Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 26, Issue 1 (2024) ISSN$_{Print}$: 1816-9171

SCAD Variable Selection factorial experiments design with an application

Bahr Kadhim Mohammed
Maha Hadi Abed
Al-Qadisiyah University, College of Administration and Economics

Corresponding Author : Maha Hadi Abed

Abstract : In this paper, we will employ one of the factors reduction methods SCAD with the regression model of factorial experiments with two levels of degree 〖 2〗^4, as this method was compared with previous methods, and the MSE criterion was employed to compare the methods, as simulation methods and real data showed that the proposed method had its results better .

.Keywords: Full factorial experiment design, variable selection , SCAD, Regression experiments design.

Introduction: The design of experiments is one of the branches of advanced statistics that is concerned with conducting agricultural, industrial, medical and other experiments. The method of designing experiments aims to give the best possible agreement between factor levels (treatments) and the experimental units on which the treatments are to be tested, and thus the best response can be obtained and improved in the future**.**

Factorial experiments are one of the means of scientific research, which plays an important and effective role in studying and researching many characteristics of experimental materials. It is intended to know the effect of factors and their partial levels, as well as the interaction of those levels on the studied phenomenon and to know the extent of the response of experimental units to those factors involved. There are several designs through which factorial experiments are carried out. Including completely randomized design (CRD), randomized complete block design (CRBD), Latin square design....etc, and one of the most prominent features of the factorial experiment is that it produces for the researcher an experiment of more than one factor at the same time to reduce effort and cost and ease the analysis of the implemented experiment .

When designing experiments, the problem of high dimensions of the data appears in the matrix (X) after converting the mathematical model of the design into a multiple regression model through the use of the general linear model by transforming the effect of the levels of factors and the effect of interactions between the levels of these factors in independent variables. A common method for dealing with high-dimensional data is the penal least squares method, which is based on the principle of minimizing the sum of squares of error according to a certain parameter constraint. Therefore, the problem of selecting variables in regression models has received extensive study by researchers to overcome such problems. Penal methods (Lasso, SCAD, MCP) have gained great importance in recent times because of their speed in selecting explanatory variables and estimating parameters at the same time.

There are many types of factorial experiments, whose dimensions are determined by the factors involved in them and the number of levels for each factor. For example, the factorial experiment that consists of two factors, one with four levels and the other with five levels, is called a 4x5 factorial experiment (Yusuf, 2015). Then (Yates) came and used the statistical analysis method for factorial experiments of type 2^n and 3^n in an in-depth and comprehensive manner and developed statistical analysis methods, but these methods seemed difficult and became more complex when the number of factors involved in the experiment increased (Ghazi, 2018).

In (2001) Fan & Li introduced a new regularization approach known as smooth absolute deviation cut (SCAD). It is a particularly important method because of its computational features. SCAD is estimated to have an oracle property if the penalty parameter is chosen correctly (Fan & Li, 2001).

In (2010) Li & Lin performed variable least squares selection with SCAD penalty. Since an algorithm is proposed to find the penalized least squares solution, a standard error formula is derived for the penalized least squares estimation. With the correct selection of the regularization parameter, the resulting estimate turns out to be root n consistent and possesses the Oracle property, which works just as well if the correct sub-model were known (Li & Lin 2010).

This paper is organized into 6 sections as follows:

In the first section there is an introduction to the topic, in the second section we present the full factorial experimental design with two levels, and in the third and fourth sections we explain the regression model for the factorial design and the two criteria EER and IER . In Section 5, we describe variable selection and some methods (Lasso, SCAD, MCP) for analyzing factor experiments when the response variable follows a normal distribution. In Section 6 we

summarize the results of the simulation study and present an analysis of the sample data. Followed by the conclusions in Section 7.

Full factorial experiment design with two levels

Many experiments involve examining the effects of two or more factors. In general, factorial designs are the most efficient for this type of experiment. By factorial design we mean that in each complete experiment or iteration of the experiment, all possible combinations of factor levels are investigated. For example, if there are (a) levels of factor (A) and (b) levels of factor (B), then each replication contains all of the ab processing groups. When factors are arranged in a factorial design, they are often said to be intersecting (Wn & Hamada, 2009) .

For example, consider the simple experiment in Figure (1). This is a two-factor experiment with each of the design factors on two levels. We have called these levels "low" and "high" and denoted them as "-" and "+", respectively (Milliken & Johnson,1989).

The main effect of factor (A) in this two-tier design can be considered as the difference between the average response at the low level (A) and the average response at the high level (A) numerically, that is(Salim,2022)

$$\boldsymbol{A = \frac{40+52}{2} - \frac{20+30}{2} = 21}$$

That is, increasing the factor (A) from the low level to the high level leads to an increase in the average response (21) units. Similarly, the main effect of (B) is

$$\boldsymbol{B = \frac{30+52}{2} - \frac{20+40}{2} = 11}$$

If the factors appear on more than two levels, the above procedure must be modified since there are other ways to determine the influence of the factor. In some experiments, we may find that the difference in response between levels of one factor is not the same across all levels of other factors. When this happens, there is an interaction between the factors. For example, consider the two factor experiment shown in Figure (2). At a low level of factor B (or $-B$), the effect of (A) (Montgomery,D.C.,2017)

$$\boldsymbol{A = 50 - 20 = 30}$$

At a high level of factor B (or B+), the effect (A) is

$$\boldsymbol{A = 12 - 40 = -28}$$

Because the effect of (A) depends on the level chosen for factor (B), we see that there is an interaction between (A) and (B). The effect size of an interaction is the average difference between these two effects, or

$$\boldsymbol{AB = (28 - 30)/2 = -29}$$

Obviously, the interaction is significant in this experiment

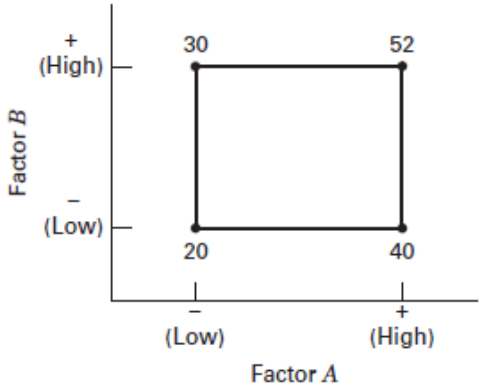


Figure (1): A two-factorial reality experiment with the response (Y) showing the angles

(Montgomery,D.C.,2013)

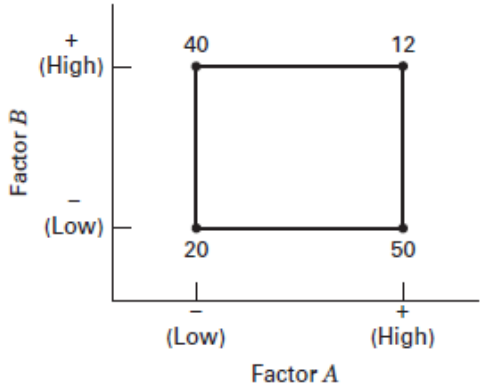


Figure (2): A two- factorial reality experiment with interaction

(Montgomery,D.C.,2013)

These ideas can be illustrated graphically. Figure (3) plots the response data, in Figure (4) against factor (A) for both levels of factor (B). Note that lines (B−) and (B+) are nearly parallel, indicating no interaction.

between factors (A) and (B). Similarly, Figure (4) plots the response data in Figure (2). Here we see that lines (B−) and (B+) are not parallel. This indicates that there is an interaction between factors (A) and (B) (Montgomery,D.C.,2013) .

Bilateral interaction diagrams such as these are often very useful in interpreting significant interactions and in reporting results to statistically untrained staff. However, it should not be used as the only data analysis technique because its interpretation is subjective, and its appearance is often misleading (Mee,2009).

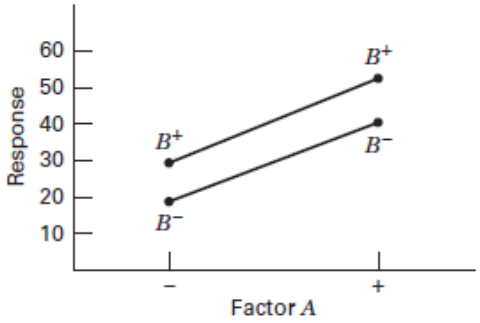


Figure (3) Factor A is an experiment without interaction

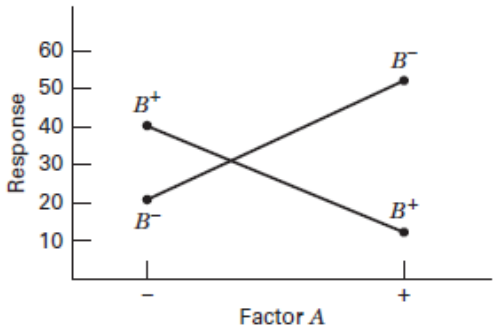


Figure (4) Factor A reaction experiment

There is another way to illustrate the concept of interaction. Let's assume that both of our design factors are quantitative (eg temperature, pressure, time). Then the regression model representation of the two-factor experiment can be written as:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon} \qquad \textbf{.... (1)}$$

where (y) is the response, β is the parameters whose values are specified,(x_1) is the variable representing the factor A, (x_2) is the variable representing the factor (B), ε is a random error term. The variables (x_1) and (x_2) are defined on a coded scale from (-1) to $(+1)$ (low and high levels of (A) and (B), $(x_1 x_2)$ represents the interaction between x_1 and x_2 **(Montgomery,D.C.,2017, Hinkelmann, & Oscar,2005).**

Regression Model for Factorial Design

Factorial design is important for studying models that were built using two or more factors. Factorial designs also have the advantage of being able to measure the effect of interaction between all factors through interaction between factors. Measurement of the response variable is affected by changing factor levels. The effect of the main factor is completely different compared to its interaction with other factors. Another possible approach is the possibility of using the main effects and interference effects of the two-level factorial designs as a regression model after transforming the levels of factors and the interactions between the levels of these factors into new variables called explanatory variables. We use the full world experience consisting of

Four operators with two levels for each factor denoted by 2^4, the factors are represented by capital letters A, B, C and D, there are $2^4 = 16$ treatment or level groups. By converting the levels of factors and the interaction between factors into explanatory variables, and according to the formula in the equation below (Montgomery,2009) :-

$$Yi=\beta_0+\beta_1\chi_1+\beta_2\chi_2+\beta_3\chi_3+\beta_4\chi_4+\beta_{12}\chi_1\chi_2+\beta_{13}\chi_1\chi_3+\beta_{14}\chi_1\chi_4+$$
$$\beta_{23}\chi_2\chi_3+\beta_{24}\chi_2\chi_4+\beta_{34}\chi_3\chi_4+\beta_{123}\chi_1\chi_2\chi_3+\beta_{124}\chi_1\chi_2\chi_4+\beta_{134}\chi_1\chi_3\chi_4+\beta_{234}\chi_2\chi_3\chi_4+\beta_{1234}\chi_1\chi_2\chi_3\chi_4+\varepsilon \quad \ldots.(2)$$

y response variable , β_s represent the parameters , (x_1, x_2, x_3, x_4) represent factors (A, B, C, D) respectively ,ε is a random error term . Often , When constructing a statistical model in the desired factor trials, the goal is to find a model for the estimated values of the response variable to be as close as possible to the actual values (Mohammed,2018) .

Error rates

The experimentwise error rate (EER) is proportion of the simulations when one or more effects is declared active. When conducting multiple statistical tests simultaneously, there is an increased risk of obtaining false positives, even when each individual test has a reasonably low alpha level. The (EER) is a way to address this issue and maintain the overall error rate at an acceptable level. Suppose there are k active effects , and,). Let P_i denote the proportion of simulations for i inert effects declared active by a specific procedure for which $i\,(i = 0, 1, \ldots, n-1)$,the experimentwise error rate (EER) will be defined as (Hamada & Balakrishnan 1998):

$$EER = 1 - P_k$$

This definition of individual error rate (IER) as the average proportion all effects are inactive can be effects declared active by suitably changing $n-1$ to the number of inactive effects. However, if we interpret "individual error rate" as the error rate associated with an individual hypothesis test or comparison, it would refer to the per-comparison error rate. This is the probability of making a Type I error in a single test, where a Type I error occurs when a null hypothesis is incorrectly rejected, the individual error rate (IER) will be defined as (Hamada & Balakrishnan 1998):

$$\sum_{i=k+1}^{n-1} P_i \left(\frac{i}{n-k-1}\right)$$

variable selection

5.1 Lasso variable Selection

Prepare the lasso method (The Least Absolute sgrinkage and selection operator) method, which was proposed by (Tibshirani) in (1996, 1997), is one of the most famous punitive methods used in estimating and choosing the variables of the linear regression model, where the Lasso is used to reduce the estimates of some regression parameters, and others are equal to zero, and thus It is possible to estimate and select variables in one step, as it was indicated that the coefficient estimated by Lasso is biased towards large parameters (Jabber 2020) and the Lasso estimator is obtained through the following formula:-

Whereas:

$$\widehat{\boldsymbol{\beta}}\boldsymbol{lasso} = \textbf{argmin}\ \boldsymbol{\beta}\ \|\boldsymbol{y}-\boldsymbol{\chi\beta}\|^2 + \boldsymbol{\lambda}\ \|\boldsymbol{\beta_1}\| \qquad \textbf{... (3)}$$

sum of squared Error (SSE) $\|\boldsymbol{y}-\boldsymbol{\chi\beta}\|^2 = (\boldsymbol{y}-\acute{}\boldsymbol{\chi\beta})(\boldsymbol{y}-\boldsymbol{\chi\beta})$

It is the sum of the absolute values of the regression parameters and is called the L1-norm of the vector $\beta\ \|\boldsymbol{\beta_1}\| = \sum_{j=1}^{p} |\boldsymbol{\beta j}|$

5.2 Lasso with fractional factorial design

In the part will be employed (lasso) from equation (3) to equation (2), we get:

$$\boldsymbol{Yi}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1\chi_1}+\boldsymbol{\beta_2\chi_2}+\boldsymbol{\beta_3\chi_3}+\boldsymbol{\beta_4\chi_4}+\boldsymbol{\beta_{12}\chi_1\chi_2}+\boldsymbol{\beta_{13}\chi_1\chi_3}+\boldsymbol{\beta_{14}\chi_1\chi_4}+$$
$$\boldsymbol{\beta_{23}\chi_2\chi_3}+\boldsymbol{\beta_{24}\chi_2\chi_4}+\boldsymbol{\beta_{34}\chi_3\chi_4}+\boldsymbol{\beta_{123}\chi_1\chi_2\chi_3}+\boldsymbol{\beta_{124}\chi_1\chi_2\chi_4}+\boldsymbol{\beta_{134}\chi_1\chi_3\chi_4}+$$
$$\boldsymbol{\beta_{234}\chi_2\chi_3\chi_4}+\boldsymbol{\beta_{1234}\chi_1\chi_2\chi_3\chi_4}+\boldsymbol{\lambda}[|\boldsymbol{\beta_0}|+|\boldsymbol{\beta_1}|+|\boldsymbol{\beta_2}|+|\boldsymbol{\beta_3}|+|\boldsymbol{\beta_4}|+|\boldsymbol{\beta_{12}}|+|\boldsymbol{\beta_{13}}|+|\boldsymbol{\beta_{14}}|+|\boldsymbol{\beta_{23}}|+|\boldsymbol{\beta_{24}}|+$$
$$|\boldsymbol{\beta_{34}}|+|\boldsymbol{\beta_{123}}|+|\boldsymbol{\beta_{124}}|+|\boldsymbol{\beta_{234}}|+|\boldsymbol{\beta_{1234}}|] \qquad \text{... (4)}$$

5.3 MCP Variable Selection

The Minimax Concave Penalty (MCP) is another alternative to obtain less biased regression coefficients in sparse models. Zhang (2010) proposed the MCP method that estimates and selects linear regression variables simultaneously, and overcomes the lasso method in terms of its inconsistency in selecting variables. The MCP estimator is obtained with the following formula: (Choon, 2012)

$$\hat{\beta}_j^{MCP} = argmin_p \|y - X\beta\|^2 + \sum_{j=1}^{p} P_{h,\gamma}^{MCP} \qquad ...(5)$$

when :

$\sum_{j=1}^{p} P_{h,\gamma}^{MCP}$ is the MCP penalty function

The MCP penalty takes the following form:

$$P(|\beta|) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma}, & |\beta| < \lambda\gamma \\ \frac{\lambda^2\gamma}{2} & |\beta| \geq \lambda\gamma \end{cases} \qquad ...(6)$$

Where: $\gamma > 1$ (Breheny, 2016)

Many concave penalties are based on λ , as well as including a fine-tuning parameter (γ) that controls the concavity of the penalty (i.e. how quickly the penalty is reduced).

MCP starts by applying the same penalty rate as the lasso, then relaxes the rate down smoothly to zero as the absolute value of the coefficient increases as compared to SCAD, however, MCP relaxes the penalty rate immediately whereas with adjusted SCAD it stays flat for a while before decreasing.So MCP is simpler than SCAD as it uses a single node instead of a single node to achieve the required properties (Breheny, 2016).

5.4 MCP with Factorial Design Regression Model

In the employed part (MCP) of equation (6) to equation (2), we get:

$$\boldsymbol{y_i}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1x_1}+\boldsymbol{\beta_2x_2}+\ldots+\boldsymbol{\beta_4x_4}+\boldsymbol{\beta_{12}x_1x_2}+\cdots+$$
$$\boldsymbol{\beta_{14}x_1x_4}+\boldsymbol{\beta_{23}x_2x_3}+\boldsymbol{\beta_{24}x_2x_4}+\ldots+\boldsymbol{\beta_{123}x_1x_2x_3}+\boldsymbol{\beta_{124}x_1x_2x_4}+\ldots+\boldsymbol{\beta_{1234}x_1x_2x_3}+\ldots+\boldsymbol{\beta_{234}x_2x_3x_4}+$$
$$\boldsymbol{\beta_{1234}x_1x_2x_3x_4}+\Big[\boldsymbol{\lambda}\,(|\boldsymbol{\beta_0}|+|\boldsymbol{\beta_1}|+|\boldsymbol{\beta_2}|+\cdots+|\boldsymbol{\beta_4}|+|\boldsymbol{\beta_{12}}|+|\boldsymbol{\beta_{13}}|+\cdots+|\boldsymbol{\beta_{14}}|+|\boldsymbol{\beta_{123}}|+|\boldsymbol{\beta_{124}}|+\cdots+$$

$$|\boldsymbol{\beta_{1234}}|) - \frac{1}{2\gamma}(|\boldsymbol{\beta_0}|^2 + |\boldsymbol{\beta_1}|^2 + |\boldsymbol{\beta_2}|^2 + \cdots + |\boldsymbol{\beta_4}|^2 + |\boldsymbol{\beta_{12}}|^2 + |\boldsymbol{\beta_{13}}|^2 + |\boldsymbol{\beta_{14}}|^2 + +|\boldsymbol{\beta_{123}}|^2 + |\boldsymbol{\beta_{124}}|^2 + +|\boldsymbol{\beta_{345}}|^2 + |\boldsymbol{\beta_{1234}}|^2)\Big] \quad \ldots(7)$$

5.5 SCAD

The SCAD method (smoothly clipped Absoulute Deviation) was proposed in 2001 by (Fan & Li), which works to estimate and select the variables of the linear regression model simultaneously using the SCAD penalty function (Fan,1997; Fan & Li, 2001). The SCAD estimator was obtained through the Punitive Least Squares Function (pLSF) as follows:

$$\hat{\beta}^{SCAD} = argmin_{\beta}\left[\|y - x\beta\|^2 + \sum_{j=1}^{p} P_{\lambda,\gamma}^{SCAD}\right] \quad \ldots.(8)$$

The SCAD function takes the following formula:

(Clarke et al ; 2009)

$$P_{\lambda,\gamma}^{SCAD} = \begin{cases} \lambda|\beta| & if \ \ |\beta| \leq \lambda \\ \frac{(2\gamma\lambda|\beta| - \beta^2 - \lambda^2)}{2(a-1)} & if\ \lambda < |\beta| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & if\ |\beta| \geq \gamma\lambda \end{cases} \quad \ldots. (9)$$

Since$\lambda \geq 0, \gamma > 2$ are the two adjustment parameters

Fan & Li (2001) set the value of the adjustment parameter $\gamma = 3.7$

The SCAD penalty function takes the first derivative according to the following formula:

$$P_{\lambda,\gamma}^{SCAD} = \begin{cases} \lambda & if \ \ |\beta| \leq \lambda \\ \frac{(\gamma\lambda - |\beta|)}{(\gamma-1)} & if\ \lambda < |\beta| < \gamma\lambda \\ 0 & if\ |\beta| \geq \gamma\lambda \end{cases} \quad \ldots.(10)$$

The researchers agreed (Fan & Li, 2001; Fan & peng, 2004; kim et. al., 2008) that the punishment function SCAD is characterized as almost an oracle with the choice of the adjustment parameter λ, and it was used by (Garcia et al; 2010) in the presence of missing data, and used by (Qiu et al;2015) in variable coefficient models with self-correlated errors.

All these studies agreed to set the control parameter $\gamma = 3.7$ for the SCAD punishment function, because this value gives a satisfactory performance for the various variable selection issues.

5.6 SCAD with Fractional Factorial Design

In the SCAD part of equation (10) to equation (2), we get:

$$\boldsymbol{Yi}=\boldsymbol{\beta_0}+\boldsymbol{\beta_1}\boldsymbol{\chi_1}+\boldsymbol{\beta_2}\boldsymbol{\chi_2}+\boldsymbol{\beta_3}\boldsymbol{\chi_3}+\boldsymbol{\beta_4}\boldsymbol{\chi_4}+\boldsymbol{\beta_{12}}\boldsymbol{\chi_1}\boldsymbol{\chi_2} + \boldsymbol{\beta_{13}}\boldsymbol{\chi_1}\boldsymbol{\chi_3}+\boldsymbol{\beta_{14}}\boldsymbol{\chi_1}\boldsymbol{\chi_4}+ \boldsymbol{\beta_{23}}\boldsymbol{\chi_2}\boldsymbol{\chi_3}+\boldsymbol{\beta_{24}}\boldsymbol{\chi_2}\boldsymbol{\chi_4}+\boldsymbol{\beta_{34}}\boldsymbol{\chi_3}\boldsymbol{\chi_4}+\boldsymbol{\beta_{123}}\boldsymbol{\chi_1}\boldsymbol{\chi_2}\boldsymbol{\chi_3}+\boldsymbol{\beta_{124}}\boldsymbol{\chi_1}\boldsymbol{\chi_2}\boldsymbol{\chi_4}+\boldsymbol{\beta_{134}}\boldsymbol{\chi_1}\boldsymbol{\chi_3}\boldsymbol{\chi_4}+\boldsymbol{\beta_{234}}\boldsymbol{\chi_2}\boldsymbol{\chi_3}\boldsymbol{\chi_4}+\boldsymbol{\beta_{1234}}\boldsymbol{\chi_1}\boldsymbol{\chi_2}\boldsymbol{\chi_3}\boldsymbol{\chi_4}+ \frac{1}{(\gamma\lambda-1)}[\boldsymbol{\gamma\lambda} - (|\boldsymbol{\beta_0}| + |\boldsymbol{\beta_1}| + \cdots + |\boldsymbol{\beta_4}| + |\boldsymbol{\beta_{12}}| + \cdots + |\boldsymbol{\beta_{14}}| + \cdots + |\boldsymbol{\beta_{34}}| + |\boldsymbol{\beta_{123}}| + |\boldsymbol{\beta_{124}}| + \cdots + |\boldsymbol{\beta_{1234}}|)]$$

....(11)

Application

6.1 Simulation study

Simulation is the creation of a representation or imitation of actual reality using specific models written according to a software method in order to obtain experimental results aimed at validating the results obtained from the theory in order to determine the most appropriate method of analysis. In this section, simulated experiments will be created based on the Monte Carlo technique in order to study the performance of the two proposed methods (SCAD, MCP) with factorial experiments and compare them with the current methods (lasso, adaptive lasso). The performance of these methods is evaluated on the basis of the MSE criterion. A good method is the one with the smallest MSE value. Programming (R) is used to analyze the data. The main model for the simulation study is given as follows:

$$y_i=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_5x_5 + \beta_{12}x_1x_2 + \cdots + \beta_{15}x_1x_5+\beta_{23}x_2x_3+\beta_{24}x_2x_4+\ldots+\beta_{45}x_4x_5+\beta_{123}x_1x_2x_3+\beta_{124}x_1x_2x_4+\ldots+\beta_{345}x_3x_4x_5+\beta_{1234}x_1x_2x_3x_4+\beta_{1235}x_1x_2x_3x_5+ \ldots+\beta_{2345}x_2x_3x_4x_5 + \beta_{12345}x_1x_2x_3x_4x_5 + \varepsilon$$

we used 1,000 iterations generated for each combination of factors $A, B, C, D, and\ E$, which are represented by the parameters $(\beta_1, \beta_2, \ldots \beta_{12}, \ldots \beta_{12345})$. The R program simulation algorithm is as follows:

Step 1: Create a response variable based on a normal distribution with a known mean and known variance $Y \sim N\ (\mu, \sigma^2)$ with mean equal to (3) and variance equal to (1) .

Step 2: Generate a random error term based on the standard normal distribution , a known mean and known variance $\varepsilon \sim N\ (o, \sigma^2)$

Step 3: Create a design matrix of five factors and two levels (high level [+1] and low level [-1]).

Step 4: Create a different number of samples in each iteration using (n = 100,n=200,n=300).

Step 5: Calculate the main effects and reaction effects according to equation (2).
Step 6: Applying the suggested approach by using SCAD method in equations (10) and (2).
Step 7: Applying the suggested approach by using MCP method in equations (6) and (2).
Step 8: Calculate the comparison methods MSE, EER and IER for all methods of study.

6.2 First simulation experiment

In this experiment, the Monte Carlo simulation technique was relied upon to study the performance of the proposed methods (SCAD, MCP) and compare them with other methods (Lasso, Adaptive lasso) for the purpose of reaching the best penalty method that works as a method of selecting variables. Table (1) below represents the estimation and selection of the coefficients of the factorial experiment of the methods.

Table (1) : Estimation and selection of factorial experiment coefficients of methods (Third simulation experiment) for $\sigma^2 = 1$

Effects	$n = 100$				$n = 200$				$n = 300$			
	Variable selection methods				Variable selection methods				Variable selection methods			
	Lasso	Adaptive lasso	SCAD	MCP	Lasso	Adaptive lasso	SCAD	MCP	Lasso	Adaptive lasso	SCAD	MCP
β_1	0.705	0.386	0.491	-0.150	0.707	0.606	0.868	0.174	0.550	-0.044	-0.053	0.145
β_2	0.287	0.624	0.406	-0.193	0.303	-0.247	0.000	0.782	0.569	0.849	-0.449	0.709
β_3	0.419	-0.029	-0.749	0.984	0.852	0.430	0.847	0.000	0.568	0.410	0.001	0.970
β_4	-0.449	0.460	0.569	-0.683	0.257	0.819	0.582	0.620	0.734	0.836	0.951	0.451
β_5	0.408	0.800	0.859	0.006	0.004	-0.557	0.150	0.662	0.728	0.134	-0.620	0.000
β_{12}	0.374	0.000	0.000	0.000	0.806	0.000	0.000	0.000	0.118	0.000	0.000	0.000
β_{13}	-0.169	0.111	-0.732	0.373	-0.126	0.373	0.673	0.477	0.685	0.664	0.312	0.031
β_{14}	0.204	0.455	0.921	0.000	0.357	0.054	0.000	0.107	0.238	0.067	0.000	0.135
β_{15}	0.921	0.453	0.194	0.368	0.332	0.454	0.000	0.188	0.893	-0.033	-0.745	0.887
β_{23}	0.100	0.201	0.000	0.734	0.012	0.505	0.430	0.956	0.097	0.624	0.707	0.979
β_{24}	0.553	0.596	0.401	0.287	0.640	0.579	0.642	0.204	0.920	0.287	0.493	0.560
β_{25}	0.809	0.156	-0.974	0.393	0.359	0.016	0.521	0.095	0.681	-0.761	0.000	0.000
β_{34}	0.000	0.000	0.270	-0.729	0.015	0.149	0.000	0.128	0.401	-0.669	0.000	0.513
β_{35}	0.090	0.834	0.997	0.857	0.041	0.315	0.262	0.056	0.194	0.159	0.994	-0.075
β_{45}	0.257	0.239	0.000	-0.261	-0.522	0.061	0.542	0.710	0.871	0.418	0.000	1.884
β_{123}	0.757	0.966	0.685	0.837	0.369	0.425	0.690	0.617	0.758	0.662	0.312	0.386
β_{124}	2.815	0.936	0.529	0.267	0.138	0.866	-0.580	0.251	0.785	0.696	0.000	0.000
β_{125}	0.403	0.000	0.000	0.000	0.677	0.625	0.000	-0.167	0.924	0.006	0.862	0.324
β_{134}	0.888	0.139	0.633	0.486	0.160	0.144	0.979	0.713	0.314	-0.864	0.000	0.574
β_{135}	0.762	0.182	0.000	0.000	0.229	0.198	0.000	0.749	0.068	0.806	0.779	-0.322
β_{145}	0.675	-0.587	0.686	-0.110	0.240	0.329	0.000	0.777	0.146	0.321	0.095	0.000
β_{234}	0.148	0.911	0.370	0.301	0.680	0.000	0.946	0.084	0.828	0.481	0.482	0.233
β_{235}	-0.325	-0.140	0.957	0.000	-0.158	0.644	0.760	0.000	0.522	-0.780	0.677	0.204
β_{245}	-0.436	-0.011	0.367	-0.842	0.850	0.000	0.580	0.450	0.425	0.032	0.236	0.681
β_{345}	0.865	0.463	0.498	-0.881	-0.140	0.759	0.724	0.146	0.484	0.976	0.000	0.489
β_{1234}	0.303	0.020	0.204	-0.066	0.238	0.283	0.000	0.476	0.377	0.479	0.000	0.861
β_{1235}	0.857	0.846	0.000	0.736	0.246	0.597	0.392	0.000	0.000	0.280	0.365	0.800
β_{1245}	0.963	0.736	0.000	0.929	0.317	0.492	0.287	0.726	0.127	0.715	0.111	0.000
β_{1345}	0.238	0.382	0.295	0.830	0.240	0.516	0.789	0.000	0.417	0.698	0.308	0.000
β_{2345}	0.942	0.344	0.000	0.264	0.169	0.833	0.324	0.611	0.405	0.683	0.000	0.691
β_{12345}	0.136	-0.696	0.088	0.576	0.673	0.826	0.000	0.000	0.249	0.436	0.740	0.458
MSE	**0.893**	**0.888**	**0.673**	**0.734**	**0.764**	**0.7348**	**0.573**	**0.622**	**0.66**	**0.621**	**0.487**	**0.51**
EER	**0.037**	**0.034**	**0.003**	**0.016**	**0.148**	**0.134**	**0.021**	**0.044**	**0.076**	**0.024**	**0.006**	**0.021**
IER	**0.009**	**0.008**	**0.004**	**0.006**	**0.0084**	**0.007**	**0.004**	**0.006**	**0.008**	**0.005**	**0.001**	**0.003**

From table (1) above, the following can be seen:

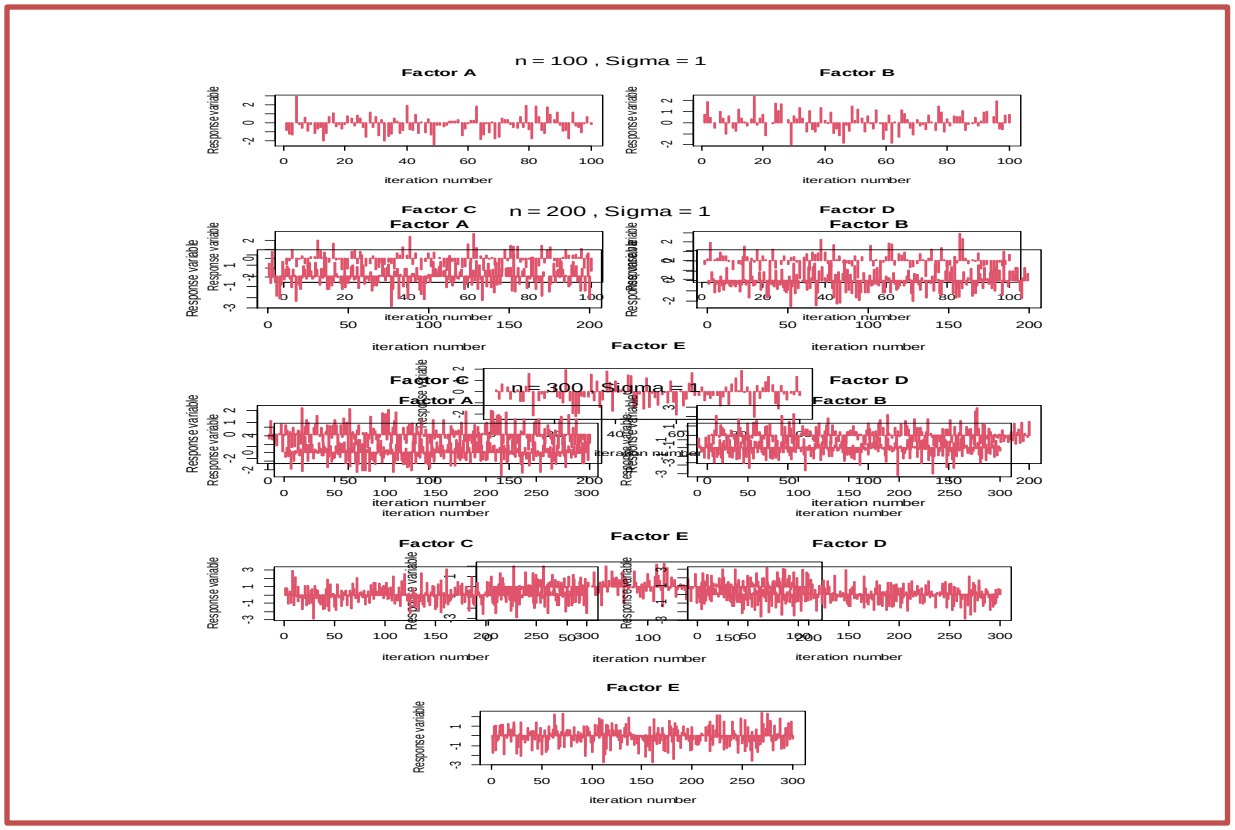
1- When the value of $\sigma^2 = 1$ with a sample size of $n = 100$, we note that the best proposed method is SCAD, as the value was ($MSE = 0.673$), while the value of EER was (0.003), while the value of IER was (0.004), which is the lowest. Compared to the methods (Lasso, Adaptive lasso), then comes the MCP method, as the value of ($MSE = 0.734$), while the value of EER was (0.016) while the value of IER was (0.006).

2- When the value of $\sigma^2 = 1$ and when increasing the sample size of $n = 200, n = 300$, it is clear that the values of the criteria (MSE, EER, IER) are less than the previous one, and this indicates that the proposed methods (MCP SCAD,) are likely Be relevant in explaining the model and factors compared to other methods previously mentioned.

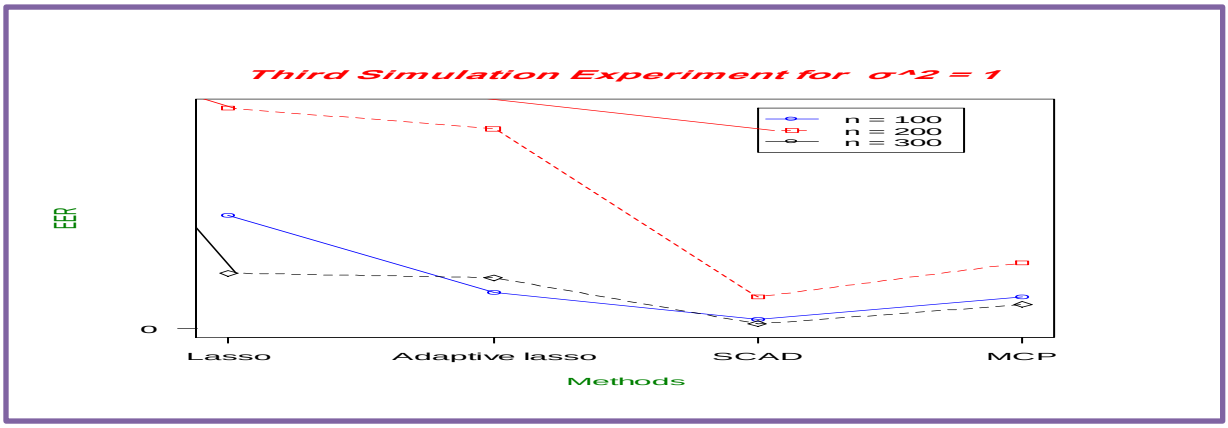
3- It was found from the results of the four methods (Lasso, Adaptive Lasso, SCAD, MCP) for estimating and choosing the factors of the factorial experiment model consisting of five factors with two levels for each factor that some parameters are equal to zero for some of the four methods, but my method (SCAD, MCP) gave better results. This is evidenced by the mean squared error of the simulation experiments, where the mean squared error was less likely to be compared with other methods.

Below is a trace plot of five factors A, B, C, D, E, when $n = 100$, and the values of $\sigma^2 = 1$. We note that the series values of the above factors are stable in one direction. When increasing the sample size $n = 200$ with the values of $\sigma^2 = 1$ constant, we notice that the series values for the above factors are closer and more stable than the previous case to the default values of the simulation. Whereas Figure (5) below shows that the values are centered near the factor averages. As for the figure (6) below, it represents the EER standard among the methods

Drawn (5) trace plot is for the third simulation of a factorial experiment



the figure (6) represents the EER standard among the methods when $\sigma^2 = 1$



6.2.1 The First application

This experiment was conducted in the Sheep and Goat Research Station of the general authority for agricultural research / ministry of agriculture in Abu Ghraib for the years 2010-2011, using 133 records that included veterinary records, breeding records, and pedigree records available at the station, (68) sheep infected with inflammatory bowel disease were selected which represents the response variable for the purpose of determining the important factors that led to infection with this disease in sheep, namely: breed, year of infection, age at infection, season of infection, and each factor has two levels (Anam et al., 2015).

Table (2) The Factors and Levels for each factor

Factors	Factors levels	
	High level :+1	Low level : -1
A= strain	Turkish Awassi	Awassi local
B= year of injury	2011	2010
C= age at injury	More than a year	less than one year
D= injury season	winter season	Spring season

6.2.2 Testing the data distribution(The second application)

There are many statistical tests used to determine the distribution of data for the studied phenomenon, including Kolmogorov-Smirnov test, Shapiro-Wilk test, and Boxplots. To find out the distribution of these data (responses), using the R program, a graph was plotted with the data distribution curve as shown in Figure (7), and Boxplots were included as shown in Figure (8) which shows the relevance of the data and its distribution by factor levels. Both models show that the data follow a normal distribution.

Figure (7) The histogram with the distribution curve of the (RIBO) (the Second application)

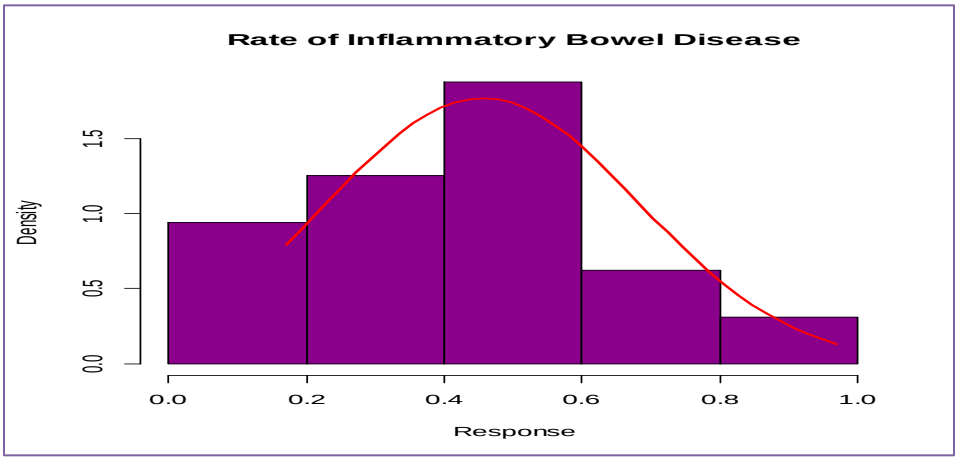
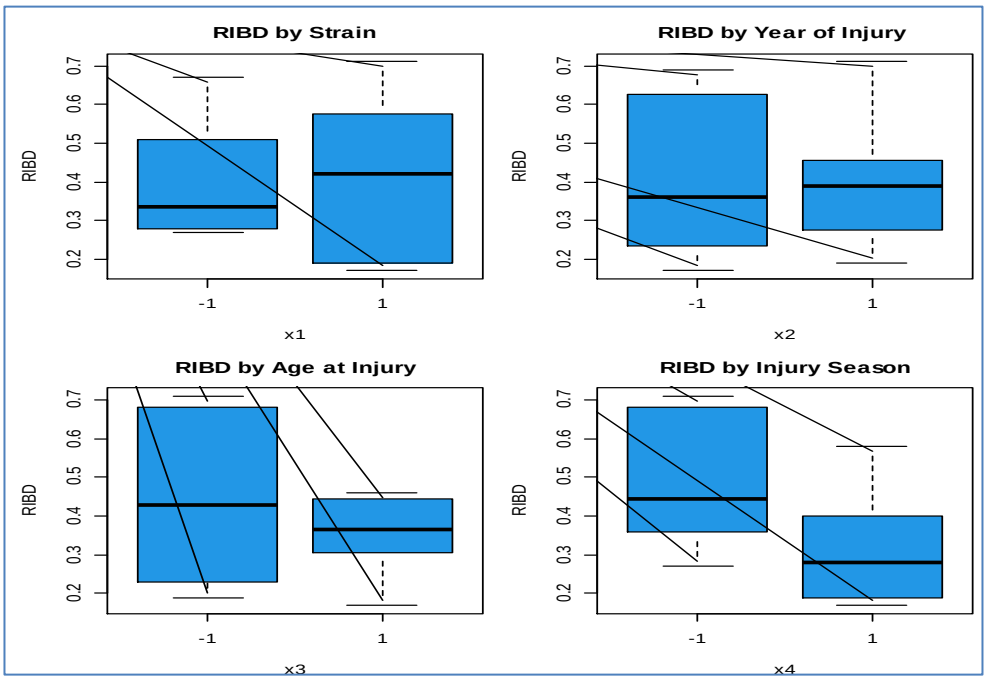


Figure (8) Boxplot charts for factors (A,B,C,D,E) according to RIBD



6.3 Real data and analysis

The data was analyzed using a program R to determine the most important factors that affect (the infection of inflammatory bowel disease in sheep), and we used SCAD the variable selection method to determine the most important factors that lead to the infection of this disease. The following table (3) show the results obtained

Table (3) Main effects (factors)

Factor	A	B	C	D
SCAD Method	-0.00435	-0.10987	0.00000	0.0000

Source: author's computations.

Through the results of the table (**3**) we note that the value of factor C (age at infection) and the value of factor D (season of infection) are equal to zero, which indicates that these factors have no effect on the response variable (enteritis in sheep). We also note that factor A (strain) and factor B (year of infection) are the main factors that have a significant significant impact on the response variable.

Table (4) Two-factor interactions

Factor	AB	AC	BC	AD	BD	CD
SCAD Method	0.764	0.875	0.543	0.000	0.0654	0.000

Source: author's computations.

Through the results of the table (4) we note that the interactions AD (strain and season of infection), CD (age at infection and season of infection) are equal to zero, which indicates that these interactions have no effect on the response variable (enteritis in sheep). We also note that the interactions AB (strain and year of injury), AC (strain and season of injury), BC (year of injury and age at injury), BD (year of injury and season of injury) represent the binary interactions that have a significant significant effect on the response variable**.**

Table (5) Three and Four -factor interactions

Factor	ABC	ABD	ACD	BCD	$ABCD$
SCAD Method	0.684	0.876	0.000	0.000	0.875

Source: author's computations.

Through the results of the table (5) we note that the interactions ACD (strain, age at infection and season of infection), BCD (year of infection, age at infection and season of infection) are equal to zero, which indicates that these interactions have no effect on the response variable (enteritis in sheep). We also note that the interactions ABC (strain,

year of infection, and age at injury), ABD (strain, year of injury, and season of injury), $ABCD$ (strain, year of injury, age at injury, and season of infection) represent the triple and quadruple interactions that have a significant significant effect on the response variable.

7. Conclusions

The last chapter, the six chapter, included identifying some of the results and conclusions reached from the theoretical and practical side and future studies of this study.

1- We conclude from the simulation results that the SCAD method is the best method in the process of estimating and selecting important factors, and then comes the MCP method, when compared to the penalty methods (Lasso, ALasso) as they achieve less MSE.

2- We conclude that the proposed remedial methods showed important factors in the interactions between these factors. Also, the results of the proposed remedial methods (SCAD, MCP) were more accurate than the results of the retributive methods (Lasso, ALasso) that were used in the study, and this can be clearly seen through the comparison criteria.

3- The results of the proposed model (factorial design regression model) showed that the factors affecting the incidence of inflammatory bowel disease in sheep are: factor C (age at infection) and the value of factor D (season of infection) had no effect on the response variable (enteritis in sheep). And that factor A (strain) and factor B (year of infection) are the two main factors that have a significant impact on the response variable.

4- The results showed that the following interactions: (AD, D, AB, AC, BC, BD) represent the bilateral interactions that have a significant impact on the response variable.

5- The results showed that (ACD, BCD) interactions had no effect on the response variable (enteritis in sheep).

6- The results showed that the interactions $(ABD, ABCD, ABC)$ represent the triple and quadruple interactions that have a significant effect on the response variable.

8- Recommendations

1- The study recommends the need to use modern punitive methods (SCAD, MCP) in the medical, industrial and agricultural fields.

2- The study recommends using (EER, IER) standards and not being satisfied with (MSE) standard for the purpose of obtaining more accurate and reliable results.

3- In experiments that contain a large number of components and interactions, the proposed model can be used to clarify the important factors as well as the important interactions between the factors.

REFERENCES

1 -Breheny, (Patrick,2016). “Adaptive lasso, MCP, and SCAD” .

2 -Choon, (Chua Lai, 2012).” Minimax concave bridge penalty function for variable selection”. A Dissertation Presented to the DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY NATIONAL UNIVERSITY OF SINGAPORE In Partial Fulfillment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY- National University of Singapore.

3 - Clarke, B., Fokoue, E., and Zh-ang, H.H. (2009), Principles and Theory for Data Mining and Machine Learning, Springer, New York.

4 -D. C. Montgomery, Design and Analysis of Experiments, 7th Edition, NewYork: Wiley, 2009.

5 - Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Pr-operties," Journal of the American Statistical Association, 96, pp- .1348-1360.

6 -Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Pr-operties," Journal of the American Statistical Association, 96, pp- .1348-1360.

7 -Fan, J. (1997), "Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis," Journal of the Italian Statistical Society, 2, pp-.13- 1-138.

8 - Hinkelmann, K. & K. Oscar ; (2005). Design and Analysis of Experiments. Volume 2 ; John Wily & Sons , Inc , New York

9- Jabbar, E.(2020). A non-linear multi-dimensional estimation and variable selection via regularized MAVE method. Thesis submitted to college of administration and economics. University of Al- Qadisiyah. Iraq.

10-Li, Runze., & Lin, Dennis K. J. (2010). “ Variable Selection for Screening Experiments”. Qualcomm Technol Quant Manag, 6(3), 271-280.

11-Mohammed, Bahr Kadhim(2018))," "Robust Lasso Variable Selection for Factoria Experiments Analysis with Application"International Journal of Statistics an Applications 2018, 8(2): 79-87 .

12- Montgomery, D.C. (2017), Design and Analysis of Experiments. 9th Edition. John Wily & Sons. Inc New York.

13- Montgomery , D.C.,”Design-and-analysis-of-experiments” ,2013.p(186)

14- MILLIKEN, G.A., JOHNSON, D.E. 1989. Analysis of Messy Data – Volume 2:Nonreplicated Experiments. Van Nostrand Reinhold. New York. pp 92-108.

15- Mee, RW (2009). A Comprehensive Guide to Factorial Two Level Experimentation, springer, p(7) .

16- Nayef (Anam Abd al-Wahed), Taha (Ahmed Aladdin), Hadi (Fendiyeh Hussein), Saleh (Ibrahim Kazem), 2015 "A study of environmental factors and genetic parameters related to inflammatory bowel disease in Awassi sheep at the Abu Ghraib research station," Wealth Department Animals - College of Agriculture - University of Baghdad, Journal of Al-Mustansiriya Sciences, Vol. 26, Issue 2.

17- Saleem ، Wisam Wadullah ،(2022)، "Modeling of High-Dimensional Factors in the Design of Factorial Experiments Using Penal Methods with Practical Application", A Dissertation Submitted to the: Council of College of Administration and Economics /University of Baghdad in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Statistics.

18-Wu, C. F. J. and Hamada, M. (2009) Experiments: Planning, Analysis and Parameter Design Optimization, 2nd edition. Wiley, New York.

19- Ghazi, Marwa Haider, (2018), "Using some statistical methods to study the effect of green tea to reduce salt stress in cucumber plants," Master's thesis in Statistics Sciences / College of Administration and Economics / University of Karbala.

20- Yates,F.(1937) ,"Design analysis of factorial Experiment", Imprial Burean of soil scienes Harpenden Engeland , Vol. 35 , pp.77.

21- Yousef, Raad Raad (2015), "Analysis of factorial experiments n2 with an exponential distribution of the response variable with application", Master of Science thesis in Statistics - College of Administration and Economics - University of Baghdad.