International Journal of Statistics and Applications 2018, 8(2): 79-87
DOI: 10.5923/j.statistics.20180802.07

Robust Lasso Variable Selection for Factorial Experiments Analysis with Application

Bahr Kadhim Mohammed

Department of Statistics and Econometrics, The Bucharest University of Economic Studies, University of AL-Qadisiyah, Iraq

Abstract Many experiments have encountered problems in selecting the best model when the responses are non-normally distributed and especially when the number of factors is small. In this article, we propose to combine the lasso variable selection and robust method (Huber loss function), when the responses are distributed according to epsilon-skew-Laplace (ESL) distribution. The proposed modification (robust lasso) is compared to the traditional lasso and adaptive lasso for the analysis factorial experiment with two level for each factor. A simulation study and real data are conducted to investigate the performance of the proposed method. We employed the mean square errors to select the best model and the results show that the proposed method using robust lasso variable selection performs well.

Keywords Factorial experiment design, Variable selection, Epsilon-skew Laplace, Robust experiment design, Lasso

1. Introduction

In many of the methods used for estimating the Factorial experiment models, such as classical Least Squares (LS), several assumptions are required such as normality, constant variances and independency. Those assumptions can be violated due to several causes, such as the presence of outliers observations, or non-normal data distribution. The last few decades there has been a growing interest in finding alternative ways to the classical methods, especially in applications that require asymmetric distributions (non-normal distributions). Also, many applications contain a large number of factors believed to be relevant in the study, but the actual effects of these factors are often few and unimportant. [1] discussed how to define extreme values in experimental design. It was known that the experiment outside the laboratory is an observation that with values that do not match the pattern of the values produced by the rest of the data. To address these problems, many researchers have proposed methods (or functions) used with LS, which are called "robust methods". Some of these methods are M-estimation, Huber function, [2], [3], etc. In [4] proposed the construction of a model using a B - technique and Box-Cox to data conversion or using General Linear Models (GLM) to eliminate the not normal data. The graphs were compared to the estimated responses through the length of the confidence interval for the mean responses in the design of industrial experiments. [5] introduced a new flexible regression model by considering an error term distributed according to the epsilon skew normal (ESN) distribution. In addition to the estimate of the classical parameters, the skewness parameter has been estimated.

In [6] the author introduced Robust Regression Shrinkage and consistent variable selection through the least absolute deviation (LAD)-Lasso, using the robust lasso with Huber loss function. The results of the study showed that the proposed method is resistant to outliers or heavy-tailed errors. In [7] proposed the modification of the 2^n factorial experiments involving a Poisson response variable for comparison (Comparative Study of Analysis Factorial Experiments with a Poisson Distributed Response Variable, based on the criteria log-transformation (LOG), SQRT, ANOVA and GLM). The results of the study showed that the modified GLM approach exhibits the best performance with respect to all the three criteria in the entire parameter space and particularly so the expected response is likely to be very small. In [8] suggested using the general linear model (GLM) and log-transformation (LOG) for the purpose of experiments analysis. The response variable is non-normal distribution and the sample size is small. The work compared between the general linear model with log-transformation and ANOVA method on the basis of the results of the confidence limits and expected length confidence limits of E (LOCI) to the response variable which has an exponential distribution. [9] introduced the D-optimal design for when the error term in the simple linear regression follows the skew normal distribution. [10] utilized test Kraemer's and Schaefer's adaptive lasso on small samples and examined their effectiveness in designs with complex aliasing via simulations. In [11] introduced the case when the response variable follows the log-epsilon-skew-normal (LESN)

* Corresponding author:
baherm@yahoo.com (Bahr Kadhim Mohammed)
Published online at http://journal.sapub.org/statistics

distribution, which is an asymmetric probability distribution, and the parameters can be estimated using the maximum likelihood (MLE) method. The reliability for an experimental design that contains two factors with two levels, using simulation and real data are conducted to investigate the performance of the proposed method. The results show that the newly proposed method performs well.

In the present work we aim at combining the lasso variable selection and robust method (Huber loss function), for the analysis of the factorial experiment with two levels for each factor for the particular case when the response variable is distributed according to epsilon-skew-Laplace (ESL). Some concepts are explained in the relationship between factorial experiments and variables selection (factors). We also analyze two methods of variable selection – Lasso and adaptive Lasso. The simulation results indicated that this method gave acceptable results compared to Lasso and adaptive Lasso, depending on the criterion (MSE).

This paper is organized as follows. In Section 2 we present the fundamentals of the two level factorial experiments design and their advantages; in Section 3 we briefly introduce the concept of the epsilon-skew-Laplace distribution; in Section 4 we illustrate the variable selection and some methods (lasso, adaptive lasso) and the combination of the lasso variable selection with robust method (Huber function) for the analysis of factorial experiments when the response variable follows an ESL distribution. In Section 5 we summarize the results of a simulation study and present a data sample analysis. A brief conclusion is included in Section 6.

2. Two-level Factorial Experiment Design

Factorial experimental designs are used for study the joint effect of a number of factors on a response variable at the same time. There are many cases of the factorial experiments design that are used in applied studies. One of these cases is when there are a number of factors (p) with two levels in experiment. A full factorial experiment of such a design requires $2\times2\ldots\times2=2^p$ observations and is called a 2^p factorial experiment design [12]. The 2^p factorial experiment design is particularly useful in application fields such as medical, agricultural and industry, when there are many factors to be investigated. Therefore, these designs are used widely in these applications, because there are only two levels for each factor. In these designs we will refer to the levels as high and low, coded with +1 and -1, to denote the high and the low level of each factor. In most cases the levels are quantitative. Sometimes they are qualitative, such as gender, or two types of variety, brand, process or medical data.

This type of factorial experiment design has many methodological advantages:

It's orthogonal design that achieves the condition

$X^TX=nl$

where

$X=(x_{ij})$: design matrix and x_{ij} represents a factor j and level i.

I: identity matrix of the degree $q\times q$

The estimates of parameter β_j (j=1, 2…, q) resulting from orthogonal designs are unbiased and have lower variances.

Finally, the orthogonal design matrix (X) achieves the possibility of measuring the main effects of the factors independently; the effects do not overlap with each other.

The possibility of estimating the interactions of the first order, the second, etc.

Table 1. The full factorial 2^5 design with two levels

Run. no	Factors					Responses
	A	B	C	D	E	
1	-1	-1	-1	-1	-1	y_1
2	+1	-1	-1	-1	-1	y_2
3	-1	+1	-1	-1	-1	y_3
4	+1	+1	-1	-1	-1	y_4
5	-1	-1	+1	-1	-1	y_5
6	+1	-1	+1	-1	-1	y_6
7	-1	+1	+1	-1	-1	y_7
8	+1	+1	+1	-1	-1	y_8
9	-1	-1	-1	+1	-1	y_9
10	+1	-1	-1	+1	-1	y_{10}
11	-1	+1	-1	+1	-1	y_{11}
12	+1	+1	-1	+1	-1	y_{12}
13	-1	-1	+1	+1	-1	y_{13}
14	+1	-1	+1	+1	-1	y_{14}
15	-1	+1	+1	+1	-1	y_{15}
16	+1	+1	+1	+1	-1	y_{16}
17	-1	-1	-1	-1	+1	y_{17}
18	+1	-1	-1	-1	+1	y_{18}
19	-1	+1	-1	-1	+1	y_{19}
20	+1	+1	-1	-1	+1	y_{20}
21	-1	-1	+1	-1	+1	y_{21}
22	+1	-1	+1	-1	+1	y_{22}
23	-1	+1	+1	-1	+1	y_{23}
24	+1	+1	+1	-1	+1	y_{24}
25	-1	-1	-1	+1	+1	y_{25}
26	+1	-1	-1	+1	+1	y_{26}
27	-1	+1	-1	+1	+1	y_{27}
28	+1	+1	-1	+1	+1	y_{28}
29	-1	-1	+1	+1	+1	y_{29}
30	+1	-1	+1	+1	+1	y_{30}
31	-1	+1	+1	+1	+1	y_{31}
32	+1	+1	+1	+1	+1	y_{32}

Results are appropriate for a large number of experimental conditions.

Our study will be limited to designs of type 2^p, which is one of the types of factorial designs in which p is chosen from the possible factors and then only two levels are assigned to each factor. The lower level is indicated by (-1) and the high level by (+1). The experiment model is

implemented by representing 2^p of the factors, which allows us to estimate the main effects of factors X_j, ($j = 1, 2, ..., p$) as well as the possible interactions between the factors.

In this paper we use full factorial experiment consisting of five factors with two levels for each factor symbolized by 2^5. The factors are expressed in capital letters A, B, C, D, and E. There are 2^5=32 treatments or level combinations. Table 1. shows the full factorial design for five factors.

The above table 1 shows the effect of the interaction indicating whether the influence of one factor depends on the levels of the other factors or not.

The analysis of experiments can be formulated in the form of the general linear regression [13]; [14]. A common regression model for studying the main effects and interactions is:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ &+ \beta_{12} x_1 x_2 + \ldots + \beta_{15} x_1 x_5 + \ldots + \beta_{23} x_2 x_3 + \ldots \\ &+ \beta_{25} x_2 x_5 + \beta_{34} x_3 x_4 + \ldots + \beta_{123} x_1 x_2 x_3 + \ldots \\ &+ \beta_{12345} x_1 x_2 x_4 x_5 + e \cdot \end{aligned} \tag{1}$$

Here y is the response variable, the βs are unknown parameters, x_1, x_2, x_3, x_4, x_5 represent factors A, B, C, D and E, respectively, and ϵ is a random error term. The variables x_1, x_2, x_3, x_4, x_5, are coded as 1 and −1, for the high and low levels for their respective factors. The interaction between x_1 and x_2 is denoted as x_1x_2, and the other interaction effects are similarly defined. Often, when building a statistical model in the factorial experiments is desirable, the aim is to find a model to the estimated values for the response variable to be as close to the actual values as possible.

3. Epsilon Skew Laplace Distribution (ESL)

The random variable y has ESL distribution denoted by y~ ESL(λ,σ,ε), if there exist parameters λ ∈ R. σ>0, and -1<ε<1 such that the pdf of x is (Elsallouk, 2008) [21]:

$$f_{ESL}(y) = \frac{1}{2\sigma\sqrt{2}} \begin{cases} \exp\left\{-\left(\dfrac{y-\lambda}{\sigma\sqrt{2}(1+\varepsilon)}\right)\right\}, y \geq \lambda \\ \exp\left\{-\left(\dfrac{y-\lambda}{\sigma\sqrt{2}(1-\varepsilon)}\right)\right\}, y < \lambda \end{cases} \tag{2}$$

where λ, σ and ε are location, scale, and skewness parameters, respectively.

The distribution function (cdf) of y is given by:

$$F(y) = \begin{cases} 1 - \dfrac{1+\varepsilon}{2} \exp\left(\dfrac{y-\lambda}{\sqrt{2(1+\varepsilon)\sigma}}\right), y \geq \lambda \\ \dfrac{1-\varepsilon}{2} \exp\left(\dfrac{y-\lambda}{\sqrt{2(1-\varepsilon)\sigma}}\right), y < \lambda \end{cases} \tag{3}$$

and the quintile distribution function of y is:

$$Q_0(u) = \begin{cases} \lambda + \sqrt{2}(1+\varepsilon)\sigma \ln\left(\dfrac{2u}{1+\varepsilon}\right), 0 < u < (1+\varepsilon)/2 \\ \lambda + \sqrt{2}(1-\varepsilon)\sigma \ln\left(\dfrac{2(1-u)}{1+\varepsilon}\right), (1+\varepsilon)/2 \leq u < 1 \end{cases} \tag{4}$$

4. Variable Selection

4.1. Lasso Variable Selection Structure

The Lasso (Least Absolute Shrinkage and Selection Operator) is a linear model estimation method proposed by Tibshirani [16]. It refers to a group of methods that use an L_1 penalty to shrink parameter estimates and perform automatic variable selection. It is an L_1 penalised least squares regression. Like garrote, it shrinks some of the coefficients while setting the rest of them exactly to zero. Tibshirani considers the Lasso to be superior to ordinary least squares (OLS) regression for two reasons: Firstly, an over specified OLS model often has little bias but large variance, adversely affecting its prediction accuracy. This can be improved by shrinking or setting to zero some of the coefficients, trading some bias for a lower model variance. Secondly, OLS models may sometimes have a large number of small coefficients, adding little value to the model and complicating the interpretation of the effects.

The formally define the method for a linear regression model:

$$Y = X\beta + \varepsilon \tag{5}$$

where $Y=(Y_1, Y_2, ..., Y_n)^r$ is the n dimensional vector of observed responses, X is the $n \times p$. The design matrix, $\beta=(\beta_1, \beta_2, ..., \beta_p)^r$ is a p-dimensional vector of the unknown regression coefficients and $\varepsilon=(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^r$ is the vector of random errors from a multivariate normal distribution $N(0,\sigma^2 In)$. The columns of X are denoted as $(X_1, X_2, \ldots, X_P)$; these represent the p independent variables. The lasso estimate of β

$$\begin{aligned} \hat{\beta}^* &= ar\min\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \\ &= \textit{Residual sumof square+penalty} \end{aligned} \tag{6}$$

where $\sum_{j=1}^{p}\lambda\left|\beta_j\right|$ is called lasso penalty.

In this work we can employ Lasso method in the equation (1) to factorial experiment model as follows:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ &+ \beta_{12} x_1 x_2 + \ldots + \beta_{15} x_1 x_5 + \ldots + \beta_{23} x_2 x_3 + \ldots + \beta_{25} x_2 x_5 \\ &+ \beta_{34} x_3 x_4 + \beta_{123} x_1 x_2 x_3 + \ldots + \beta_{12345} x_1 x_2 x_3 x_4 x_5 \\ &+ \lambda\left[\sum_{j=1}^{31}\left[\left|\beta_0 + \beta_1 + \ldots + \beta_5 + \ldots + \beta_{12345}\right|\right]\right], \\ &j = 1, 2, \ldots, 2^5 - 1 \end{aligned} \tag{7}$$

For estimating the parameters in equation (7) we used the R program and the "Lasso2 package", [15].

4.2. Adaptive Lasso Variable Selection

In [16] Zou introduced a new version of the Lasso method [16] based on the adaptive weights which in turn lead to different penalization and to different coefficients in the ℓ_1 penalty. Adaptive Lasso, as a regularization method, avoids over fitting penalizing large coefficients. It has the same advantage as Lasso: it can shrink some of the coefficients to exactly zero. Zou [17] has given different weights to different coefficients. The adaptive lasso can be defined as:

$$\arg\min\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{i=1}^{p}wj\left|\beta j\right|\right\} \quad (8)$$

where ($w_1, w_2,..., w_p$) are the adaptive weights. Zou [17] has shown that if the weights are efficiently chosen in a data-dependent way then the adaptive lasso can achieve the oracle properties. He suggested the use of estimated weights, $\hat{w}_j=\left|\beta_j\right|^{-y}$, where β= {bj: $j = 1, p$} is a root-n-consistent estimator of β and $\gamma > 0$ is a user-chosen constant.

In this work we can employ adaptive Lasso method in the equation (1) to factorial experiment model as follows:

$$\begin{aligned}&y_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\\&\beta_{12}x_1x_2+...+\beta_{15}x_1x_5+...+\beta_{23}x_2x_3+...+\beta_{25}x_2x_5+\\&\beta_{34}x_3x_4+\beta_{123}x_1x_2x_3+...+\beta_{12345}x_1x_2x_3x_4x_5\\&+\lambda\left[\sum_{j=1}^{31}w_j\left[\left|\beta_0+\beta_1+...+\beta_5+...+\beta_{12345}\right|\right]\right]\end{aligned} \quad (9)$$

For estimating the parameters in equation (9) we used the program R. "parcor" package, [18].

4.3. Robust Lasso Factorial Experiment When the Response Variable Follows the ESL Distribution

Many researchers developed methods of robust regression shrinkage with selection methods (like lasso). In [6] Wang introduced the robust regression shrinkage and consistent variable selection through the LAD-Lasso, using the robust lasso with Huber loss function. In this section, we attempt to develop an integrated methodology to combine the variable selection (Lasso) method and robust method (Huber loss function) for the analysis of factorial experiments with two levels for each factor when the response variables follow ESL distribution.

Suppose that the response variable, y, follows epsilon-skew-Laplace distribution. The probability density in equation (2) can be expressed as:

$$f_{ESL}(y)=\frac{1}{2\sigma\sqrt{2}}e^{\left\{\left(\frac{-(y-\lambda}{\sqrt{2}(1+\epsilon)}\right)I_{(y\geq\lambda}+\left(\frac{-(y-\lambda}{\sqrt{2}(1-\epsilon)}\right)+I_{(y\geq\lambda}\right\}} \quad (10)$$

where λ represents the mean (location parameter) of the epsilon-skew Laplace, and ϵ

$$\begin{aligned}&\lambda_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\\&\beta_{12}x_1x_2+...+\beta_{15}x_1x_5+...+\beta_{23}x_2x_3+...+\beta_{25}x_2x_5+\\&\beta_{34}x_3x_4+\beta_{123}x_1x_2x_3+...+\beta_{12345}x_1x_2x_3x_4x_5\end{aligned} \quad (11)$$

As we mentioned in the section (4.1), it can be solved by using the technique lasso for analysis factorial experiment of type 2^5 design, as shown in equation (7).

Now, we will employ one of the robust methods (Huber loss function, 1973) also used by Yi and Huang, [19] in the regression models as follows:

$$h_\gamma(z)=\begin{cases}\dfrac{z^2}{2\gamma}, if\left|z\right|\leq\gamma\\ \left|z\right|-\dfrac{\gamma}{2}, if\left|z\right|>\gamma\end{cases} \quad (12)$$

where γ>0 is given constant. This function is quadratic for |z|≤γ and linear for |z|>γ.

As previously mentioned the Lasso formula used for factorial experiments analysis is given in the equation (7), and the Huber loss function of factorial experiment model is employed with Lasso variables selection. We can reach the following form:

$$\begin{aligned}&y_i=\frac{1}{n}\sum\nolimits_i h_\gamma\left[\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\right.\\&\beta_5x_5+\beta_{12}x_1x_2+...+\beta_{15}x_1x_5+...+\beta_{23}x_2x_3+...\\&\left.+\beta_{25}x_2x_5+\beta_{34}x_3x_4+\beta_{123}x_1x_2x_3+...+\beta_{12345}x_1x_2x_3x_4x_5\right]+\\&\lambda\left(\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_{12}+...+\beta_{15}+\right.\\&\left....+\beta_{23}+...+\beta_{25}+\beta_{34}+\beta_{123}+...+\beta_{12345}\right)\\&j=1,2,...,2^p-1\end{aligned} \quad (13)$$

where h_γ is the Huber loss function, and:

$$\begin{aligned}&y_i=\frac{1}{n}\sum\nolimits_i h_\gamma\left[\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\right.\\&\beta_5x_5+\beta_{12}x_1x_2+...+\beta_{15}x_1x_5+...+\beta_{23}x_2x_3+...\\&\left.+\beta_{25}x_2x_5+\beta_{34}x_3x_4+\beta_{123}x_1x_2x_3+...+\beta_{12345}x_1x_2x_3x_4x_5\right]+\\&\lambda\left(\left|\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5+\beta_{12}+...+\beta_{15}+\right.\right.\\&\left.\left.+...+\beta_{23}+...+\beta_{25}+\beta_{34}+\beta_{123}+...+\beta_{12345}\right|\right)\\&j=1,2,...,2^p-1\end{aligned} \quad (14)$$

where 2^p-1=2^5-1=31 which effects the factors. For estimating the parameters in equation (14) we used the program R and "hqreg" package [20].

5. Application

5.1. Simulation Study

In this section, the simulation factorial experiments with

non-normal responses which follow Epsilon-skew Laplace distributions are presented to illustrate the performance of the proposed approach. The simulation experiments are generated by the R program. The selected factorial experimental design is used to study five factors with two-level referred to as symbols A, B, C, D and E with interactions. For these factors, we represent the parameters (β_1, β_2,...β_{12},...β_{12345}). To compare the effectiveness of the robust lasso variable selection with lasso and adaptive lasso methods, we used the mean square error (MSE) to determine the most appropriate method for such data. For the model we used 1000 iterations generated for each combination of factors A, B, C, D, and E, represent by the parameters (β_1, β_2,...β_{12},...β_{12345}). Thee simulation of the program R is as follows:

Step 1: Generate the response data based on epsilon skew Laplace, using equation (4).

Step 2: For the lasso method, find the estimate parameter and selection variable (factors) using equation (7).

Step 3: For the adaptive lasso method, find the estimate parameter and selection variable (factors) using equation (9).

Step 4: For the robust lasso method, find the estimate parameter and selection variable (factors) using equation (14).

Step 5: Calculate the MSE and R Square (R^2).

Table 2. Estimate and selection effects on simulation data for Factorial 2^5 analysis design (n=100)

Effects	n=100		
	Variable selection methods		
	Lasso	Adaptive lasso	Robust lasso
β_1	6.83367232	6.26481613	16.086549
β_2	22.0452930	27.4054739	31.6505038
β_3	-44.7039743	-53.8209557	-54.309208
β_4	-283.97093	-295.867813	0.0000000
β_5	-82.493880	-93.1401845	0.0000000
β_{12}	27.7706178	34.7388770	37.3759174
β_{13}	-7.76260452	-7.2681230	0.0000000
β_{23}	12.1310295	12.5612774	21.7363291
β_{14}	3.28801061	2.1791133	11.6394281
β_{34}	2.06046671	1.0122355	9.4994640
β_{15}	14.2370394	15.7656099	23.8423390
β_{25}	1.31654538	0.53402047	7.64578812
β_{35}	-14.6154471	-16.371540	0.0000000
β_{45}	0.00000000	0.00000000	0.0000000
β_{123}	10.0221410	9.82391687	19.5854276
β_{124}	2.43434588	1.33909469	10.2246349
β_{134}	31.4868761	39.1483128	0.0000000
β_{234}	0.94153788	0.33654135	6.55810299
β_{125}	0.05420517	0.00000000	0.91925163
β_{135}	0.21390105	0.02558244	3.09451607
β_{235}	-1.00669629	-0.3406328	0.0000000
β_{145}	0.00000000	0.00000000	0.0000000
β_{245}	0.20188900	0.02748814	2.8549468
β_{345}	-0.06672547	0.00000000	000000000
β_{1234}	0.21783534	0.03653715	3.17136184
β_{1235}	-0.39383602	-0.1465863	-4.3039569
β_{1245}	-0.39047734	-0.13683801	-4.2848700
β_{1345}	0.0000000	0.00000000	0.0000000
β_{2345}	-0.16886899	-0.01353800	-2.4071656
β_{12345}	1.80071185	0.81554076	0.0000000
MSE	0.8712517	0.7111041	0.7027808
R^2	0.5918896	0.6825224	0.7577845

Table 3. Estimate and selection effects on simulation data for Factorial 2^5 analysis design (n=200)

Effects	n=200		
	Variable selection methods		
	Lasso	Adaptive lasso	Robust lasso
β_1	5.3436728	5.2295481	11.023741
β_2	19.023620	20.434235	26.561256
β_3	-51.359770	-62.238554	-58.286180
β_4	-303.54380	-354.54980	0.00000000
β_5	-72.432918	-98.432180	0.00000000
β_{12}	22.5489421	28.7689210	31.673420
β_{13}	-8.8531091	-8.3984320	0.0000000
β_{23}	10.1342187	10.431289	19.542398
β_{14}	2.45892100	1.78341201	9.7834129
β_{34}	1.78432090	0.9543210	8.2319805
β_{15}	13.6734212	13.783421	21.732109
β_{25}	0.43678910	0.45620923	6.4281258
β_{35}	-14.615447	-16.371540	0.0000000
β_{45}	0.0000000	0.00000000	0.0000000
β_{123}	9.0894143	8.76320900	18.321986
β_{124}	2.0432878	1.0543209	9.0432195
β_{134}	29.673210	35.983218	0.0000000
β_{234}	0.8315429	0.30432109	5.3219879
β_{125}	0.0254210	0.00000000	0.87632103
β_{135}	0.1165987	0.01652103	2.08431090
β_{235}	-2.0463282	-0.5052108	0.0000000
β_{145}	0.0000000	0.00000000	0.0000000
β_{245}	0.1943219	0.01765321	1.7632109
β_{345}	-0.1873254	0.00000000	0.000000
β_{1234}	0.19832153	0.02873211	2.1542098
β_{1235}	-0.4760432	-0.7620963	0.0000000
β_{1245}	-0.4873572	-0.2764301	-5.2109970
β_{1345}	0.0000000	0.00000000	0.0000000
β_{2345}	-0.2453094	-0.27635402	-3.0127956
β_{12345}	0.95637832	0.79832801	0.0000000
MSE	0.8012517	0.705692	0.6803218
R^2	0.6098216	0.6978234	0.7843215

Table 2. and Table 3. summarize the results of the three methods (Lasso, adaptive lasso, robust lasso) for estimating

and selecting factors to the factorial experiment model consisting of five factors with two levels for each factor. We can see that some parameters have zero for all three methods, but robust lasso method gave better results. This is evident through the mean square error simulation experiment, as the mean square error was less likely to compare with other methods. This indicates that the robust lasso method is more likely to be relevant in the explanation of the model and factors compared to the other methods mentioned earlier.

5.2. Method Illustration with Sample Data

A sample data was collected from Women and Children Hospital in Diwaniyah in Iraq. The size of the sample included 64 cases of newborns of both sexes aged 1 to 10 days, all of these being children with early neonatal jaundice. The aim is to determine the most important factors leading to the disease of jaundice in newborns (Neonatal jaundice). The factorial experiments have been used to determine the most important factors affecting the percentage of jaundice in newborns, which represents the response variable (y) with the interactions of these factors at two levels for each factor: high level (+1) and low level (-1). There are 26 common factors are produced from the five main factors: diabetes mellitus of the mother, blood pressure, duration of pregnancy, mother anemia and child weight at birth. There are 10 two-factor interactions, 10 of the three-way interactions, 5 of the four-way interactions and one five-way interaction between the factors.

Table 4. Factors and levels for each factor

Factors	Factor Levels	
	Low level: -1	High level: +1
A=x_1: diabetes mellitus	90 mg/dl.	140 mg/dl.
B=x_2 blood pressure	100/60 mm Hg	140/90 mm Hg
C=x3: duration of pregnancy	≤36 weeks	>36 weeks
D=x_4: anemia	Anemia ≤11 g/dl.	Anemia >11 g/dl.
E=x_5: child weight at birth	≤2.5 Kg	>2.5 Kg

The main factors ant their interactions are detailed below:

a. Main Factors

A Diabetes mellitus – this factor has two levels: high level diabetes, the sugar level >140 mg/dl. And low level diabetes, the sugar level <90 mg/dl.

B Blood pressure - this factor his two levels: high level, the blood pressure level 140/90 mm Hg, and low level, the blood pressure level 100/60 mm Hg.

C Duration of pregnancy - this factor his two levels: high level, the duration of pregnancy >36 weeks, and low level, the duration of pregnancy ≤36 weeks.

D Anemia - this factor has two levels: high level, the anemia level >11 g/dl, and low level, the anemia level ≤11 g/dl.

E Child weight at birth - this factor his two levels: high level, child weight > 2.5 Kg, and low level, child weight≤2.5 Kg.

b. Two-factors interactions

AB Two-factor interaction between diabetes mellitus and blood pressure.

AC Two-factor interaction between diabetes mellitus and duration of pregnancy.

BC Two-factor interactions between blood pressure and duration of pregnancy.

AD Two-factor interaction between the diabetes mellitus and anemia.

BD Two-factor interaction between blood pressure and anemia.

CD Two-factor interaction between duration of pregnancy and anemia.

AE Two-factor interaction between the diabetes mellitus and child weight at birth.

BE Two-factor interaction between blood pressure and child weight at birth.

CE Two-factor interaction between duration of pregnancy and child weight at birth.

DE Two-factor interaction between anemia and child weight at birth.

c. Three-factors interactions

ABC Three-factor interaction between diabetes mellitus, blood pressure and duration of pregnancy.

ABD Three-factor interaction between diabetes mellitus, blood pressure and anemia.

ACD Three-factor interaction between diabetes mellitus, duration of pregnancy and the anemia.

BCD Three-factor interaction between blood pressure, duration of pregnancy and anemia.

ABE Three-factor interaction between diabetes mellitus, blood pressure and child weight at birth.

ACE Three-factor interaction between diabetes mellitus, duration of pregnancy and child weight at birth.

BCE Three-factor interaction between blood pressure, duration of pregnancy and child weight at birth.

ADE Three-factor interaction between diabetes mellitus, anemia and child weight at birth.

BDE Three-factor interaction between blood pressure, anemia and child weight at birth.

CDE Three-factor interaction between duration of pregnancy, anemia and child weight at birth.

d. Four-factors interactions

ABCD Four-factor interaction between diabetes mellitus, blood pressure, duration of pregnancy and anemia.

ABCE Four-factor interaction between diabetes mellitus, blood pressure, duration of pregnancy and child weight at birth.

ABDE Four-factor interaction between diabetes mellitus, blood pressure, anemia and child weight at birth.

ACDE Four-factor interaction between diabetes mellitus, duration of pregnancy, anemia and child weight at birth.

BCDE Four-factor interaction between blood pressure, duration of pregnancy, anemia and child weight at birth.

e. Five-factors interactions

ABCDE Five-factor interaction between diabetes mellitus, blood pressure, duration of pregnancy, anemia and child weight at birth.

y_i is the response variable, which represents the percentage of jaundice in newborns.

5.3. Data Analysis and Results

The data were analyzed using R program in order to determine the most important factors affecting the increase in the incidence of jaundice disease in newborns. We used the robust lasso variable selection method to determine the most important factors leading to the disease of jaundice in newborns. Data analysis results appear in a single table, but it has been divided into several tables for further clarification. Table 5., Table 6., Table 7., and Table 8. show the results obtained:

Table 5. Main effects (factors)

Factor	A	B	C	D	E
Robust lasso	-0.18357	-0.371071	0.3085714	0.1835534	0.0000000

Table 6. Two-factor interaction

Factor	AB	AC	BC	AD	BD
Robust lasso	-0.18371	-0.121064	-0.308571	0.0000000	0.0585611
Factor	CD	AE	DE	CE	BE
Robust lasso	-0.21061	0.3085715	-0.121064	0.3085715	-0.210614

Table 7. Three-factor interactions

Factor	ABC	ABD	ACD	BCD	ABE
Robust lasso	-0.12101	-0.246057	0.1835121	0.000000	0.18357110
Factor	ACE	BCE	ADE	BDE	CDE
Robust lasso	-0.37115	-0.433525	0.1218154	0.3085713	0.000000

Table 8. Four and five-factor interactions

Factors	ABCD	ABCE	ABDE	ACDE	BCDE	ABCDE
Robust lasso	-0.58571	0.000000	0.00000	0.0588	-0.24605	0.433921

Main effects

The value for factor E (child weight at birth) are equal to zero, indicating that they are not significant and do not lead to the appearance of the disease of jaundice in newborns. We also note that A (diabetes mellitus), B (blood pressure), C (duration of pregnancy), and D (anemia) represent the main factors having major effects and leading to the appearance of the disease of jaundice in newborns.

Two-factor interactions

Through the results of Table 6 we note the interactions AD (diabetes mellitus and anemia) we observe that their value are equal to zero. This indicates that they are not significant and do not lead to the appearance of the disease of jaundice in newborns. We also note that factor interactions AB (diabetes mellitus and blood pressure), AC (diabetes mellitus and duration of pregnancy), BC (blood pressure and duration of pregnancy), BD (blood pressure and anemia), CD (duration of pregnancy and anemia), AE (diabetes mellitus and child weight at birth), DE (anemia and child weight at birth), CE (duration of pregnancy and child weight at birth) and BE (blood pressure and child weight at birth) have significant effect and lead to the appearance of the disease of jaundice in newborns.

Three-factor interactions

From the Table 7, we find that the values of interactions BCD (blood pressure, duration of pregnancy and anemia) and CDE (duration of pregnancy, anemia and child weight at birth) have values equal to zero. This indicates that they are not significant and do not lead to the appearance of the disease of jaundice in newborns. We also note that factor interactions ABC (Diabetes mellitus, blood pressure and duration of pregnancy), ABD (diabetes mellitus, blood pressure and anemia), ACD (diabetes mellitus, duration of pregnancy and anemia), ABE (diabetes mellitus, blood pressure and child weight at birth), ACE (diabetes mellitus, duration of pregnancy and child weight at birth), BCE blood pressure, duration of pregnancy and child weight at birth), ADE (diabetes mellitus, anemia and child weight at birth) and BDE (blood pressure, anemia and child weight at birth), have significant effect and lead to the appearance of the disease of jaundice in newborns.

Four and five -factor interactions

Table 8. presents the results of four-factor interactions, from the results we can see that the ABCE (Diabetes mellitus, blood pressure, duration of pregnancy and child weight at birth) and ABDE (diabetes mellitus, blood pressure, anemia

and child weight at birth) has the value equal to zero, indicating that it is not significant and does not lead to the disease of jaundice in newborns. We also note that the factor interactions ABCD (diabetes mellitus, blood pressure, duration of pregnancy and anemia), ACDE (diabetes mellitus, duration of pregnancy, anemia and child weight at birth), and BCDE (blood pressure, duration of pregnancy, anemia and child weight at birth) have significant effect and lead to the disease of jaundice in newborns. And also we can note that the value of five-factors interactions (ABCDE) (diabetes mellitus, blood pressure, duration of pregnancy, anemia and child weight at birth) has the value different than zero. This means that it is significant and leads to appearance of the disease of jaundice in newborns.

6. Conclusions

In this paper we have presented a new methodology to analyze factorial experiments when the response variable follows epsilon skew Laplace distribution. We employed a Huber function with lasso variable selection (robust lasso), through the results simulation and real data. We observed that the proposed method (robust lasso) gave better results than lasso and adaptive lasso methods. This inference is clear from the results of the MSE and R squared in the simulation experiments as well as in the results of the application of real data, we observed that the robust lasso method can be used to estimate and select all five main effects, all 10 two-factor interactions, 10 three-factor interactions, 5 four-factor interactions and one five-factor interaction. Through the obtained results, we have reached a set of results pertaining to the study of the influence of the main factors and interactions that lead to the disease of jaundice in newborns and the significant factors with interactions are:

a. Main effects for factors

A (diabetes mellitus), B (blood pressure), C (duration of pregnancy) and D (anemia).

b. Two-factor interactions

Interactions factors AB (diabetes mellitus and blood pressure), AC (diabetes mellitus and duration of pregnancy), BC (blood pressure and duration of pregnancy), BD (blood pressure and anemia), CD (duration of pregnancy and anemia), BE (blood pressure and child weight at birth), AE (diabetes mellitus and child weight at birth), CE (duration of pregnancy and child weight at birth) and DE (anemia and child weight at birth).

c. Three-factor interactions

Interactions factors ABD (diabetes mellitus, blood pressure and anemia), ABE (diabetes mellitus, blood pressure and child weight at birth), ACD (diabetes mellitus, duration of pregnancy and anemia), ABE (diabetes mellitus, blood pressure and child weight at birth), ACE (diabetes mellitus, duration of pregnancy and child weight at birth), BCE blood pressure, duration of pregnancy and child weight at birth), ADE (diabetes mellitus, anemia and child weight at birth) and BDE (blood pressure, anemia and child weight at birth).

d. Four and five-factor interactions

Interactions factors ABCD (diabetes mellitus, blood pressure, duration of pregnancy and anemia), ACDE (diabetes mellitus, duration of pregnancy, anemia and child weight at birth) and BCDE (blood pressure, duration of pregnancy, anemia and child weight at birth) have significant effect and lead to the disease of jaundice in newborns. And also we can note that the value of five-factors interactions (ABCDE) (diabetes mellitus, blood pressure, duration of pregnancy, anemia and child weight at birth).

REFERENCES

[1] C. Daniel, "Locating outliers in factorial experiments," Technometrics, vol. 2, no. 2, pp. 149-156, 1960.

[2] P.J. Huber, Robust Statistics, 1st ed., New York: Wiley, SBN-10: 0471418056, pp: 308, 1981.

[3] P. J. Huber, "Robust regression: Asymptotics, conjectures and Monte Carlo," The Annals of Statistics, vol. 1, *ed.* 5, pp. 799–821, 1973.

[4] G. M. Oyeyemi, "Treatment of non-normal responses from designed experiments," Department of Statistics University of Iorin, Nigeria, JNSA, no. 17, pp. 8-19, 2004.

[5] A. Hutson, "Utilizing the flexibility of the epsilon-skew-normal distribution for common regression problems," Journal of Applied Statistics, vol. 31, no. 6, pp. 673-683, 2004.

[6] H. Wang, G. Li, and G. Jiang, "Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso," Journal of Business & Economic Statistics, vol. 25, no. 3, pp. 347-355, Jul. 2007.

[7] H. V. Kulkarni, and V. V. Patil, "A comparative study of analyses of 2^n factorial experiments with a Poisson distributed response variable," Communications in Statistics-Simulation and Computation, vol. 39, no. 8, pp. 1530-1547, 2010.

[8] S. C. Patil, and H. V. Kulkarni, "Analysis of 2^n factorial experiments with exponentially distributed response variable," Applied Mathematical Sciences, vol. 5, no. 10, pp. 459-476, 2011.

[9] H. Jafari, and R. Hashemi, "Optimal designs in a simple linear regression with skew-normal distribution for error term," Journal of Applied Mathematics, vol. 1, no. 2, pp. 65-68, 2011.

[10] A. Kane, and A. Mandal, "A new analysis strategy for designs with complex aliasing," Submitted to the American Statistician, 2016.

[11] B. K. Mohammed, M. Roman, M. H. Odah, and A. S. M. Bager, "Testing reliability: factorial design with data from a log-epsilon-skew-normal distribution," Economic Computation & Economic Cybernetics Studies & Research, vol. 51, no. 3, 2017.

[12] D. C. Montgomery, Design and Analysis of Experiments, New York: John Wiley & Sons, 2001, pp. 64-65.

[13] R. L. Mason, R. F. Gunst, and J. L. Hess, Statistical Design and Analysis of Experiments: with Applications to Engineering and Science, vol. 474, John Wiley & Sons, 2003.

[14] D. C. Montgomery, Design and Analysis of Experiments, 7th Edition, New York: Wiley, 2009.

[15] J. Lokhorst, B. Venables, and B. Turlach, Lasso2. constrained estimation aka "lasso". R package version 1.2-19, 2015.

[16] R. Tibshirani, "Regression shrinkage and selection via the Lasso," Journal of the Royal Statistical Society, series B, no. 58, pp. 267–288, 1996.

[17] H. Zou, "The adaptive lasso and its oracle properties," Journal of the American Statistical Association, vol. 101, pp. 1418–1429, 2006.

[18] N. Kraemer, and J. Schaefer, Parcor: Regularized estimation of partial correlation matrices R package version 0.2-6, 2015.

[19] C. Yi, and J. Huang, "Semi smooth newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression," Journal of Computational and Graphical Statistics, vol. 26, no. 3, pp. 547-557, 2017.

[20] C. Yi, Hqreg: Regularization Paths for Lasso or Elastic-Net Penalized Huber Loss Regression and Quantile Regression, R package version 1.4, 2017.

[21] Elsalloukh, The Epsilon-Skew Laplace Distribution. In B. Section, editor: The 2008 proceedings of the American Statistical Association, 2008.