحية، وبنفس الوقت التوصل الى تكافؤ نتائج التحليل للنموذجين في الدراسة.
الكلمات المفتاحية: معامل الانحدار ذو العامل الواحد لتحليل التباين، التصميم التام التعشية، اختبار Shapiro-Wilk، الإحصاءة Levene، مستوى الكوليسترول في الدم.

1. Introduction:

The concept of experimental design was broadly explained (montgomery, 2014). Several researchers use the concept of multiple regression analysis and ANOVA in their statistical research analysis. Multiple regression analysis are frequently used in different aspects of life. (oberkirchner et al, 2010) uses multiple regression to analyses and develop a model for the effect of essential material and process parameters to weight and moisture content of impregnated papers. (Bajpai, 2013) analyses university model using multiple regression and ANCOVA and found very essential. (Syla, 2013)

Study the significance of active-employment programs on employment levels using multiple regression. (Everarda et al, 2005) uses multiple regression result to study the importance of engaging student to graphical user interface in teaching statistical courses. (Oswald, 2012) shows how viewing multiple regression results through multiple lenses can give a better assessment to the researchers. (Kelley et al, 2003) shows that in multiple regression obtaining accurate parameter contributes more than having statistical significancy. (Pazzani et al, 1981) shows how independent sign regression generate linear model that are almost accurate as multiple regression. (Ludlow, 2014)

Study suppressor variables and suppression effects in building regression model. (Moya-laraño et al, 2008) encourages ecological researchers to use partial regression in their studies [7, 9].

In addition, dummy variables are defined with two single values of (0 or 1).

They may be explanatory or outcome variables with various different usage and applications in many different fields like time series analysis with seasonality data, analysis of qualitative data, such as survey responses;

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

categorical data with different levels. In economic forecasting, bio-medical research, credit scoring, response modeling, and other fields [1, 2].

In the regression model, a dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept[9].

This kind of description in a linear regression model is convenient to describe subgroups of data that have different intercepts and/or slopes without the creation of separate models. In logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, and increases the stability and significance of the coefficients.

In addition to the direct benefits to statistical analysis, representing information in the form of dummy variables is makes it easier to turn the model into a decision too[9]. Synonyms for dummy variables are design variables [Hosmer and Lemeshow, 1989], Boolean indicators, and proxies [Kennedy, 1981]. Related concepts are binning [Tukey, 1977] or ranking, because belonging to a bin or rank could be formulated into a dummy variable. Bins or ranks can also function as sets and dummy variables can represent non-probabilistic set membership. Set theory is usually explained in texts on computer science or symbolic logic. See [Arbib, et. al., 1981] or [MacLane, 1986][11].

2. Analysis of Variance in Linear Model[12]:

From the mathematical point of view, linear regression and ANOVA are identical: both break down the total variance of the data into different "portions" and verify the equality of these "sub-variances" by means of a test ("F" Test). Analysis of variance can be handled by a general regression analysis routine, if the model is correctly identified[9].

2.1 Multiple Linear Regression Model[3, 9, 10]:

When the k^{th} of independent variables are used to predict the dependent variable the model to studied is of the form.

$$Y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{ik} + \epsilon_{ij} \quad \ldots(1)$$

This model is called a multiple regression model. Where $y_1, y_2, \ldots, y_n$ be n independent observation on $\boldsymbol{Y}$, x_{ij} is the independent

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

variable for the ith observation, $i = 1, 2, \dots, n$,and $\varepsilon_{ij}'s$ are identically independent error.

Thus, the n equations representing the linear equations can be rewritten in the matrix form as.

$$\underline{Y} = X\underline{\beta} + \underline{\in}$$

Fitting a straight line in matrix term for estimating of $\beta_i's$ is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{21} & X_{22} & \dots & X_{k2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}_{k\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

Where $\boldsymbol{Y}$ is a vector of observations Y, $\boldsymbol{X}$ is a matrix of independent variables, $\boldsymbol{\beta}$ is the vector of parameters to be estimated, $\boldsymbol{\varepsilon}$ is to be a vector of errors, and **1** is to be a vector of ones.

The least squares estimates of **$\boldsymbol{\beta}$'s** are given by:

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

Where $\widehat{\boldsymbol{\beta}}$ is the vector of estimates of β, provided that $(X'X)$ is nonsingular.

2.2 Completely Randomized Design Model[2]**:**

A Completely Randomized Design Model is probably the simplest experimental design, in terms of data analysis and convenience. The usual fixed effect of analysis of variance is[8]

$$Y_{ij} = \mu + t_i + \varepsilon_{ij}\ ; i = 1. \dots . k \ and \ j = . \dots n_k \ \dots(2)$$

Where Y_{ij} is the realized value of the random variable of jth response for ith, μ is the true value of mean, $t_i$is the ith treatment effect, and $\varepsilon_{ij}$is the random effect. The null hypothesis test for equality of treatments is.

$$H_0: t_i = 0$$

$$H_1: t_i \neq 0$$

2.3 Regression Treatment of the- One Way Model[9, 10]**:**

The model given in equation (2) involves the parameters $\mu, t_1, t_2, \dots, t_k$, a natural first step in regression approach is to write down the model.

Tikrit Journal of Administration and Economics Sciences, (31/3/2022);Vol. 18, No. 57, Part (2): 505-519
Doi: www.doi.org/10.25130/tjaes.18.57.2.30

$$Y = \mu x_0 + t_1 X_1 + t_2 X_2 + \cdots + t_k X_k + \varepsilon \ldots (3)$$

Consider the variables $X_i$that reproduce equation (2) is the use of dummies

$X_0 = 1$

and[1]

$$X_i = \begin{cases} 1, \ if \ the \ observed \ value \ of \ Y_{ij} \ belong \ to \ i^{th} \ traetment \\ 0, if \ the \ observed \ value \ of \ Y_{ij} \ belong \ to \ other \ traetment \end{cases}$$

The explanatory or indicator variable is a dummy variable which takes the values of (0 and 1) where the matrix model of analysis of variance is[11].

$$\begin{bmatrix} Y_{11} \\ . \\ . \\ . \\ Y_{1n} \\ . \\ . \\ . \\ Y_{kn} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ . & . & & & . \\ . & . & & \ldots & . \\ . & . & & & . \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ . & . & & & \\ . & . & 1 & \ldots & 1 \\ . & . & & & \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ t_1 \\ t_2 \\ . \\ . \\ . \\ t_k \end{bmatrix} + \begin{bmatrix} \mu \\ \varepsilon_{11} \\ \varepsilon_{12} \\ . \\ . \\ . \\ \varepsilon_{kn} \end{bmatrix} \ldots (4)$$

The natural equation for least square method estimation in the analysis of variance is:

$$\begin{bmatrix} Y.. \\ Y_1. \\ Y_2. \\ . \\ . \\ . \\ Y_k. \end{bmatrix} = \begin{bmatrix} nk & n & n & \ldots & n \\ n & n & 0 & \ldots & 0 \\ n & 0 & n & \ldots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ n & 0 & 0 & \ldots & n \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{t}_1 \\ \hat{t}_2 \\ . \\ . \\ . \\ \hat{t}_k \end{bmatrix} \ldots (5)$$

And this is consistent with:

$$(X'X)\underline{\hat{\beta}} = X'\underline{Y}$$

From this, the estimation relationship between regression and analysis of variance is formed as follows.

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

Where $\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ . \\ . \\ . \\ \hat{\beta}_{k-1} \end{bmatrix} = \begin{bmatrix} \bar{Y}_{k.} \\ \bar{Y}_{1.} - \bar{Y}_{k.} \\ \bar{Y}_{2.} - \bar{Y}_{k.} \\ \bar{Y}_{3.} - \bar{Y}_{k.} \\ . \\ . \\ . \\ \bar{Y}_{k-1} - \bar{Y}_{k} \end{bmatrix}$ …(6)

2.4 The One- Way Randomized Complete Design[3, 5]:

To test the null hypothesis of no differences between means treatment effects $\sum_{i=1}^{k} t_i = 0$ give us

$\mu_1 = \mu + t_1, \ \mu_2 = \mu + t_2, \ldots, \ \mu_k = \mu + t_k$

will be under the constrain of $\sum \hat{t}_i = 0$, then the estimated mean differences are computed by:

$$\hat{\mu} = \frac{Y..}{nk} = \bar{Y}.., \ \hat{t}_i = \frac{Y_i.}{n} - \frac{Y..}{nk} = \bar{Y}_i. - \bar{Y}.., \ i = 1,..,k$$

But in the case of unequal replications the statistical test $H_0: t_i = 0$ to the mean differences consuming least square methods under the condition $\sum n_i \hat{t}_i = 0$ is:

$$\hat{\mu} = \frac{Y..}{N} = \bar{Y}.., \ \hat{t}_i = \frac{Y_i.}{n_i} - \frac{Y..}{N} .$$

First the sum of square due to (uncorrected) regression is:

$$R(\mu, t_1, t_2, \ldots, t_k) = \hat{\mu} Y.. + \hat{t}_1 Y_1. + \hat{t}_2 .Y_2. + \ldots + \hat{t} Y_k .$$
$$= \bar{Y}..Y.. + (\bar{Y}1. + \bar{Y}..)Y_1. + (\bar{Y}2. + \bar{Y}..)Y_2. + \ldots + (\bar{Y}k. + \bar{Y}..)Y_k .$$
$$= \left(\frac{Y^2..}{nk} + \frac{\sum Y_{i.}^2}{n}\right) - \bar{Y}..\left(\sum_{i=1}^{k} Y_i.\right)$$
$$= \frac{\sum Y_{i.}^2}{n} + \frac{Y^2..}{nk} - \frac{Y^2..}{nk}$$
$$= \frac{\sum Y_{i.}^2}{n} = \hat{\beta}' X' \underline{Y} \ \ldots(7)$$

In other words, the model in the equation (7) represents the sum of the product of the $\hat{\mu}. \hat{t}_1. \ldots . \hat{t}_k$, assuming that the null hypothesis is true, it means that the mean of $\sum t_i = 0$ with the method of least squares estimation in equation (5) will be reduced to the:

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

$$\hat{\mu}=\frac{Y..}{nk} \text{ or } nk\,\hat{\mu}=Y..$$

$$R(\mu)=\left(\frac{Y..}{nk}\right)(Y..)=\frac{Y_{..}^2}{nk}$$

Where μ is known multiplying the right hand of equation (5) by $\hat{\mu}$, then the sum square due to $\hat{t}_1 . \ldots . \hat{t}_k$ due to (corrected) regression is:

$$R(t_1,t_2,\ldots,t_k \mid \mu)=R(\mu,t_1,t_2,..,t_k)-R(\mu)$$
$$=\frac{\sum Y_{i.}^2}{n}-\frac{Y_{..}^2}{nk}$$
$$=SSTrea..$$

With (k-1) degrees of freedom, the sum square due to residual is:

$$SSE=\left(\sum_{i=1}^{k}\sum_{j=1}^{n}Y_{ij}^2-\frac{Y_{..}^2}{nk}\right)-R(t_1,t_2,\ldots,t_k \mid \mu)$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}Y_{ij}^2-\frac{Y_{..}^2}{nk}-\frac{\sum_{i=1}^{k}Y_{i.}^2}{n}+\frac{Y_{..}^2}{nk}$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}Y_{ij}^2-\frac{\sum_{i=1}^{k}Y_{i.}^2}{n} \quad \ldots(8)$$

3. Testing Assumptions: Normality and Equal Variance[5]:

To test hypotheses about population parameters, we must assume that the population distribution of the variable being measured is normal in form with equal variance. The two-stage procedure depending on are Shapiro-Wilk test and Leven test.

3.1 Levene Test for Equality of Variances[4]:

Levene's test (Levene 1960) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene test can be used to verify that assumption:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: \sigma_i^2 \neq \sigma_j^2$$

Given a variable X with sample of size N divided into k subgroups, where N_i is the sample size of the ith subgroup, the Levene test statistic is defined as:

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2} \qquad \dots(9)$$

Where $\bar{Z}_{i.}$ are the group means of the Z_{ij} and $\bar{Z}_{..}$ is the overall mean of the Z_{ij} and

$$Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$$

Critical Region:

The Levene test rejects the hypothesis that the variances are equal if $W > F_{\alpha, k-1, N-k.}$ and $F_{\alpha,k-1,N-k}$ is the upper critical value of the F distribution with k-1 and N-k degrees of freedom at a significance level of α.

3.2 Shapiro–Wilk Normality Test [6]:

The Shapiro–Wilk goodness-of-fit test is used to determine if a random sample, $X_i, i = 1, 2, \dots, n$, is drawn from a normal Gaussian probability distribution with true mean and variance, μ and σ^2, respectively. That is, X~N(μ, σ^2). Thus, we wish to test the following hypothesis:

H_0: The random sample was drawn from normal population, N(μ, σ^2)

H_1: The random sample does not follow N(μ, σ^2).

To test this hypothesis, we use the Shapiro–Wilk test statistic, which is given by

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \dots(10)$$

where $x_{(i)}$ are the ordered sample values and a_i are constants that are generated by the expression.

$$(a_1, a_2, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}}$$

With $m = (m_1, m_1, \dots, m_1)^T$ being the expected values of the ordered statistics that are independent and identically distributed random variables that follow the standard normal, N(0,1), and V is the covariance matrix of the order statistics.

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

4. Research Objective:

Dummy variables are variables that take the values of only 0 or 1. They may be explanatory or outcome variables; however, the focus of this article is explanatory or independent variable construction and usage to demonstrate the possibility of using linear regression as an alternative model to the fixed complete randomized design ANOVA in the case of equal replications.

In the practice, healthy eating means eating a choice of diets that provide you the nutrients you want to retain your healthiness, feel good, and have energy. These nutrients include protein, carbohydrates, fat, water, vitamins, and minerals. Nutrition is important for everyone.

For comparing the effect of four different diets on reduction of serum cholesterol levels, $N = nk = 32$ men were assigned at random to one of the four treatment groups as shown in Table No. (1). Where the criterion of classification is diet, it is of interest to test whether the means of all four diets are significantly different, that is, do these diets have an effect on cholesterol level in normal men.

Table (1): Provides Natural Cholesterol Levels (in mg/100 ml)

Diet1	Diet2	Diet3	Diet4
260	195	180	240
265	245	220	210
220	260	245	220
235	235	230	235
240	230	235	250
225	250	280	235
255	240	260	225
250	200	225	180

At the end of 3 months, serum cholesterol determinations were made on each of the participants. Therefore, the appropriate statistical model is that of the completely randomized design. These procedures require a large amount of computation, especially in the case of complicated classifications using SPSS which is useful tool in visualizing the relationship between an explanatory variables that is estimated to the regression treatment of the one-way analysis of variance ANOVA method.

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

5. Statistical Tests:

According to Shapiro-Wilk test of normality in equation 10 the result were obtained in Table (2) and explained by Figure (1), that gives the existence of normality assumption since the p-value for each diet group are greater than α =0.05, in other words Sig = 0.720, 0.286, 0.874 are 0,462 greater than α =0.05, the null hypothesis that the data are normality distributed is accepted. The results mentioned above were obtained through the program of SPSS 26.

Table (2): Shapiro-Wilk Test of Normality

	Diet	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Xi	D1	.148	8	.200*	.951	8	.720
	D2	.218	8	.200*	.899	8	.286
	D3	.188	8	.200*	.967	8	.874
	D4	.226	8	.200*	.924	8	.462
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

According to Figure (1) shows the normality curve of standard residual in SPSS 26, which is approximately symmetric, which is one of more important assumptions to be existed.

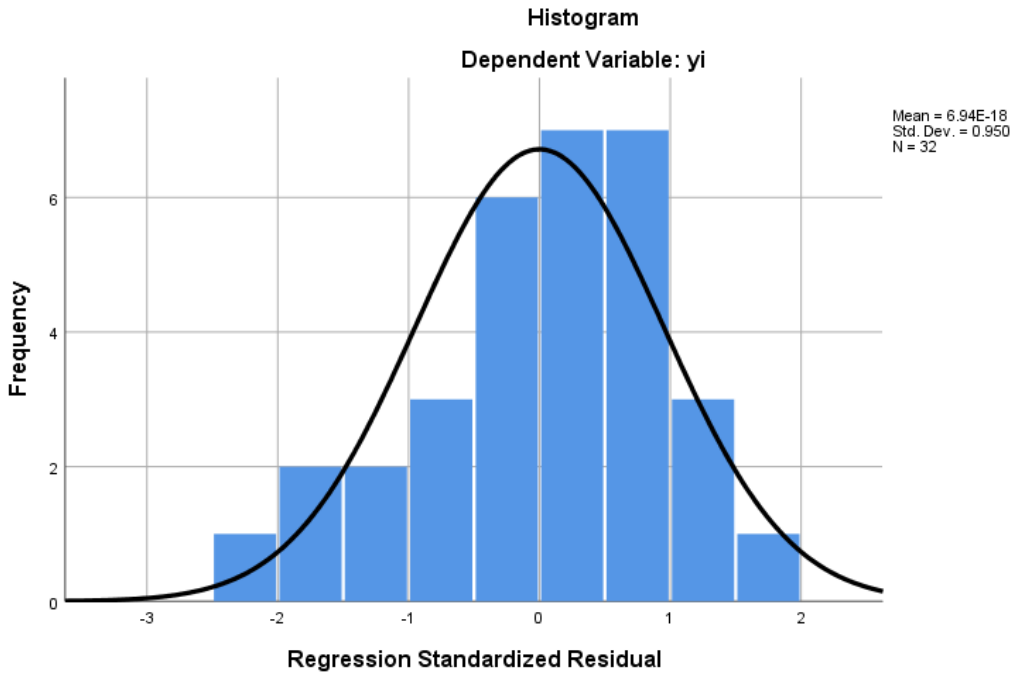

Figure 1: Normality of Standard Residual

Tikrit Journal of Administration and Economics Sciences, (31/3/2022);Vol. 18, No. 57, Part (2): 505-519
Doi: www.doi.org/10.25130/tjaes.18.57.2.30

According to the null hypothesis in SPSS 26 of Leven statistic for homogeneity of variance test as in equation (10) $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ were accepted as explained in Table (3) since the sig value or p-value is greater than α =0.05 level of significant .

Table (3): Leven statistic Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Xi	Based on Mean	.395	3	28	.758
	Based on Median	.348	3	28	.791
	Based on Median and with adjusted df	.348	3	22.071	.791
	Based on trimmed mean	.393	3	28	.759

Advanced statistical analysis requires a full-fledge statistical software. Depending on the nature of work, we use SPSS-26 for advanced analysis. Graphically a compression was made on cholesterol level as in Figure (2), and it shows stability of cholesterol level which affects the homogeneity of variance which it has an effect on achieving homogeneity of variance.

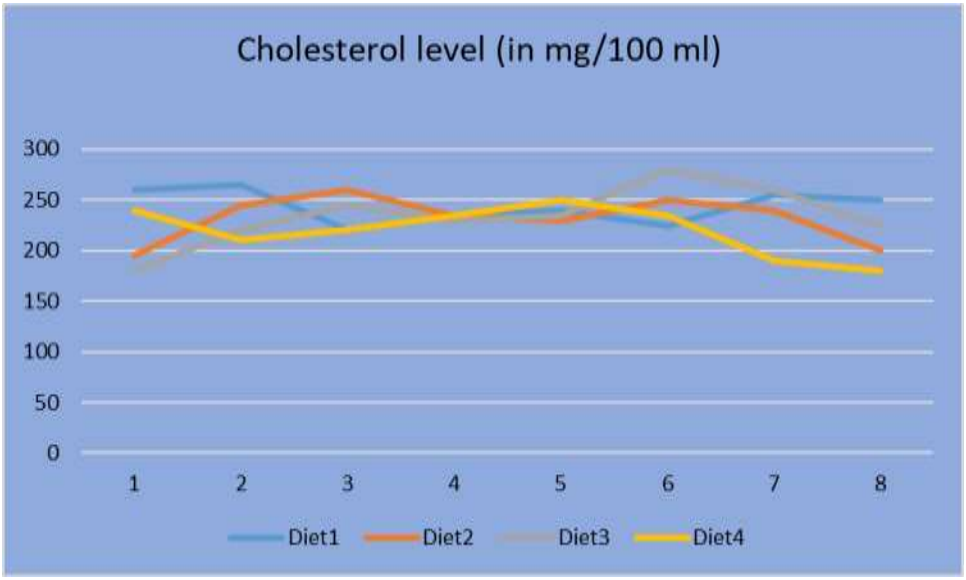

Fig. 2: Cholesterol level (in mg/100 ml) for four different diets

To compare the mean diets differences according to the hypothesis test of $H_0: t_i = 0$. Table 4 consist the F- test statistic and we can decide that there is no statistically significance difference between these four serum cholesterol groups since Sig > α =0.05. The population means are not different then it is assumed that diet has no effect on serum cholesterol level in this group of (32) normal males. Since the null hypotheses is

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

accepted, there is insufficient evidence to show that the population means are not equal. It cannot be said that the diets studied have an effect on cholesterol level in the study. This result is the same as that obtained using the linear regression model in Table 6.

Table (4): Mean Cholesterol Level (in mg/100 ml) ANOVA

Yi					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2293.750	3	764.583	1.329	.285
Within Groups	16106.250	28	575.223		
Total	18400.000	31			

5.1Multiple Linear Regression Applications:

Multiple linear regression analysis the sum square of regression parameters ($\beta_0 . \beta_1$, β_2 *and* β_3) the estimated regression equation (1) and (5) takes the following form;

$$\hat{Y} = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

$$\begin{bmatrix} 32 & 8 & 8 & 8 \\ 8 & 8 & 0 & 0 \\ 8 & 0 & 8 & 0 \\ 8 & 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 7440 \\ 1950 \\ 1855 \\ 1875 \end{bmatrix}$$

The coefficient of regression can be estimated by solving this equation.

$$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \bar{Y}_{4.} \\ \bar{Y}_{1.} - \bar{Y}_{4.} \\ \bar{Y}_{2.} - \bar{Y}_{4.} \\ \bar{Y}_{3.} - \bar{Y}_{4.} \end{bmatrix} = \begin{bmatrix} 220 \\ 23.750 \\ 11.875 \\ 14.375 \end{bmatrix}$$

Table (5) contains the regression coefficient of the estimated regression equation in the study:

Table (5): Multipole Linear Regression Coefficients

Regression Coefficients a						
Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	220.000	8.480		25.945	.000
	x1	23.750	11.992	.429	1.981	.058
	x2	11.875	11.992	.214	.990	.331
	x3	14.375	11.992	.260	1.199	.241
a. Dependent Variable: yi						

Doi: www.doi.org/10.25130/tjaes.18.57.2.30

Through Table (4) and (6), it can be concluded that the regression analysis is identical to the analysis of variance in the case of completely randomized design experiment with equal replications.

Table (6): ANOVA of Multiple Linear Regression

ANOVAa						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2293.750	3	764.583	1.329	.285b
	Residual	16106.250	28	575.223		
	Total	18400.000	31			
a. Dependent Variable: yi						
b. Predictors: (Constant), x3, x2, x1						

6. Conclusions:

- Through Table (4) and (6), it can be concluded that the regression analysis is identical to the analysis of variance in the case of completely randomized design experiment with equal replications. Otherwise, it cannot be applied with unbalanced design.
- To test the hypothesis test $H_0: t_i = 0$ in table (4), it can be conclude that there is no statistically significance difference between the serum cholesterol groups since Sig > α =0.05. it is assumed that diet has no significant effects on serum cholesterol level for 32 normal males in the study.
- In the situation of multiple linear regression model predictor variables possibly will be related through respectively and moreover with the dependent variable. However the analysis of variance outside the problematic of the related explanatory variables as extended as contracts with predectable qualitative variables this marks designs informal in analysis of variance associating to regression analysis.
- Dummy variable have variety of uses, mostly being used to model quantitative effects. They are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup, takes only the value 0 or 1 to indicate the absence or presence of some categorical effects.

7. Recommendations:

- The analysis of variance (ANOVA) is a method that occurs within regression models. Thus the balanced complete randomized design

ANOVA with equal replications can be 'treated as' linear regression by writing that data with binary variables, on they are known for their wide application to medical and biological data.

- ❖ Also, to the extent of the importance and necessity of the issue, we realize why regression analysis is so prominent in use in all kinds of fields, and not just in biology. We recommend in the clinical study; unequal sample size can be considered revising "Maximum Likelihood" estimation in multiple different samples since ANOVA does not include the investigation of a relation among binary or additional variables obviously. Slightly it instructions whether two or more samples from different populations have the equivalent mean.
- ❖ It can be describing ANOVA as a regression with dummy variables, that this is the case in the simple regression with categorical variables. A categorical variable will be encoded as an indicator matrix (a matrix of 0/1 depending on whether a subject is part of a given group or not) and then used directly for the solution of the linear system described by a linear regression.

Sources:

1. Abdullahi U. Usman, Hassan S. Abdulkadir and Kabiru Tukur, 2015, Application of Dummy Variables in Multiple Regression Analysis, Jodhpur National University, Jodhpur, Rajasthan, India. International Journal of Recent Scientific Research Vol. 6, Issue, 11, pp. 7440-7442, November, 2015.
2. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg, 2013, Applied Regression Analysis and Other Multivariable Methods, 2013. ISBN 10: 1285051084, Page 257.
3. Ivan N. Vuchkov and Lidia N Boyadjieva, 2001, Quality Improvement with Design of Experiments A Response Surface Approach, Kluwer Academic Publishers, London, ISBN 13: 97894009097.
4. Joseph L. Gastwirth, Yulia R. Gel and Weiwen Miao, 2010, The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. Statistical Science 2009, Vol. 24, No. 3, 343-360 DOI: 10.1214/09-STS301.
5. John W. Tukey, 1991, Exploratory Data Analysis: Past, Present, and Future, Technical Report No. 302 Princeton University, 408 Fine Hall, Washington Road, Princeton, NJ 08544-1000.
6. Kandethody M. Ramachandran, Chris P. Tsokos, 2021, Categorical data Analysis and goodness-of-fit tests and applications, in Mathematical Statistics with Applications in R, 3rd edition, 2021. Pages (461-490).
7. Karl P. Yule, G.U.; Blanchard, Norman; Lee, Alice, 1903, The Law of Ancestral Heredity, Biometrika, 2(2): 211- 236, doi:10.1093/ biomet/2.2.211. JSTRO 2331683

8. Mason R. L., Gunst Richard F, Hess James L., 2003, Statistical Design and analysis of Experiments with Applications to Engineering and Science, Second edition a John Willy and Sons Publication, Copyright at 2003.
9. Norman R. Draper, Harry Smith, 2014, Applied Regression Analysis. John Wiley & Sons, ISBN 0-471-02995-5, 3rd edition, Page (409).
10. Mohamed T. Abdelmoneim, 2004, Design and Analysis of Experiments. Angelo Egyptian Bookshop, ISBN 9770520608, page 27-39.
11. Suazan Garavaglla, Asha Sharma, 1998, A Smart Guide to Dummy Variables: Four Applications and A Macro. Dun and Bradstreet, Murray Hill, New Jersey 07974. Corpus ID: 8652875.
12. Wayne W. Daniel, and L. Chad Cross, 2013. Biostatistics A Foundation for Analysis in the Health Sciences, Copyright 2013. pp 305-327, and pp 490-510, John Wiley & Sons, Inc.