AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Lasso Factorial Design Quantile Regression Model with Application

[1]Fadel Hamid Hadi Alhusseini, [2]Bahr k. Mohammed, [3]Saif Hosam Raheem

[1]*PhD candidate in Faculty of Economics and Business Administration, University of Craiova, (Department of Statistics and Economic Informatics), Craiova, Romania.*
[2]*Faculty of Economics and Business Administration, Bucharest University of Economic Studies, (Department of Statistics and Econometrics), Bucharest, Romania.*
[3]*Faculty of Economics and Business Administration, Al-Qadiseya University, Department of Statistic, Iraq*

Address For Correspondence:
Fadel Hamid Hadi Alhusseini, Department of Statistics and Economic Informatics, College of Economics and Business Administration, University of Craiova, Romania

ARTICLE INFO

Article history:
Received 18 December 2016
Accepted 16 February 2017
Available online 28 February 2017

Keywords:
factorial experiment, quantile regression, design quantile regression, Lasso

ABSTRACT

In this paper, we propose a new approach that integrates the topic of factorial experiment and regression model after transforming the treatments to new variables represent all main variables and all possible Interactions for building new model and do variables selection for independent variables inside of model. from selected independent variables, it is possible to build a regression model. The one line regression by using quantile regression was employed. To obtain relevant variable selection using the lasso approach with factorial design quantile regression model (used for retrieving the important variable with effect in blood pressure). The main aim is focus on important variable and exclude unimportant variables from structure model at same time. In this paper, we will use four Lasso Factorial design quantile Regression levels according to quantile ratios (0.20, 0.40, 0.60, 0.80) respectively. From results, we conclude that Lasso Factorial design quantile Regression levels at quantile ratio (0.20, 0.40) present weakness in representation of data under study because of the value of the pseudo-R squared is lower than 0.50. But Lasso Factorial design quantile regression levels at quantile ratio (0.60, 0.80) is stronger in representation of the data under study because the value of the pseudo-R squared is greater than 0.50%. Therefore, we focus in Lasso Factorial design quantile Regression model under levels (0.60, 0.80) in explanation. Also, from results we see that the methods under study can make variable selection as follows: Lasso Factorial design quantile Regression at level (0.20) can exclude 12 unimportant independent variables for constriction model, Lasso Factorial design quantile Regression at level (0.40) can exclude 11 unimportant independent variables for constriction model in an algorithm in program R was built, based on the package (quantreg). Lasso Factorial design quantile Regression at level (0.60) can exclude 12 unimportant independent variables for constriction model. Lasso Factorial design quantile Regression at level (0.80) can exclude 14 unimportant independent variables for constriction model. For analyzing these result, a new algorithm in program R was built, based on the package (quantreg).

INTRODUCTION

In spite of (Fisher, 1926), first credited the development of factorial design and analysis, but is (Yates, 1937)who has great credit in promoting the development and analysis of the factorial experiment. Yates (1937) used the method of statistical analysis of the factorial experiments of a wide range of type 2^2 , 3^2. This is a very difficult method, especially when the factors involved in the experiment are numerous. Searle (1971) submitted the dividing of the variance, which is due to factors input in the experiment. These factors consist of a number

To Cite This Article: Fadel Hamid Hadi Alhusseini, Bahr k. Mohammed, Saif Hosam Raheem., Lasso Factorial Design Quantile Regression Model with Application. ***Aust. J. Basic & Appl. Sci.,*** *11(2): 104-114, 2017*

of levels and describing data in a linear model parameters that can be estimated in different ways. One of these methods is the least squares estimate which is usually used in the ANOVA estimates. Kirk (1995) explained the interpretation of factorial experiment and how to identify the main effects and interactions of all the experiment. The combining among level of factor with every level of the other factor inside with him in the experiment. Tinsson and Dossou (2005) had used experimental designs and adapted general linear model; and explained the function orthogonally linking method from the design matrix which was obtained under the assumption of a Gaussian. The authors investigated its properties using simulation [13]. Luis et.al. (2012) used beta regression models from frequents and Bayesian perspectives to estimate means response for 2^{k-p} factorial experiment with response (0, 1). The study compared it to normal regression models, and the results showed that the coverage of intervals with logit transformation was as good as the best function link for beta regression. Also, Bayesian and frequents lengths of intervals are similar for legit link. The Bayesian model with probit link presented the worst performance with respect to lengths of intervals. The main goal of the experiential design is to test the significance of the main factors that affect the response variable (treatments). Factorial design is considered extensive when compared to experiments design, because it considers the interaction between the main factors (Wu and Hamada, 2009). From known, the number of treatments is linked to the number of factors and levels by positive relationship. Sometimes, the number of treatments reaches a huge value; therefore, dealing with all treatments can be quite challenging. It is though possible to transform these treatments into new variable as covariates and then use these covariates for building quantile regression model, as proposed by Koenker and Bassett (1978). Hence, we use the lasso method to determine the important variables in factorial design quantile regression model. The variable selection methods were discovered at the middle of the last century. There are many classical variables selection methods that can be used to decide the important variables in the model, such as AIC (Akaike, 1973) and BIC (Schwarz, 1978). But all these traditional methods are inadequate and have many drawbacks, such as long time to achieve variables selection, because they deal with 2^p models, where p is the number of covariates. Many variable selection methods have been introduced during the past decades. One type of these methods are the lasso methods which were proposed by Tibshirani (1996). The lasso method has some advantages: firstly, it does not need a long time for performing variable selection and secondly it improves predictive precision through reducing the variance for model coefficients. Generally, the lasso method makes coefficient estimation and variable selection automatic. It is used in many fields of statistics. The first researcher who employed the lasso method in quantile regression model (QReg) was Koenker (2004).

In the present study, the lasso method was employed in the factorial design quantile regression model for selecting important variables in blood pleasure. Six factors are used (A, B, C, D, E, F), more details being in section five where each factor has two levels (high and low). Therefore, the number of treatments is equal $(2^6 - 1 = (2\times2\times2\times2\times2\times2) - 1 = 63$ treatments. These treatments are dealt with as covariates. From this approach, we will build a new quantile regression model containing one response variable and 63 covariates. Not all the covariates have effects on blood pleasure (response variable); therefore, we employ the lasso method for selecting strong covariates through our algorithm. The paper is organized as follows: section two presents the factorial experiment with level (2X2); we study the factorial design of the quantile regression model in section three; section four includes the study pf the lasso factorial design of the quantile regression model; in section five a data sample and analysis is presented and a brief conclusion is included in section six.

MATERIALS AND METHODS

Factorial Experiment with Level (2X2):

Design of experiment (DOE) analysis focuses on understanding the effects of a set of covariates on the response variable. DOE is employed for finding cause-and-effect relationships between a set of covariates and the response variable. In the methodology of DOE (Mettas and Guo, 2007), the covariates and dependent variables are named factor and response variable, respectively. Each factor has possible values (levels) and the experimental unit takes all combinations of these values (levels) over all factors. The set of factors is called treatment, and the combination of treatments is named replication. Factorial experiments study the effects of interaction between factors on the response variable. The presence of interactions in the factorial experiment can have important implications on data interpretation in the experiment. Generally, the number of treatments in the factorial experiment is the multiplication of the number of levels for each factor across all the experiment factors. For example, suppose we should conduct an experiment with two factors, factor one with M levels, and factor B with N levels. The number of treatments is $(N\times M)$ and the experiment is factorial design. The full factorial experiment comes with all the combinations $(N\times M)$ and the partial factorial experiment comes with some of these combinations. In the full factorial experiment, all the factors and their interactions can be investigated, while in the partial factorial experiment, some of the interactions can be eliminated, where part of the treatment cannot be run. If there are (2) factors and (3) levels, then the number of treatment is $(3\times3 = 9)$. If there are 3 factors and 3 levels, then the number of treatment is $(3\times3\times3 = 27)$ and so on.

In the foregoing experiments performed either in Completely Randomized Design (C.R.D). Randomized Block Design (R.B.D) or Latin Square Design (L.S.D.), we were primarily concerned with the comparison and estimation of the effects of a single set of treatments like varieties of wheat, manure of different methods of cultivation etc. Such experiments which deal with one factor only may be called simple experiments. In factorial experiment, as the adjective factorial indicates, the effects of several factors of variation are studied and investigated simultaneously, the treatment being all the combinations of different factors under study. In these experiments an attempt is made to estimate the effects of each of the factors and the interaction effects. The variation in the effect of one factor is a result to different levels of other factors (Kapoor and Gupta, 2010).

Types of Factorial Experiments:

The simplest factorial experiment can be illustrated as an experiment that studies the effects of two factor, each with two levels. This experiment is called (2^2) factorial experiment - the base stands for number of levels and the exponent represents the number of factors. Therefore, 2^n represents a factorial experiment with (n) factors, each taking two levels. The factorial experiment (2×4) means that there are two factors, the first one has two levels and the second factor has 4 levels.

2^2 Factorial Design:

In this type of factorial design there are two factors, each at two levels, so that there are 2×2=4 treatment combinations in all. Following the notations proposed by Yates, the capital letters *A* and *B* indicate the names of the two factors under study and the small letters *a* and *b* represent one of the two levels of each of the corresponding factors (called the second level). The first level of *A* and *B* is generally expressed by the absence of the corresponding letter in the treatment combinations. The four treatment combinations can be enumerated as follows (Kapoor and Gupta, 2010).

a_0b_0 or 1 : factors *A* and *B*, both at first level
a_1b_0 or a : *A* at second level and *B* at first level
a_0b_1 or b : *A* at first level and *B* at second level
a_1b_1 or ab : *A* and *B* both at second level

These four treatment combinations can be compared by laying out the experiment in C.R.D., with *r* replication, for example, each replicate containing 4 units. ANOVA can be carried out accordingly. In the above cases, there are 3 *d.f.* associated with the treatment effects. In factorial experiment our main objective is to carry out separate tests for the main effects *A*, *B* and the introduction *AB*, splitting the treatment *S.S* with 3 *d.f.* into three orthogonal components each with 1 *d.f.* and each associated either with the main effects *A* and *B* or the introduction.

General Full Factorial Designs:

Experiments with two or more factors are encountered frequently. The best way to carry out such experiments is by using full factorial experiments (Douglas, 2001). These are experiments in which all combinations of factors are investigated in each replicate of the experiment. Full factorial experiments are the only means to completely and systematically study interactions between factors in addition to identifying significant factors. One-factor-at-a-time experiments (where each factor is investigated separately by keeping all the remaining factors constant) do not reveal the interaction effects between the factors. Further, in one-factor-at-a-time experiments, full randomization is not possible. The model and analysis of a multi-way factorial are similar to those of a two-way factorial. In experiments, which owns more than a factor and several levels. factorial experiment must be dependent on a kind of design. We will use a completely random design for this type of experiment (2^6), CRD is the name given to this design, because it is randomly factorial distribution of treatment on the experimental units. After that is allocated experimental plots per treatment (Douglas, 2001). This design is used extensively because it is the simplest kinds of designs, which is the foundation for the construction of the experimental designs. When the experimental units are homogeneous.

Factorial Design Quantile Regression Model Study:

Factorial design is important for studying the models built by using two or more factors. Also, factorial designs have the advantage of being able to measure the overlapping effect among all factors through interaction. The measurement of the response variable is affected by changing the levels of factor - the effect of the main factor is quite different compared to its interaction with other factors. Another possible approach is using the main effects and interaction effects for two-level factorial designs as regression model after transforming the main factor and interaction into new variables, called explanatory variables. Assuming we have a full factorial design with six factors (A, B, C, D, E and F) with two levels, there are 6^2 = 64 treatments divided according to the following: 6 main factors (A, B, C, D, E and F), 15 two-factor interactions (AB, AC, AD, AE,AF, BC, BD, BE, CD, CE , DE, …EF), 20 three-factor interactions (ABC, ABD, ABE, ACD, ACE, ADE,

BCD, BCE, BDE , CDE,…DEF), 15 four-factor interactions (ABCD, ABCE, ACDE, BCDE , ABDE,…CDEF) and 6 five-factor interactions (ABCDE,…BCDEF) and one -factor interaction (ABCDEF).

The equations below are generally used to determine the number of effects:

$$type\ of\ effects = C_i^N = \frac{N!}{i!\,(N-i)!}\ .\ were\ N\ total\ \ number\ of\ factors\ , i\ number\ of\ effects$$

According to our assumption, it is possible to use another approach to show the main effect and interaction effects for two-level factorial in response variable. This approach is using the regression model. Therefore, the regression model is structured through six factors (A, B, C, D and E). The interaction will take the following formula (Wu and Hamada, 2009):

$$\begin{aligned} y_i = {} & \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_2 \\ & + \beta_{13} x_1 x_3 . + \beta_{123} x_1 x_2 x_3 + \beta_{124} x_1 x_2 x_4 + \cdots + \beta_{1234} x_1 x_2 x_3 x_4 \\ & + \beta_{1345} x_1 x_3 x_4 x_5 + \beta_{1245} x_1 x_2 x_4 x_5 \ldots + \beta_{123456} x_1 x_2 x_3 x_4 x_5 x_6 \\ & + u \ \ , j = 1,2, \ldots . 2^N - 1 \end{aligned} \tag{1}$$

where

y_i is the response variable ($i=1, 2, \ldots\ldots n$)

$(x_1, x_2, x_3, x_4, x_5, x_1x_2, x_1x_3 \ldots\ldots, x_1x_2x_3x_4x_5)$ are covariates

β_o is intercept

$(\beta_1, \beta_2, \ldots\ldots, \beta_{12345})$ are regression parameters

N is the number of factors (here equal to 5, according the assumption)

u_i is a random error term distributed according to normal distribution with mean zero and variance σ^2.

In this paper, we will build the treatments of the response variables through quantile regression for structure factorial design quantile regression, which takes the following formula:

$$\begin{aligned} y_i = {} & \beta_o + \beta_{(\tau)1} x_1 + \beta_{(\tau)2} x_2 + \beta_{(\tau)3} x_3 + \beta_{(\tau)4} x_4 + \beta_{(\tau)5} x_5 + \beta_{(\tau)12} x_1 x_2 \\ & + \beta_{(\tau)13} x_1 x_3 . + \beta_{(\tau)123} x_1 x_2 x_3 + \beta_{(\tau)124} x_1 x_2 x_4 + \cdots \\ & + \beta_{(\tau)1234} x_1 x_2 x_3 x_4 + \beta_{(\tau)1345} x_1 x_3 x_4 x_5 + \beta_{(\tau)1245} x_1 x_2 x_4 x_5 \ldots \\ & + \beta_{(\tau)123456} x_1 x_2 x_3 x_4 x_5 x_6 + u_\tau \ \ , j = 1,2, \ldots . 2^N - 1 \end{aligned} \tag{2}$$

where τ ,$0 < \tau < 1$.

To estimate the parameters of the model in equation (2) linear programming algorithm can be used (Koenker and Orey, 1987). In this paper, we are using the lasso approach for coefficient estimation and variable selection at the same time.

Lasso Factorial Design Quantile Regression Model:

Variable selection is a statistic tool which is used for discriminating between important and unimportant independent variables in regression models. One of these models is the quantile regression model (QR), which was introduced by Koenker and Bassett (1978). This type of regression model is not required to satisfy least squares assumptions. For instance, the error terms are not necessary to be in normal distribution. Moreover, QR is not sensitive to the presence of vertical outliers in dataset. The QR model provides full coverage to all observations of response variable, thus it splits the data into two subsets based on the value of τ where $0 < \tau < 1$.

The general form of QR is given by:

$$T_i = x_i^t \beta_\tau + u_i \qquad \tau \in (0,1) \tag{3}$$

where

β_τ is a vector of unknown quantities

τ is the quantile level, th

u_i is the residual terms with density function restricted to τth quantile equal to 0

The estimation of quantile regression coefficients β_t can be computed by minimizing the following equation:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^t \beta_\tau) \tag{4}$$

where $\rho_\tau(u)$ is the check function defined by:

$$\rho_\tau(u) = \begin{cases} \tau u & if\ u \geq 0 \\ -(1-t)u & if u < 0 \end{cases} \tag{5}$$

Since the equation (3) is not differentiable at the origin, there is no exact solution for equation (3). The equation (3) can be solved by using linear programming algorithm (Koenker and Orey, 1987). The variable selection methods can improve forecasting precision. Also, it is often used for identifying a small subset of covariates from a large set of covariates to obtain improved explanations. Many researchers have studied the subject of variable selection. For instance, Tibshirani (1996) proposed the lasso method (least absolute shrinkage and selection operator) in linear regression. By this technique variable selection and coefficients

estimation can be achieved concomitantly. The lasso method is represented by L_1-norm (reglarizion). The lasso estimation can be achieved through the following equation:

$$\hat{\beta}^{lasso} = \underset{\beta}{minimize} \quad \left\|y - X\hat{\beta}\right\|_2^2 + \lambda\|\beta\|_1 \tag{6}$$

here $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$

The equation (4) is reformulated as in equation (5):

$$\hat{\beta}^{lasso} = \text{minimize}\sum_{i=1}^{n}\left(y - X\hat{\beta}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad ,\lambda \geq 0 \tag{7}$$

Koenker (2004) is the first researcher who employed lasso penalty method in quantile regression model using the following formula:

$$\underset{\beta_\tau}{min}\sum_{i=1}^{n}\rho_\tau\left(T_i - x_i^t\beta_\tau\right) + \lambda\sum_{j=1}^{p}|\beta_j| \tag{8}$$

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

this paper, we add the penalty lasso to the model in equation (2). The new model becomes as follows:

$$\begin{aligned} y_i = {} & \beta_{(\tau)o} + \beta_{(\tau)1}x_1 + \beta_{(\tau)2}x_2 + \beta_{(\tau)3}x_3 + \beta_{(\tau)4}x_4 + \beta_{(\tau)5}x_5 + \beta_{(\tau)12}x_1x_2 \\ & + \beta_{(\tau)13}x_1x_3 + \beta_{(\tau)123}x_1x_2x_3 + \beta_{(\tau)124}x_1x_2x_4 + \cdots \\ & + \beta_{(\tau)1234}x_1x_2x_3x_4 + \beta_{(\tau)1345}x_1x_3x_4x_5 + \beta_{(\tau)1245}x_1x_2x_4x_5 \ldots \\ & + \beta_{(\tau)123456}x_1x_2x_3x_4x_5x_6 + u_\tau \\ & + \lambda\left[\left|\beta_{(\tau)o}\right| + \left|\beta_{(\tau)1}\right| + \left|\beta_{(\tau)2}\right| + \left|\beta_{(\tau)3}\right| + \left|\beta_{(\tau)4}\right| + \left|\beta_{(\tau)5}\right|\right. \\ & + \left|\beta_{(\tau)12}\right| + \left|\beta_{(\tau)13}\right| + \left|\beta_{(\tau)123}\right| + \left|\beta_{(\tau)124}\right| + \cdots \ldots . + \left|\beta_{(\tau)1234}\right| \\ & \left. + \left|\beta_{(\tau)1345}\right| + \left|\beta_{(\tau)1245}\right| + \cdots . + \left|\beta_{(\tau)123456}\right|\right] \qquad ,j \\ & = 1,2 \ldots 2^N - 1 \end{aligned} \tag{9}$$

For estimating the parameters in equation (9) we used the lasso method in quantile regression model with the following formula:

$$\underset{\beta_\tau}{min}\sum_{j=1}^{2^N-1}\sum_{i=1}^{n}\rho_\tau\left(y_i - x_{ij}^t\beta_{(\tau)j}\right) + \lambda\sum_{j=1}^{2^N-1}|\beta_j| \tag{10}$$

where, the equation (10) is not differentiable at 0.

Sample Study Description and Data Analysis:

Sample Study Description:

The data was collected from the Al-Shamia hospital in Iraq, with a sample size of 300 people of both genders (male and female). This study contains a set of variables and one of the response variables is represented by the blood pressure of each person. This variable is continuous. There are 63 covariates and the origin of these covariates is in the treatments for six factors representing the main factors. Each factor has two levels and two, three, four, five, six-factor interactions, classified according the following:

x_1 =A - Cholesterol (LDL) - this factor has two levels (high level, LDL>200mlg/dl and low level, LDL<200mlg/dl**)**

x_2 =B - Blood viscosity (PCV) - this factor has two levels (high level, PCV> 49 and low level, PCV<49)

x_3 =C - Heart rate (H.R) - this factor has two levels (high level, H.R>72 and low level, H.R<72)

x_4 =D - Age (A) - this factor has two levels (high level, A>45 and low level, A<45)

x_5 =E- weight (W) - this factor has two levels (high level, W>85 and low level, W<85)

x_6 =F= Diabetes (D) - this factor has two levels (high level, D>180 and low level of D<180)

x_1x_2 =AB= two-factor interactions between Cholesterol and Blood viscosity

x_1x_3 =AC= two-factor interactions between Cholesterol and Heart rate

x_1x_4 =AD= two-factor interactions between Cholesterol and Age

x_1x_5 =AE= two-factor interactions between Cholesterol and Weight

x_1x_6 =AF= two-factor interactions between Cholesterol and Diabetes

x_2x_3 =BC= two-factor interactions between Blood viscosity and Heart rate

x_2x_4 =BD= two-factor interactions between Blood viscosity and Age

x_2x_5 =BE= two-factor interactions between Blood viscosity and Weight

x_2x_6 =BF= two-factor interactions between Blood viscosity and Diabetes

x_3x_4 =CD= two-factor interactions between Heart rate and Age

x_3x_5 =CE= two-factor interactions between Heart rate and Weight

x_3x_6 =CF= two-factor interactions between Heart rate and Diabetes
x_4x_5 =DE= two-factor interactions between Age and Weight
x_4x_6 =DF= two-factor interactions between Age and Diabetes
x_5x_6 =EF= two-factor interactions between Weight.and Diabetes
$x_1x_2x_3$ =ABC= three-factor interactions between Cholesterol, Blood viscosity and Heart rate
$x_1x_2x_4$ =ABD= three-factor interactions between Cholesterol, Blood viscosity and Age
$x_1x_2x_5$ =ABE= three-factor interactions between Cholesterol, Blood viscosity and Weight
$x_1x_2x_6$ =ABF= three-factor interactions between Cholesterol, Blood viscosity and Diabetes
$x_1x_2x_6$ =ACD= three-factor interactions between Cholesterol, Heart rate and Age
$x_1x_2x_6$ =ACE= three-factor interactions between Cholesterol, Heart rate and Weight
$x_1x_3x_6$ =ACF= three-factor interactions between Cholesterol, Heart rate and Diabetes
$x_1x_4x_5$ =ADE= three-factor interactions between Cholesterol, Age and Weight
$x_1x_4x_6$ =ADF= three-factor interactions between Cholesterol, Age and Diabetes
$x_1x_5x_6$ =AEF= three-factor interactions between Cholesterol, Weight and Diabetes
$x_2x_3x_4$ =BCD= three-factor interactions between Blood viscosity, Heart rate and Age
$x_2x_3x_5$ =BCE= three-factor interactions between Blood viscosity, Heart rate and Weight
$x_2x_3x_6$ =BCF= three-factor interactions between Blood viscosity, Heart rate and Diabetes
$x_2x_4x_5$ =BDE= three-factor interactions between Blood viscosity, Age and Weight
$x_2x_4x_6$ =BDF= three-factor interactions between Blood viscosity, Age and Diabetes
$x_2x_5x_6$ =BEF= three-factor interactions between Blood viscosity, Weight and Diabetes
$x_3x_4x_5$ =CDE= three-factor interactions between Heart rate, Age and Weight
$x_3x_4x_6$ =CDF= three-factor interactions between Heart rate, Age and Diabetes
$x_1x_2x_6$ =CEF= three-factor interactions between Heart rate, weight and Diabetes
$x_4x_5x_6$ =DEF= three-factor interactions between Age, weight and Diabetes
$x_1x_2x_3x_4$ =ABCD= four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Age
$x_1x_2x_3x_5$ =ABCE= four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Weight
$x_1x_2x_3x_6$ =ABCF= four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Diabetes
$x_1x_2x_3x_6$ =ABDE= four-factor interactions between Cholesterol, Blood viscosity, Age and Weight
$x_1x_2x_4x_6$ =ABDF= four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Diabetes
$x_1x_2x_5x_6$ =ABEF= four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Diabetes
$x_1x_3x_4x_5$ =ACDE= four-factor interactions between Cholesterol, Heart rate, Age and Weight
$x_1x_3x_4x_6$ =ACDF= four-factor interactions between Cholesterol, Heart rate, Age and Diabetes.
$x_1x_3x_5x_6$ =ACEF= four-factor interactions between Cholesterol, Heart rate, weight and Diabetes
$x_1x_4x_5x_6$ =ADEf= four-factor interactions between Cholesterol, Age, Weight and Diabetes
$x_2x_3x_5x_6$ =BCEF= four-factor interactions between Blood viscosity, Heart rate, Weight and Diabetes
$x_2x_3x_4x_5$ =BCDE= four-factor interactions between Blood viscosity, Heart rate, Age and Weight
$x_2x_3x_4x_6$ =BCDF= four-factor interactions between Blood viscosity, Heart rate, Age and Diabetes
$x_2x_3x_5x_6$ =BCEF= four-factor interactions between Blood viscosity, Heart rate, Weight and Diabetes
$x_2x_4x_5x_6$ =BDEF= four-factor interactions between Blood viscosity, Age, Weight and Diabetes
$x_3x_4x_5x_6$ =CDEF= four-factor interactions between Heart rate, Age, Weight and Diabetes
$x_1x_2x_3x_4x_5$ =ABCDE= five-factor interactions between Cholesterol, Blood viscosity, Heart rate, Age and Weight
$x_1x_2x_3x_5x_6$ =ABCEF= five-factor interactions between Cholesterol, Blood viscosity, Heart rate, Weight and diabetes
$x_1x_3x_4x_5x_6$ =ACDEF= five-factor interactions between Cholesterol, Heart rate, Age, Weight and Diabetes
$x_1x_2x_4x_5x_6$ =ABDEF= five-factor interactions between Cholesterol, Heart rate, Age, Weight, and Diabetes
$x_1x_2x_3x_5x_6$ =ABCEF= five-factor interactions between Cholesterol, Blood viscosity, Heart rate, Age, Weight, and Diabetes
$x_2x_3x_4x_5x_6$ =BCDEF= five-factor interactions between Blood viscosity, Heart rate, Age, Weight and Diabetes
$x_1x_2x_3x_4x_5x_6$ =ABCDEF= six-factor interactions between Cholesterol, Blood viscosity, Heart rate, Age, Weight and Diabetes

The values of each covariate are calculated from the main factor and factor interactions at all levels. After completing the study data, we used lasso quantile regression model to analyze these data through four quantile levels under the quantile ratio (τ_1=0.20, τ_2=0.40, τ_3=0.60, τ_4=0.80). For data analysis, we are going to build the algorithm depending on the function (rq) in the lasso found within package (quantreg), as proposed by Koenker (2016).

Data Analysis:

The lasso factorial design of the quantile regression model computes each line separately and each factorial design of the quantile regression model differs based on the quantile ratio.

Table 1 shows an estimation of coefficients for the factorial design of the quantile regression.

Table 1: Coefficients estimation for the factorial design of the quantile regression

Coefficients				
Variable	$\tau_1 = 0.20$	$\tau_2 = 0.40$	$\tau_3 = 0.60$	$\tau_4 = 0.80$
Intercept	5.648	4.373	9.140	1.293
x_1 =A	-0.2327*	0.002	0.023	0.323
x_2 =B	0.003	-0.006	-1.729	-1.799
x_3 =C	-0.02109*	0.332*	5.713*	2.173*
x_4 =D	0.348	0.896	1.388*	4.963*
x_5 =E	-0.3893*	0.007*	-4.492*	-3.339*
x_6 =F	-0.0164*	-0.053*	-1.862*	-9.791*
x_1x_2 =AB	0.0038*	-0.002*	1.266*	4.305
x_1x_3 =AC	-0.4548	-0.070*	0.000	3.169
x_1x_4 =AD	-0.0472	0.000	-4.320	-2.056
x_1x_5 =AE	0.000	0.005	1.044	0.000
x_1x_6 =AF	1.019	-0.003	0.000	-1.265
x_2x_3 =BC	0.000	1.882*	9.881	1.682*
x_2x_4 =BD	0.0273*	0.026	0.000	0.000
x_2x_5 =BE	-0.3742	0.053*	2.424	-3.144
x_2x_6 =BF	1.791	-0.023	-7.629	-2.614*
x_3x_4 =CD	0.000	0.006	0.000	0.000
x_3x_5 =CE	0.893*	-0.076*	-4.146	1.708*
x_3x_6 =CF	0.094*	0.000	-1.156*	-3.861*
x_4x_5 =DE	0.000	1.940	1.750*	-1.839*
x_4x_6 =DF	0.499*	3.989*	-5.566*	3.971*
x_5x_6 =EF	0.000	0.013*	6.142*	1.772*
$x_1x_2x_3$ =ABC	-0.0282*	-0.048*	-5.762*	-3.650*
$x_1x_2x_4$ =ABD	0.000	0.006*	-9.430*	-3.008*
$x_1x_2x_5$ =ABE	-1.392*	0.093*	2.211*	-5.480*
$x_1x_2x_6$ =ABF	-0.041*	2.299*	-8.721*	1.130*
$x_1x_3x_4$ =ACD	1.684	-0.0007*	8.755	8.239
$x_1x_3x_5$ = ACE	0.000	0.000	1.361	0.000
$x_1x_3x_6$ = ACF	0.000	-5.683	1.409	-2.928*
$x_1x_4x_5$ = ADE	3.780*	0.027	2.360	4.029
$x_1x_4x_6$ = ADF	0.000	-0.006*	-1.004	1.163*
$x_1x_5x_6$ = AEF	-0.061*	-0.044	-4.753*	0.000
$x_2x_3x_4$ =BCD	-0.072	-8.755*	0.000	-1.596*
$x_2x_3x_5$ =BCE	0.0034	1.252	1.318	0.000
$x_2x_3x_6$ =BCF	-0.036	0.000	-2.928	3.443
$x_2x_5x_5$ =BDE	-0.064	1.316*	-5.365	7.447*
$x_2x_4x_6$ =BDF	-0.002*	0.005	1.064	-3.990
$x_2x_5x_6$ =BEF	-0.014	-1.653*	1.262	0.000
$x_3x_4x_5$ =CDE	-0.007*	1.026	-5.401*	5.107
$x_3x_4x_6$ =CDF	0.006*	1.812*	5.329*	-3.069*
$x_3x_4x_6$ =CEF	-1.475	-0.125*	1.971*	2.113*
$x_4x_5x_6$ =DEF	0.031*	-1.169*	2.220*	0.000
$x_1x_2x_3x_4$ =ABCD	0.000	0.000	0.000	-3.334*
$x_1x_2x_3x_5$ =ABCE	0.0029	-0.0082	0.000	0.000
$x_1x_2x_3x_6$ =ABCF	0.000	0.000	0.000	-7.068
$x_1x_2x_3x_6$ =ABDE	-1.976*	1.057*	6.193*	1.679
$x_1x_2x_4x_6$ =ABDF	1.865	4.157	1.375	1.510
$x_1x_2x_5x_6$ =ABEF	-0.003*	0.000	4.771*	0.000
$x_1x_3x_4x_5$ =ACDE	-1.590	4.062*	0.000	1.472
$x_1x_3x_5x_6$ =ACEF	-2.208	-1.413	-1.208	2.184*
$x_1x_4x_5x_6$ =ADEf	-0.0038	0.000	2.962	-9.869
$x_1x_3x_4x_6$ =BCDE	0.046*	0.000	5.229	-6.668
$x_2x_3x_4x_6$ =BCDF	-0.049	0.000	-5.287	0.000
$x_2x_3x_5x_6$ =BCEF	1.437	-1.748*	-2.628*	0.450
$x_2x_4x_5x_6$ =BDEF	-1.686	-7.187*	0.000	7.070
$x_3x_4x_5x_6$ =CDEF	-0.002	0.000	-5.981*	4.001*
$x_3x_4x_5x_6$ =ABEF	0.081	-0.002*	2.802*	5.342*
$x_1x_2x_3x_4x_5$ =ABCDE	0.056*	1.426*	1.564*	0.000
$x_1x_2x_3x_5x_6$ =ABCEF	-0.021*	-7.600*	0.000	5.816*
$x_1x_3x_4x_5x_6$ =ACDEF	0.000	-1.559*	4.519*	3.196*
$x_1x_2x_4x_5x_6$ =ABDEF	0.0051*	1.809	0.000	2.359

$x_1x_2x_3x_5x_6$ =ABCEF	3.760	-1.390	2.464	-1.191
$x_2x_3x_4x_5x_6$ =BCDEF	4.529*	0.0003*	-2.205*	0.210
$x_1x_2x_3x_4x_5x_6$ =ABCDEF	0.011*	-4.344*	-2.443*	3.671*
The pseudo-R squared	0.26453	0.44342	0.55876	0.74423

* Means the value of Pr(>|t|) < 0.05

At quantile ratio ($\tau_1 = 0.20$):

From the coefficients estimation, there are 12 covariates with unimportance in blood pressure. These covariates are: x_1x_5 =AE,x_2x_3 =BC, x_3x_4 =CD, x_4x_5 =DE, x_5x_6 =EF, $x_1x_2x_4$ =ABD, $x_1x_3x_5$ = ACE, $x_2x_3x_6$ =BCF, $x_1x_4x_6$ = ADF, $x_1x_2x_3x_4$ =ABCD, $x_1x_2x_3x_6$ =ABCF and$x_1x_3x_4x_5x_6$ =ACDEF.

Therefore, it is possible to remove these covariates from the factorial design of the quantile regression model at quantile ratio ($\tau_1 = 0.20$) because of the coefficients estimation of these covariates is equal exact zero. The rest of the covariates have effect in building the model. There are 26 covariates that have significant relationship in blood pressure. The rest of the covariates have no significant relationship with blood pressure, but have different effect on the response variable (blood pressure). At ratio (0.20), the covariates have weakness in explaining the variation in the response variable (blood pressure). This is clear from the value of the pseudo-R squared. The important covariates can explain 26.453% of the variation in blood pressure.

At quantile ratio ($\tau_1 = 0.40$):

From coefficients estimation, there are 11 covariates that have unimportance in blood pressure as the coefficients estimation is equal zero. These covariates are: x_1x_4 =AD, x_3x_6 =CF, $x_1x_3x_5$ = ACE, $x_2x_3x_6$ =BCF, $x_1x_2x_3x_4$ =ABCD, $x_1x_2x_3x_6$ =ABCF, $x_1x_2x_5x_6$ =ABEF, $x_3x_4x_5x_6$ =CDEF, $x_1x_3x_4x_6$ =BCDE, $x_2x_3x_4x_6$ =BCDF, $x_3x_4x_5x_6$ =CDEF.

It is also possible to remove these covariates from the factorial design of the quantile regression model at quantile ratio ($\tau_1 = 0.40$). because of the coefficients estimation of these covariates is equal exact zero. The rest of the covariates have effect in building the model. There are 22 covariates that have significant relationship with the blood pressure. The rest of the covariates do not have significant relationship with blood pressure, but have different effects on the response variable (blood pressure). At ratio (0.40), the covariates have weakness in explaining the variation in response variable (blood pressure). This is clear from the value of pseudo-R squared. The important covariates can explain 44.342% from the variation in blood pressure.

At quantile ratio ($\tau_1 = 0.60$):

From Table 1, there are 12 covariates that do not have importance in blood pressure, as coefficients estimation is equal zero. These covariates are: x_1x_3 =AC, x_1x_6 =AF, x_2x_4 =BD, $x_2x_3x_4$ =BCD, $x_1x_2x_3x_4$ =ABCD, $x_1x_2x_3x_5$ =ABCE, $x_1x_2x_3x_6$ =ABCF, $x_1x_3x_4x_5$ =ACDE, $x_2x_4x_5x_6$ =BDEF, $x_1x_2x_3x_5x_6$ =ABCEF, $x_1x_2x_4x_5x_6$ =ABDEF.

It is also possible to remove these covariates from the factorial design of the quantile regression model at quantile ratio ($\tau_1 = 0.60$) because of the coefficients estimation of these covariates is equal exact zero. The rest of the covariates have effect in building the model. There are 27 covariates that have significant relationship in blood pressure. The rest of the covariates have no significant relationship with blood pressure, but have different effects on the response variable (blood pressure). At ratio (0.60), the covariates have strength in explaining the variation in the response variable (blood pressure). This is clear from the value of pseudo-R squared). The important covariates can explain 55.876% from the variation in blood pressure.

At quantile ratio ($\tau_1 = 0.80$):

From Table 1, at quantile ratio (0.80), there are 12 covariates that do not have importance in blood pressure, as the coefficients estimation is equal zero. These covariates are: x_1x_5 =AE, x_2x_4 =BD, x_3x_4 =CD, $x_1x_3x_5$ =ACE, $x_1x_5x_6$ = AEF, $x_2x_3x_5$ =BCE, $x_2x_5x_6$ =BEF, $x_4x_5x_6$ =DEF, $x_1x_2x_3x_5$ =ABCE, $x_1x_2x_5x_6$ =ABEF, $x_1x_2x_3x_4x_5$ =ABCDE, $x_2x_3x_4x_5x_6$ =BCDEF. Therefore, these covariates can be removed from building the model of factorial design of the quantile regression model at quantile ratio ($\tau_1 = 0.80$) because of the coefficients estimation of these covariates is equal exact zero. The rest of the covariates have effect on blood pressure. These covariates can explain 74.423% from the variation in blood pressure. This indicates a strong model at quantile ratio (0.80). Therefore, we focus on describing the covariates at this quantile ratio as following:

a - covariates (main effects): C, D, E, F have significant relationship with blood pressure, but the effects of A, B do not have a significant relationship to blood pressure

b - covariates (two-factors interactions): BC, BF, CE, CF, DE, DF, EF have significant relationship with blood pressure

c - covariates (three-factors interactions): ABC, ABD, ABE, ABF, ACF, ADF, BCD, BDE, CDF, CEF have significant relationship with blood pressure

d - covariates (four-factors interactions): ABCD, ACEF, CDEF, ABEF have significant relationship with blood pressure; the rest of the covariates of four-factors interactions do not have a significant relationship with blood pressure

e - covariates (five-factors interactions): ABCEF, ACDEF have significant relationship with blood pressure; the rest of the covariates of five-factors interactions no not have a significant relationship with blood pressure

f - covariate (six-factors interactions): ABCDEF has significant relationship with blood pressure.

The procedure of variable selection is clarified in Figures 1 and 2 through for quantile ratios (0.20, 0.40, 0.60, 0.80).

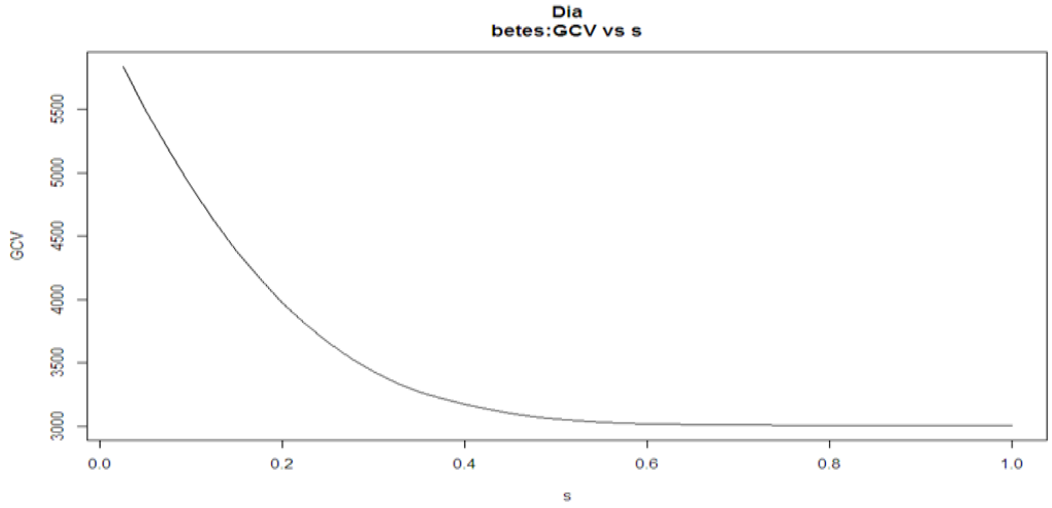

Fig. 1: GCV Score as a function of relative bound (s) for blood pressure; data set via different quantile ratio

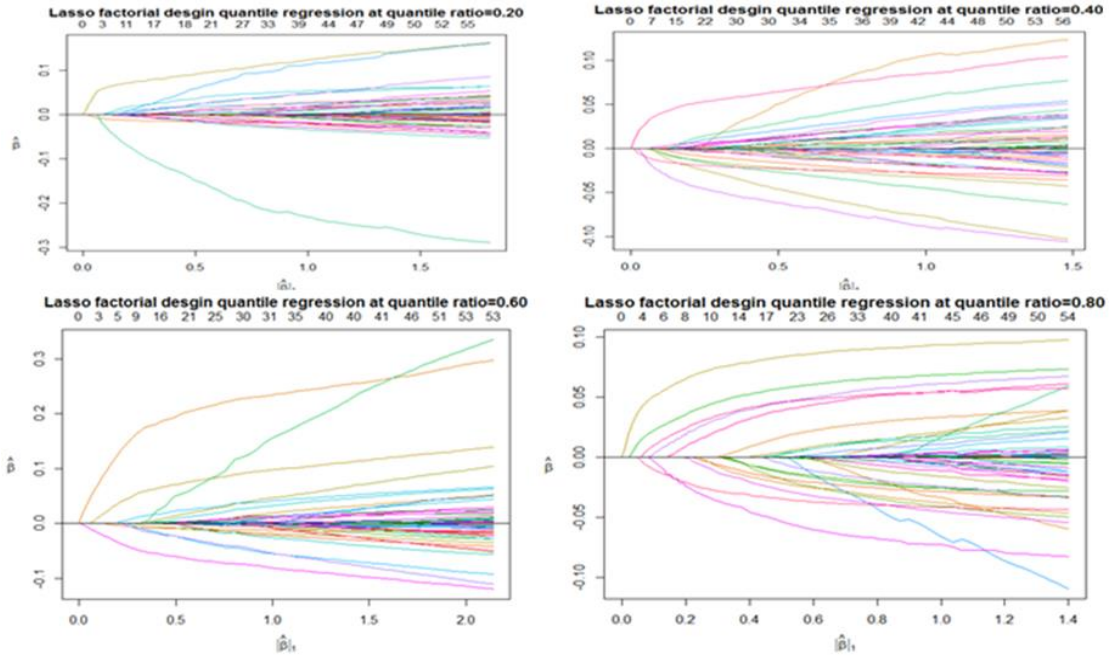

Fig. 2: LASSO coefficient in blood pressure

From Figure 1 we can see that the value of GCV will decreases until 0.4. That means that the value of *s*=0.4. If covariates path exceeds 0.4, this means that the coefficients will be nonzero (important covariates). If the independent variables path is less than 0.4, this means that the coefficients will be zero (unimportant covariates). The value of *s* is very important for distinguishing between zero and nonzero coefficients. Figure 2. illustrates all the independent variables that take specific paths as a function of relative bound (*s*).

Conclusion:

Lasso Factorial Design Quantile Regression quantile ratios (0.20,0.40) cannot represent data of blood pressure under study, as the values of the pseudo-R squared (0.26453, 0.44342) are small. This means that all covariates in models can explain (26.45%,44.34%) respectively in variation of blood pressure. This indicator

shows weakness in the representation of the data of blood pressure. The Lasso Factorial Design Quantile Regression quantile ratios (0.60,0.80) have strength in representing the data of blood pressure as the values of the pseudo-R squared (0.55876, 0.74423) are high. This means that all covariates in models can explain (55.87%,74.44%) respectively in variation of the blood pressure. This indicator shows strength in representation of the data of blood pressure. In this paper, we focused on Lasso Factorial Design Quantile Regression quantile ratio (0.80), as it is stronger in representing data of blood pressure under study. Also, there are 12 covariates at quantile ratio 0.20 that are unimportant in explaining the variation in blood pressure. Because of these covariates its coefficients estimation is exact zero. This means that 12 covariates in the philosophy of the lasso approach are unimportant in building Lasso Factorial Design Quantile Regression Model at quantile ratio (0.20). From Table (1), there are 26 covariates that have significant effect on blood pressure at quantile ratio (0.20). From Table (1) there are 11 covariates unimportant in explaining the variation in blood pressure. Because of these covariates its coefficients estimation is exact zero, and also there are 22 covariates that have significant effect in blood pressure at quantile ratio 0.40. From Table (1), there are 12 covariates unimportant in blood pressure in explaining the variation in blood pressure because of these covariates its coefficients estimation is exact zero, and also there are 27 covariates that have significant effect in blood pressure at quantile ratio 0.60. From Table (1), there are 12 covariates unimportant in explaining the variation in blood pressure because of these covariates its coefficients estimation is exact zero, as shown below:

a- x_1x_5 =AE - two-factor interactions between Cholesterol and Weight
b- x_2x_4 =BD - two-factor interactions between Blood viscosity and Age
c- x_3x_4 =CD - two-factor interactions between Heart rate and Age
d- $x_1x_3x_5$ =ACE - three-factor interactions between Blood viscosity, Heart rate and weight
e- $x_1x_5x_6$ =AEF - three-factor interactions between Cholesterol, Weight and Diabetes
f- $x_2x_3x_5$ =BCE - three-factor interactions between Blood viscosity, Heart rate and Weight
g- $x_2x_5x_6$ =BEF - three-factor interactions between Blood viscosity, Weight and Diabetes
h- $x_4x_5x_6$ =DEF - three-factor interactions between Age, Weight and Diabetes
i- $x_1x_2x_3x_5$ =ABCE - four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Weight
j- $x_1x_2x_5x_6$ =ABEF - four-factor interactions between Cholesterol, Blood viscosity, Heart rate and Diabetes
k- $x_1x_2x_3x_4x_5$ =ABCDE - five-factor interactions between Cholesterol, Blood viscosity, Heart rate, Age, and Weight
l- $x_2x_3x_4x_5x_6$ =BCDEF - five-factor interactions between Blood viscosity, Heart rate, Age, Weight, and Diabetes

The above covariates will be removed from building the Lasso Factorial Design Quantile Regression quantile ratios at quantile ratio (0.80) because these variables do not have importance in building this model. This procedure achieved good explanation for the of the rest important covariates. Also, there are 27 covariates that have significant effect in blood pressure.

Future Work:

There are many advantages in using new regularization Factorial Design Quantile Regression Model such as adaptive lasso, elastic net and fused methods. By studying the regularization Factorial Design Tobit Quantile Regression Model, when response variable is censored at zero point from left side, further contributions can be brought in restructuring of data through transfer treatment in experimental design to covariates in regression models in order to cover entire relationship between response variable and all covariates, and exclude unimportant covariates from these models by using one regularization methods

REFERENCES

Dossou-Gbété, S., W. Tinsson, 2005. Factorial experimental designs and generalized linear models, 29(2): 249-268.

Fernando, L., R. Ospina, L. Alberto, 2012. On estimated means for 2^{k-p} experiments with beta response. Journal of Statistics: Advances in Theory and Applications, 8(2): 105-130.

FISHER, R.A., 1926. "The arrangement of field Experiments ,"*J . Ministry Agric*, 33: 503-513.

Guo, H., A. Mettas, 2007. Improved reliability using accelerated degradation & design of experiments. I Proceedings Annual Reliability and Maintainability Symposium.

Kirk, R.E., 1995. Experimental design, Procedures for the behavioral sciences, 3rd ed., Pacific Grove, CA: Brooks/Cole.

Koenker, R., V. D'Orey, 1987. Algorithm AS 229: Computing regression quantiles. Journal of the Royal Statistical Society: Series C (Applied Statistics), 36: 383-393.

Koenker, R., J.A. Machado, 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American statistical association, 94(448): 1296-1310.

Koenker, R., 2005. Quantile regression, 38. Cambridge University Press.

Koenker, R., 2016. quantreg: Quantile regression. R package version 5.05.

Searle, Shayle, R., 1971. Linear Models. New York: John Wylie & Sons.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 2: 267-288.

Toutenburg Shalabh, H., 2009. Statistical Analysis of Designed Experiments. Third Edition. Springer Texts in Statistics.

Wu, C.F.J., M. Hamada, 2009. Experiments: Planning, Analysis and Parameter Design Optimization, 2nd Edition. Wiley, New York.

Wu, C.J., M.S. Hamada, 2011. Experiments: planning, analysis, and optimization, 552. John Wiley & Sons.

Yates , F. (1937) The design and Analysis of Factorial experiments. Technical Commonwealth, 35: 1-95. Bureau of soils, Harpenden.