

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
© 2025 Stats & Maths
Maths 2025; SP-10(8): 43-52
www.mathsjournal.com
Received: 12-07-2025
Accepted: 18-08-2025

Tahani Jaber Nasser
Department of Statistics, College
of Administration and
Economics, University of Al-
Qadisiyah, AL Diwaniyah, Iraq

Bahr Kadhim Mohammed
Department of Statistics, College
of Administration and
Economics, University of Al-
Qadisiyah, AL Diwaniyah, Iraq

Corresponding Author:
Tahani Jaber Nasser
Department of Statistics, College
of Administration and
Economics, University of Al-
Qadisiyah, AL Diwaniyah, Iraq

Estimate factorial experimental design by using some methods with an application

Tahani Jaber Nasser and Bahr Kadhim Mohammed

DOI: <https://www.doi.org/10.22271/math.2025.v10.i8Sa.2149>

Abstract

Factorial experimental design serves as a fundamental statistical framework for evaluating the effects of multiple factors, their levels, and the interactions between them. When standard assumptions are met, traditional analysis of variance (ANOVA) is typically used. However, when these assumptions are violated—such as in the presence of non-normality or heteroscedasticity—alternative estimation methods become necessary. This study examines the comparative efficiency of various estimation techniques, including Ordinary Least Squares (OLS), Generalized Linear Models (GLM), Quantile Regression, the Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive LASSO, and the MM-estimator. The research focuses on assessing the robustness of each method to outliers, the accuracy of estimating main and interaction effects, and the flexibility of handling complex factorial designs (specifically 2^3 structures), under challenging data conditions often encountered in agricultural experiments. Additionally, the study explores the role of data transformation techniques, model selection criteria, and diagnostic tools in improving the reliability of statistical analysis. Comparative evaluations are conducted across varying sample sizes, effect sizes, and error distributions to reflect a wide range of experimental scenarios. The overarching goal is to provide agricultural researchers with practical, evidence-based guidance for selecting appropriate statistical methodologies in factorial experiments, thereby enhancing the validity and interpretability of their results.

Keywords: Factorial design, ANOVA, quantile regression, LASSO, adaptive LASSO, MM-estimator, generalized linear model, agricultural experiments

1. Introduction

Experimental design plays a central role in applied statistics, providing a structured framework for planning experiments, controlling variability, and drawing reliable inferences. Among these designs, factorial experiments—originally introduced by Fisher (1926) ^[13] and further advanced by Yates (1937) enable the simultaneous investigation of multiple factors and their potential interactions. This multifactorial approach enhances efficiency by reducing the number of required experimental runs, thereby saving time, cost, and resources. In agricultural research, factorial designs are particularly valuable for exploring how various controllable factors—such as fertilizer types, irrigation levels, and environmental conditions—influence crop productivity and sustainability. However, traditional analytical methods like Analysis of Variance (ANOVA) often rely on assumptions such as normality and homoscedasticity, which may not hold in real-world data. When these assumptions are violated, the validity and reliability of the conclusions drawn from such methods are compromised. To address these limitations, this study evaluates the performance of six statistical approaches for analyzing data from factorial experiments in the presence of non-normal and heterogeneous errors: Ordinary Least Squares (OLS), Generalized Linear Models (GLM), Quantile Regression, the Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive LASSO, and the MM-estimator. The aim is to identify the most robust and accurate estimation technique under these challenging data conditions, thereby providing practical guidance for researchers conducting factorial experiments in applied fields such as agriculture.

1.1 Experiment Design

An experimental design identifies the independent and dependent variables and includes the method by which the statistical and randomization aspects of the experiment will be carried out. Its primary goal is to establish a causal relationship between the independent and dependent variables, while the secondary goal is to extract the maximum amount of information with minimal resource expenditure. (Kirk, 1995) ^[2]. Experiment design refers to the plan used to carry out the experiment. It involves a series of steps that include the allocation of treatments to experimental units for the purpose of collecting data. This process requires a comprehensive understanding of how to analyze the data after it has been gathered from the experiment. It is essential for the design to allow for the estimation of experimental error, as the inability to estimate this error renders the design invalid. There are various types of experimental designs, each tailored to a specific objective. The researcher's goal and the nature of the study determine the appropriate design. A well-chosen design ensures the ability to overcome challenges that may arise during statistical analysis. Thus, the optimal design for an experiment will vary based on the factors being studied.

1.2 Factorial Experimental Design

Factorial experiments are used to study the effect of two or more factors on a particular phenomenon or specific property. Instead of conducting a simple experiment for each factor separately, which requires a lot of effort, time, cost, and experimental materials, a single experiment that includes all factors at once can be performed, providing many advantages. These advantages include eliminating the complexities and problems of repeating the experiment for each factor individually, as well as reducing costs, effort, time, and materials used. Additionally, this method does not require testing the independence of factors, as the experimental results provide comprehensive information about the effect of each factor and its interactions with other factors. This experiment, which involves multiple factors, is known as a "factorial experiment" or "multifactorial experiment" (Factorial Experiment) (Kamal Alwan Al-Mashhadani, 2010) ^[1]. Factorial experiments can be conducted in a wide range of experimental designs, which are organized in a manner suitable for simple designs and are statistically analyzed. The factorial experiment provides valuable information about the interaction effects between factors, which are of great importance in interpreting the data.

1.3 Methods of estimating factorial experiments

1.3.1 Ordinary Least Squares (OLS)

It is considered one of the statistical methods used to estimate parameters in linear models by minimizing the sum of squared errors between observed and predicted values. The sample treatment mean and its observed value respectively represent the "least squares estimator" and "least squares estimate" of the treatment mean (Dean & Voss, 1999) ^[12]. In one-way analysis of variance models, least squares estimators are derived by optimizing the following sum of squared errors function:

$$\sum_{i=1}^v \sum_{t=1}^r e_{it}^2 = \sum_{i=1}^v \sum_{t=1}^{r_i} (y_{it} - \mu - T_i)^2 \quad (1)$$

Where,

e_{it} : Error term for treatment e in replication t . v : Number of treatments. r_i : Number of replications for treatment i . y_{it} : Observed value for treatment i in replication t . μ : Overall mean. T_i : Treatment effect for i . The estimated model is given by:

$$\hat{y}_i = \hat{\mu} + \hat{T}_i \quad (2)$$

Where,

$\hat{y}_i$: Predicted value for treatment i . $\hat{\mu}$: Least squares estimate of the overall mean.

$\hat{T}_i$: Least squares estimate of the treatment effect i . The calculation of least squares estimators is done by differentiating the sum of squared errors with respect to each parameter and setting the derivatives to zero, leading to the following system of normal equations:

$$y_{..} - n\hat{\mu} - \sum_i r_i \hat{T}_i = 0, \quad (3)$$

$$y_{i.} - r_i \hat{\mu} - r_i \hat{T}_i = 0, \quad (4)$$

$i=1, \dots, v,$

Where,

$y_{..}$: Grand total of all observations. $y_{i.}$: Total of observations for treatment i . n : Total number of observations.

1.3.2 Generalized Linear Models (GLM)

Are a method for modeling the relationship between a dependent variable and several independent variables. This type of model is characterized by the dependent variable following one of the exponential family distributions, making generalized linear models less restrictive than traditional regression models. These models provide a solution for modeling data that does not follow a

normal distribution. Generalized linear models were first introduced by Neider and Wedderburn in 1972 in an attempt to expand the assumptions of traditional regression models to fit practical applications (Barnett & Dobson, 2008) ^[3]. Generalized linear models are less restrictive for several reasons, including [Evelien, 2014]:

- They do not require constant variance, allowing for different variances.
- The relationship between the dependent and independent variables does not have to be linear; rather, there must be a linear relationship between the link function and the set of independent variables, which allows for the application of nonlinear models using generalized linear models.
- Model parameters are estimated using Maximum Likelihood Estimation (MLE) or traditional Ordinary Least Squares (OLS).
- In these models, the expected value of the response variable is replaced by the link function, making the error variance more stable. The link function is a linear combination of the explanatory variables. Components of the GLM Model.

1.3.3 Quantile Regression (QR)

For over two centuries, regression techniques in statistical modeling have been developed to determine the relationship between a dependent variable (outcome) and independent variables (covariates). Classical regression has been one of the most commonly used statistical methods for estimating effects at the mean, assuming that regression coefficients or covariate effects remain constant across the population. However, these average effects are not always of interest in many fields and may often be highly heterogeneous. Quantile Regression (QR), introduced by Koenker and Bassett in 1978, has gained increasing attention and has been widely used in exploring relationships such as those between foreign direct investment and economic growth, as well as in precision health and personalized medicine. QR is capable of capturing heterogeneous effects through conditional quantiles of the outcome variable, offering a more comprehensive understanding of its distribution across different levels.

In cases of heteroscedasticity or data with asymmetries and heavy tails, the sample median (the 50th percentile) often provides a more accurate measure of central tendency than the mean. Therefore, QR is more flexible and robust compared to classical mean regression models, as it does not require strict distributional assumptions. While traditional regression methods still dominate statistical modeling, QR serves as a valuable extension and complement, particularly when classical assumptions are violated. QR has attracted considerable research interest and has been widely applied to independent data and time-to-event (survival) data. More recently, its applications have expanded to include longitudinal data analysis (Huang et al., 2023) ^[9].

In the context of improving product reliability, recent studies have shown that QR is an effective tool for analyzing lifetime data, particularly when dealing with issues such as heteroscedasticity or censored data. Unlike traditional methods such as Maximum Likelihood Estimation (MLE), which require strict distributional assumptions, QR offers a more flexible and robust approach to estimating quantiles without relying on such assumptions. To further enhance the performance of QR, the Weighted Quantile Regression (WQR) model has been proposed, introducing an adaptive weight parameter to improve stability and accuracy. Both theoretical and practical results have shown that WQR outperforms traditional QR, especially in estimating lower quantiles (e.g., 0.1), which are particularly crucial in reliability analysis (Lu & Zhao, 2023) ^[8].

1.3.4 The Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) is a linear model estimation method proposed by Tibshirani (1996) ^[14]. This method shares with ridge regression the property of minimizing the sum of squared errors while imposing a constraint on the estimated coefficients. More specifically, while ridge regression uses a quadratic constraint, Lasso applies an upper bound on the sum of absolute values of the coefficients. Tibshirani presents two main advantages of Lasso compared to ordinary least squares (OLS) regression:

First: OLS models that are over specified often exhibit high variance despite low bias, which reduces prediction accuracy. Lasso addresses this issue by shrinking some coefficients or setting them to zero (removing them from the model), enabling a trade-off between bias and variance.

Second: OLS models typically contain many small coefficients that add little value to the model and increase its complexity without providing significant explanatory power. Lasso shrinks some coefficients and sets others to zero, automatically performing variable selection by including only non-zero effects in the final model.

Let (x_i, y_i) , $i=1,2,\dots,N$ represent a traditional regression dataset with $x_i=(x_{i1},\dots,x_{ip})$ as independent variables and y_i as the response variable. Assuming standardized independent variables:

$$\sum_i \frac{x_{ij}}{N} = 0 \text{ and } \sum_i \frac{x_{ij}^2}{N} = 1 \quad (5)$$

And assuming centered y_i where $\bar{y}=0$, The Lasso estimator $\hat{\beta}$ is defined as the set $\hat{\beta}=(\hat{\beta}_1,\dots,\hat{\beta}_p)$ that minimizes:

$$\min \beta \{ \sum_{i=1}^N (y_i - \sum_{j=1}^p \beta_j x_{ij})^2 \}$$

Subject to

$$\sum_{j=1}^p |\beta_j| \leq t \quad (6)$$

Where $t \geq 0$ is the tuning parameter that controls the amount of shrinkage applied to the coefficient estimates. If $\hat{\beta}^{\text{ols}}$ are the full OLS estimates and $t_0 = \sum |\hat{\beta}_i^{\text{ols}}|$, then values of $t < t_0$ lead to shrinking coefficients toward zero, effectively excluding variables whose coefficients become zero from the model.

Tibshirani recommends three different methods for estimating the bound t : K-fold cross-validation, generalized cross-validation, and a method based on Stein's unbiased risk estimator. K-fold cross-validation is used to estimate t .

Suppose

$$Y = \eta(x) + \varepsilon \quad (7)$$

Where $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2$ where ME is the mean squared error. To estimate $\hat{\eta}(x)$ as

$$\text{ME} = E[(\hat{\eta}(X) - \eta(X))^2], \quad (8)$$

And the prediction error PE as follows:

$$\text{PE} = E[(Y - \hat{\eta}(X))^2] = \text{ME} + \sigma^2 \quad (9)$$

This prediction error is then estimated using K-fold cross-validation. If

$$s = \frac{t}{\sum_{j=1}^p |\hat{\beta}_j^0|} \quad (10)$$

Leave-one-out cross-validation (where $K=n$) is used to determine the optimal value of t . The Lasso problem represents a quadratic programming problem that can be solved using numerical analysis algorithms (Efron et al., 2004, p. 37). However, the modified least angle regression (LARS) method can also be used, which provides the complete solution path for all values of t (from 0 to ∞).

1.3.5 Adaptive lasso

Adaptive LASSO is considered an advanced statistical technique used to enhance variable selection and parameter estimation in statistical models, particularly in cases involving a large number of explanatory variables, such as in factorial designs. This method was first introduced by Zou (2006), who developed an improved version of the traditional LASSO by incorporating adaptive weights calculated from initial estimates of the parameters, allowing for more precise and selective shrinkage. This method is distinguished by its ability to handle multicollinearity and improve computational efficiency, making it suitable for complex analytical scenarios (Ren & Zhang, 2013) ^[5]. The strength of Adaptive LASSO lies in its capability to simultaneously perform estimation and variable selection by integrating adaptive weights into the penalty function, which results in shrinking the coefficients of irrelevant variables to zero. This improves both model interpretability and predictive accuracy (Qian & Yang, 2013) ^[6]. The quantile regression model at a specific quantile level $\tau \in (0,1)$ is represented by the following equation:

$$Q_y(\tau|X) = X\beta_{(\tau)} = x_i^T \beta_{(\tau)} \quad (11)$$

Where,

$Q_y(\tau|X)$: is the conditional quantile function of the response variable y ,

x_i : is the i -th row vector of the design matrix X ,

$\beta_{(\tau)}$: is the parameter vector at quantile τ .

the linear regression model can be represented as follows

$$\begin{aligned} y_i = & \beta_{(\tau)0} + \beta_{(\tau)1}x_1 + \beta_{(\tau)2}x_2 + \beta_{(\tau)3}x_3 + \beta_{(\tau)4}x_4 + \beta_{(\tau)12}x_1x_2 + \beta_{(\tau)13}x_1x_3 + \\ & \beta_{(\tau)14}x_1x_4 + \beta_{(\tau)23}x_2x_3 + \beta_{(\tau)24}x_2x_4 + \beta_{(\tau)34}x_3x_4 + \beta_{(\tau)123}x_1x_2x_3 + \\ & \beta_{(\tau)124}x_1x_2x_4 + \beta_{(\tau)134}x_1x_3x_4 + \beta_{(\tau)234}x_2x_3x_4 + \beta_{(\tau)1234}x_1x_2x_3x_4 + \mu_{(\tau)} \end{aligned} \quad (12)$$

To estimate the regression coefficients while performing variable selection, the Adaptive LASSO method solves the following penalized optimization problem:

$$\hat{\beta}_{\text{AdaLASSO}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \right\} \quad (13)$$

Where

$\hat{\beta}_{\text{AdaLASSO}}$: The estimated vector of regression coefficients obtained using the Adaptive LASSO method. This estimator aims to reduce model complexity and perform variable selection by shrinking some coefficients toward zero.

$\underset{\beta}{\operatorname{argmin}}$: Indicates the values of the coefficient vector β that minimize the objective function enclosed within the brackets. y_i

: Represents the observed value of the response variable for the i^{th} observation. X_i : A row vector of explanatory variables

(predictors) corresponding to the i^{th} observation. $X_i^T \beta$: The predicted value $\hat{y}_i$ for the i^{th} observation, computed as the inner

product of X_i and β . $\sum_{i=1}^n (y_i - X_i^T \beta)^2$: The residual sum of squares, measuring the discrepancy between the observed and

predicted values. λ : The regularization parameter that controls the strength of the penalty. A larger λ leads to greater shrinkage of the coefficients. $\sum_{j=1}^p w_j |\beta_j|$: The penalty term that adds adaptive shrinkage to the coefficients, where each β_j is penalized based

on its corresponding weight w_j . p : The number of explanatory variables included in the model. w_j : The adaptive weight for the coefficient. It is typically defined as:

$$w_j = \frac{1}{|\hat{\beta}_j^{\text{init}}|^\gamma} \quad (14)$$

$\hat{\beta}_j^{\text{init}}$: The initial estimate of the coefficient β_j , usually obtained from an Ordinary Least Squares (OLS) or Ridge regression fit.

These estimates are used to compute the adaptive weights.

γ : A positive tuning parameter (commonly set to 1) that controls the influence of the initial estimates on the weights w_j . It is

often selected by the user or through cross-validation. The following equation represents the Lasso-penalized quantile regression model (Quantile Regression with Lasso Penalty), which combines the advantages of quantile regression with the regularization capability for selecting the most important variables:

$$\hat{\beta}_{(\tau)} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \rho_\tau(y_i - x_i^T \beta) + \lambda \sum_{j=1}^p w_j |\beta_j| \right\} \quad (15)$$

The quantile loss function (ρ_τ) used in the quantile regression model is defined by the following relation:

$$\rho_\tau(\mu) = \begin{cases} \tau\mu, & \mu \geq 0 \\ (\tau - 1)\mu, & \mu < 0 \end{cases} \quad (16)$$

1.3.6 Quantile Factorial Design

Quantile regression models are increasingly employed in factorial designs, particularly when three factors are involved, to overcome the limitations of traditional models-especially in the presence of heteroscedasticity and outliers. Quantile regression estimates conditional quantiles of the response, allowing researchers to understand response behavior across the distribution's tails and gain insights into the effects of different factors at various levels of the response.

The general form of the quantile regression model is as follows:

$$Q_Y(\tau|A, B, C) = \beta_0(\tau) + \beta_1(\tau)A + \beta_2(\tau)B + \beta_3(\tau)C + \beta_4(\tau)AB + \beta_5(\tau)AC + \beta_6(\tau)BC + \beta_7(\tau)ABC \quad (17)$$

$Q_Y(\tau|A, B, C)$: The conditional quantile of the response variable Y at quantile level τ , given factors A, B, C .

τ : The quantile of interest (e.g., 0.25, 0.50, 0.75).

$\beta_0(\tau)$: Intercept term; the baseline value of the τ -th quantile when all factors are at their reference level (e.g., zero).

$\beta_1(\tau)$: Main effect of factor A at quantile τ .

$\beta_2(\tau)$: Main effect of factor B at quantile τ .

$\beta_3(\tau)$: Main effect of factor C at quantile τ .

$\beta_4(\tau)$: Interaction effect between A and B at quantile τ .

$\beta_5(\tau)$: Interaction effect between A and C at quantile τ .

$\beta_6(\tau)$: Interaction effect between B and C at quantile τ .

$\beta_7(\tau)$: Three-way interaction effect between A, B and at quantile τ .

A,B,C : Experimental factors, each coded at two levels (typically -1 for low and +1 for high).

One advantage of using quantile regression in factorial experiments is that it is less sensitive to outliers compared to traditional regression (Xu, 2023) ^[10]. It allows for the analysis of treatment effects across multiple response levels and does not require the assumption of a specific distribution for the response variable (Silva, 2022) ^[11]. Despite the significant advantages offered by quantile regression, it may introduce complexity in model interpretation and requires careful selection of appropriate quantiles. This complexity can pose a challenge for some researchers, particularly those accustomed to using conventional regression methods.

1.3.7 MM Estimator

The MM Estimator method was introduced by Yohai (1987) as one of the most robust estimation techniques in regression analysis, particularly in the presence of outliers or violations of the normality assumption of model residuals. This method combines the high breakdown point of S-estimation with the efficiency of M-estimation, achieving a balance between statistical robustness and efficiency. The MM Estimator can reach a breakdown point of up to 50%, allowing it to resist the influence of a large proportion of contaminated data while maintaining efficiency close to that of Ordinary Least Squares (OLS) under normally distributed errors (Maronna, Martin & Yohai, 2006) ^[16].

The MM method consists of three sequential stages. First, regression coefficients are estimated using the S-estimator, defined as:

$$\hat{\beta}_s = \operatorname{argmin}_{\beta} \hat{\sigma} \text{ such that } \frac{1}{n} \sum_{i=1}^n \rho \left(\frac{y_i - x_i^T \beta}{\hat{\sigma}} \right) = b \quad (18)$$

Next, this estimate is used as a starting point to calculate an initial estimate using the M-estimator:

$$\hat{\beta}_0 = \operatorname{argmin}_{\beta} \sum_{i=1}^n \rho_0 \left(\frac{y_i - x_i^T \beta}{\hat{\sigma}} \right) \quad (19)$$

Finally, the M-estimator is applied again with a loss function chosen to balance efficiency and robustness, resulting in the final MM estimate:

$$\hat{\beta}_{MM} = \operatorname{argmin}_{\beta} \sum_{i=1}^n \rho_1 \left(\frac{y_i - x_i^T \beta}{\hat{\sigma}} \right) \quad (20)$$

Where,

y_i : response variable.

x_i : vector of predictors.

β : regression coefficients to be estimated.

$\hat{\sigma}$: robust scale estimator.

ρ, ρ_0, ρ_1 : loss functions (e.g., Huber or bisquare), selected to balance robustness and efficiency.

b : constant usually chosen to adjust the breakdown point.

n : sample size.

The final estimate $\hat{\beta}_{MM}$ is the regression estimate that combines robustness against outliers with high efficiency, making this method highly suitable for practical applications involving contaminated or anomalous data.

Simulation Example

We simulated data from a full factorial experimental design with 2^3 factors (A, B, and C), each taking two levels (−1 for low and +1 for high). The true model included all main effects, two-way interactions, and the three-way interaction:

$$Q_Y(\tau|A, B, C) = \beta_0(\tau) + \beta_1(\tau)A + \beta_2(\tau)B + \beta_3(\tau)C + \beta_4(\tau)AB + \beta_5(\tau)AC + \beta_6(\tau)BC + \beta_7(\tau)ABC + \varepsilon \quad (21)$$

Where the β coefficients were fixed at known values and the error term ε followed a non-normal, heavy-tailed distribution to mimic real-world deviations from Gaussian noise. The experiment was simulated under three sample sizes (N=50, 100 and 200) by replicating the factorial runs proportionally.

For each dataset, six estimation methods were applied:

1. Ordinary Least Squares (OLS)
2. Generalized Linear Model (GLM, Gaussian family)
3. Quantile Regression
4. Least Absolute Shrinkage and Selection Operator (LASSO)
5. Adaptive LASSO
6. MM Estimator (robust regression)

Model performance was assessed using mean squared error (MSE), coefficient bias, and coefficient of determination (R^2). Each scenario was replicated 500 times for the

The simulation results show that Quantile Regression consistently outperformed other methods in terms of bias and explained variance (R^2) across all sample sizes, demonstrating its robustness to the non-normal error structure. While OLS showed higher mean squared error (MSE) and lower R^2 , reflecting sensitivity to skewed noise, the Generalized Linear Model (GLM) and MM Estimator provided more stable and accurate estimates.

Regularization methods behaved differently: Adaptive LASSO improved performance over standard LASSO, particularly at the largest sample size (N=200), where it achieved the lowest MSE and competitive R^2 . The MM estimator, with its robustness to outliers and heavy-tailed errors, offered consistent performance across all sample sizes, balancing bias and variance effectively.

Table 1: The MSE, Bias and R^2 for the methods of study when, $N=50$

OLS	1.332	0.052	0.655
GLM	1.003	0.028	0.836
Quant. Reg	0.963	0.025	0.876
LASSO	0.990	0.029	0.821
Adapt. LASSO	0.985	0.026	0.844
MM Est.	0.992	0.026	0.842

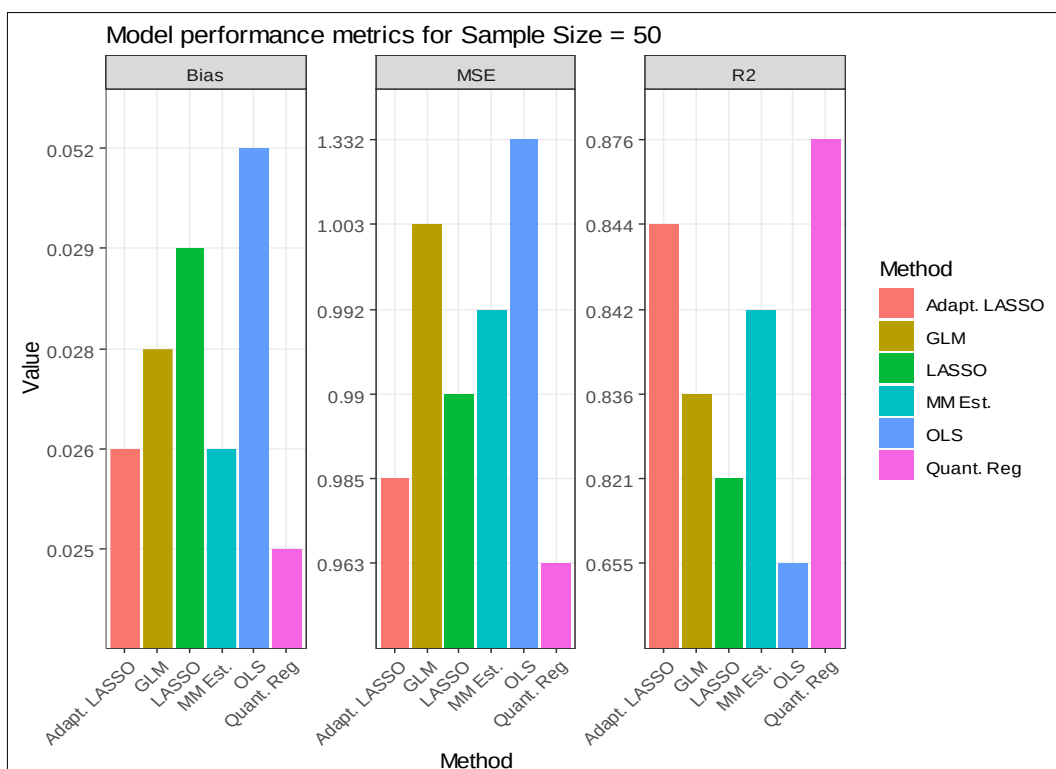
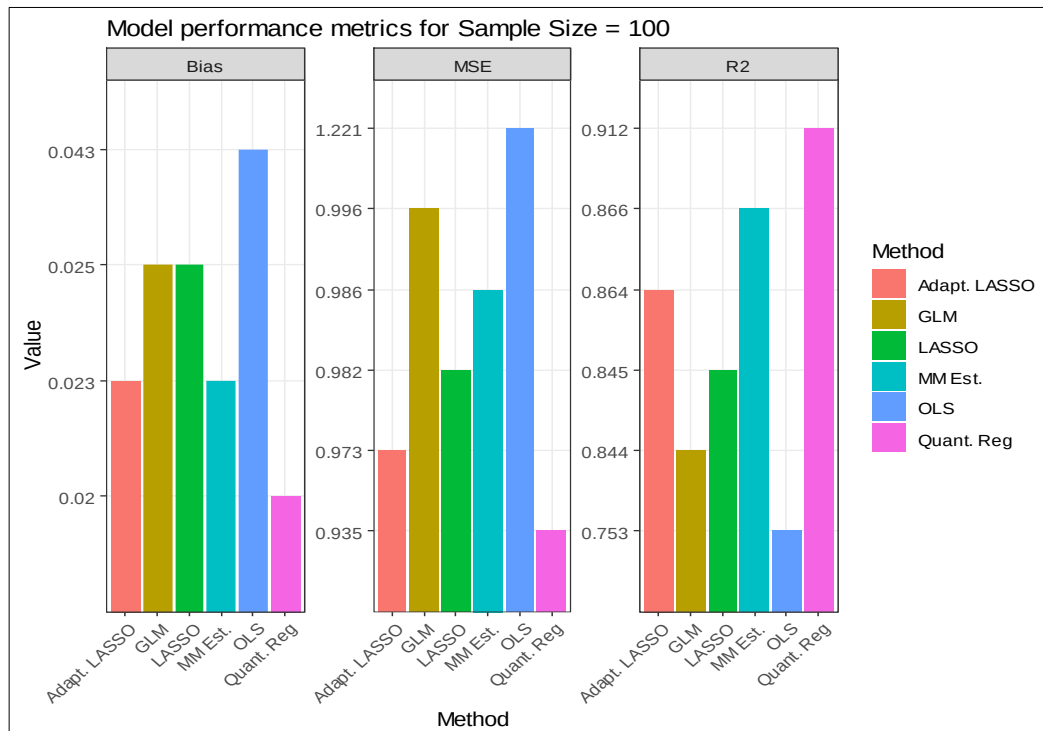


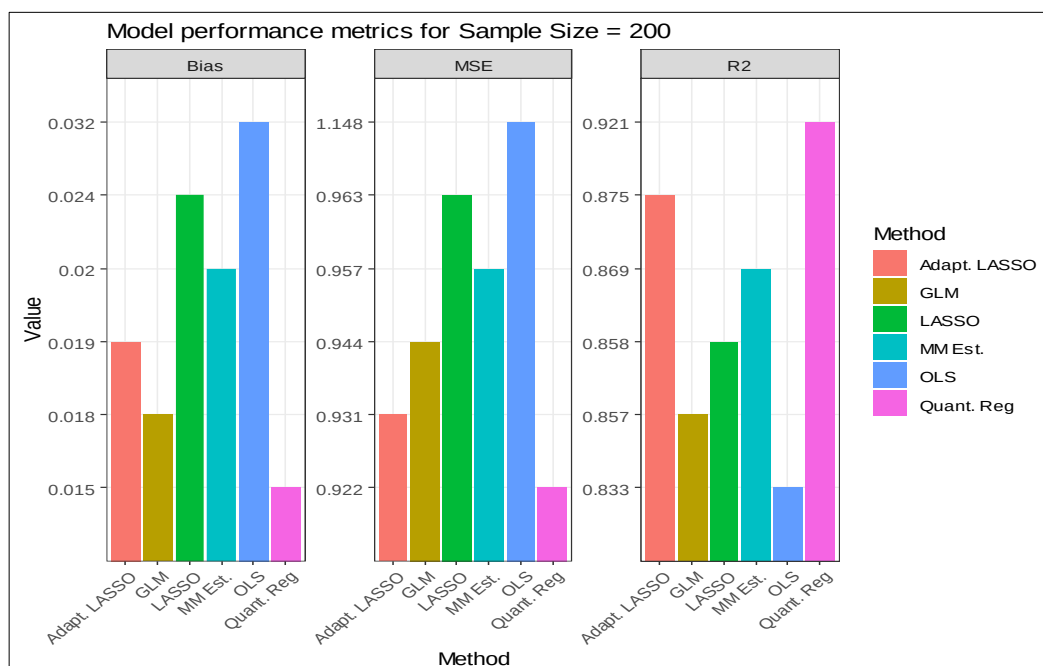
Fig 1: The comparison criteria plot for the methods when $N=50$

Table 2: The MSE, Bais and R^2 for the methods of study when $n = 100$

OLS	1.221	0.043	0.753
GLM	0.996	0.026	0.844
Quant. Reg	0.935	0.020	0.912
LASSO	0.982	0.025	0.845
Adapt. LASSO	0.973	0.023	0.864
MM Est.	0.986	0.023	0.866

**Fig 2:** The comparison criteria plot for the methods when $N=100$ **Table 3:** The MSE, Bais and R^2 for the methods of study when, $N=200$

OLS	1.148	0.032	0.833
GLM	0.944	0.018	0.857
Quant. Reg	0.922	0.015	0.921
LASSO	0.963	0.024	0.585
Adapt. LASSO	0.961	0.019	0.875
MM Est.	0.957	0.020	0.869

**Fig 3:** The comparison criteria plot for the methods when $N=200$

Real data

The dataset employed in this analysis is the classic Iris dataset, originally introduced by R.A. Fisher (1936)^[13], which contains morphological measurements of 150 iris flowers from three species: Iris setosa, Iris versicolor, and Iris virginica. For each specimen, four continuous variables are recorded: sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm). The present study utilizes sepal length as the response variable, with petal length and species as predictors, including their interaction. This configuration enables the evaluation of both main effects and interaction effects between a continuous morphological trait and species identity in predicting sepal length, within a well-balanced, real-world dataset.

Figure 4 demonstrate that the Sepal length grows with petal length in all species and the Versicolor and virginica start with lower baseline sepal lengths but have steeper slopes than setosa. Whereas, the graph will show these three lines starting at different intercepts but all increasing positively, with virginica being the steepest.

Table 4: The Coefficients of real data with lower and upper values

Intercept	4.25000	2.53727	5.81615
Petal. Length	0.50000	-0.07441	1.54688
Species versicolor	-2.06250	-3.60592	-0.45052
Species virginica	-2.74231	-3.91826	-1.88936
Petal. Length: Speciesversicolor	0.37500	-1.10568	1.15509
Petal. Length: Speciesvirginica	0.42308	-0.85308	1.29443

This scatter plot in Figure 4, shows the relationship between Petal Length (x-axis) and Sepal Length (y-axis) for the three iris species; setosa (red), versicolor (green), and virginica (blue) based on the classic *iris* dataset. A quantile regression line for the median ($\tau=0.5$) is fitted separately for each species. Overall, the plot illustrates that as petal length increases, sepal length also tends to increase, with the strength of this relationship varying by species. Quantile regression provides a robust estimate of the median relationship, making it less sensitive to outliers than ordinary least squares regression.

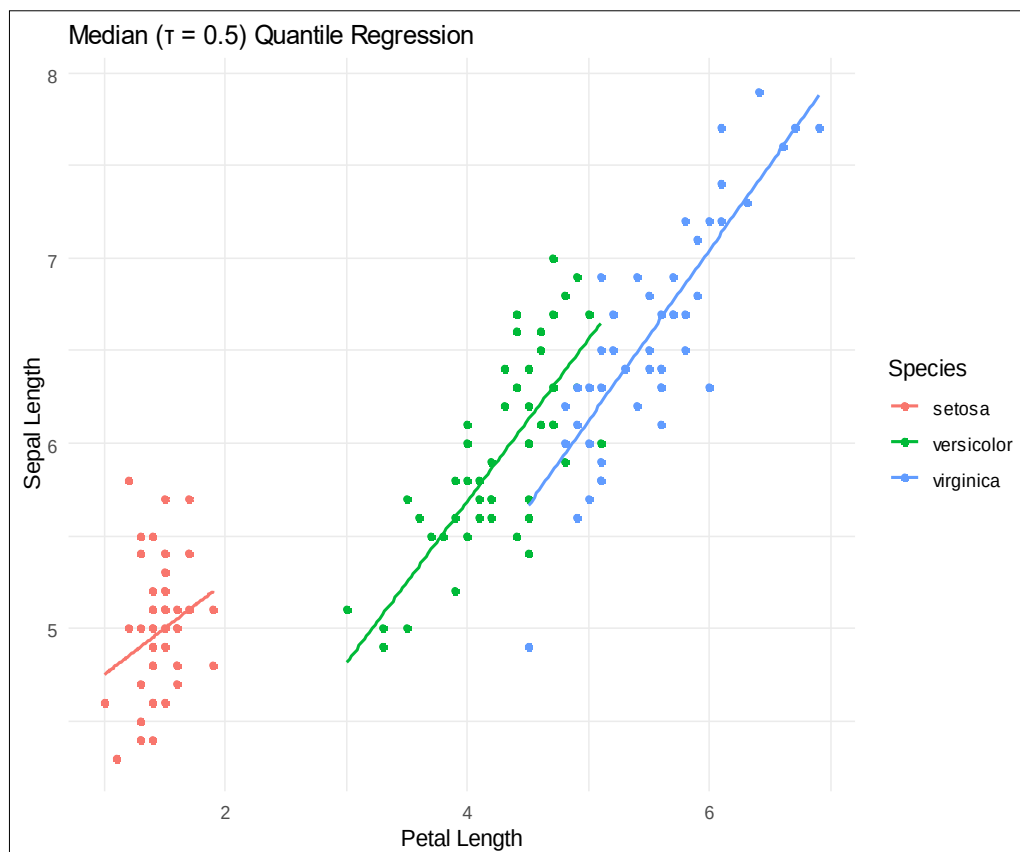


Fig 4: Shows the relationship between Petal Length (x-axis) and Sepal Length (y-axis)

Conclusion

These findings suggest that for factorial designs with non-normal errors, robust and regularized methods like Quantile Regression and Adaptive LASSO are preferable to classical least squares. Simulation results demonstrate that Quantile Regression consistently yields the lowest bias and highest explained variance (R^2) across all sample sizes under non-normal errors. Classical OLS exhibited higher mean squared error and lower R^2 , reflecting sensitivity to skewness. Adaptive LASSO outperformed standard LASSO, especially at larger samples, achieving the lowest MSE. The MM estimator provided robust and stable estimates throughout. Overall, robust and regularized approaches like Quantile Regression and Adaptive LASSO are recommended for factorial designs with non-normal noise, outperforming classical methods.

References

1. Al-Mashhadani AK. Design and analysis of experiments using computer. 1st Ed. Baghdad: Al-Jazeera Printing and Publishing Office, Deposit no. 864, House of Books and Documents, Baghdad.; 2010.
2. Kirk RE. Experimental design: Procedures for the behavioral sciences. 3rd ed. Pacific Grove (CA): Brooks/Cole; 1995.
3. Dobson AJ, Barnett AG. An introduction to generalized linear models. 3rd ed. London: Chapman & Hall; 2008.
4. Brisard E. Pricing of car insurance with generalized linear models. Manchester: University of Manchester; 2014.
5. Ren Y, Zhang X. Model selection for vector autoregressive processes via adaptive lasso. *Commun Stat Theory Methods*. 2013;42(13):2423-36. doi:10.1080/03610926.2011.611317
6. Qian W, Yang Y. Model selection via standard error adjusted adaptive lasso. *Ann Inst Stat Math*. 2013;65(2):295-318. doi:10.1007/s10463-012-0370-0
7. Pepler PT. Detecting active effects in unreplicated 2k factorial experiments: A simulation study. Stellenbosch: University of Stellenbosch; 2006.
8. Lu A, Zhao S. Robust quantile regression analysis on experimental design. Hebei: School of Mathematical Sciences, Hebei Normal University; 2023.
9. Huang Q, Zhang H, Chen J, He M. Quantile regression models and their applications: A review; 2023.
10. Xu M. Quantile regression model and its application research. *Acad J Sci Technol*; 2023. DOI: 10.54097/vt1qpm59
11. Silva PV da. Uso da regressão não linear quantílica na descrição de dados de crescimento. São Paulo: Universidade de São Paulo; 2022. DOI: 10.11606/t.11.2022.tde-10052022-155257
12. Dean A, Voss D. Design and analysis of experiments. New York: Springer-Verlag; 1999.
13. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen*. 1936;7(2):179-88.
14. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58(1):267-88.
15. Hinkelmann K, Kempthorne O. Design and analysis of experiments. Vol. 1: Introduction to experimental design. 2nd Ed. Hoboken (NJ): John Wiley & Sons; 2008.
16. Maronna RA, Martin RD, Yohai VJ. Robust statistics: Theory and methods. Hoboken (NJ): John Wiley & Sons; 2006.