Journal of Econometrics and Statistics
Vol. 1, Issue 1, 2021, pp. 1-16

URL : www.arfjournals.com/jes

ROBUST REGRESSION ANALYSIS UNDER A SPECIAL COMPOUND SYMMETRY STRUCTURE

Rabindra Nath Das[1] and Sabyasachi Mukherjee[2]
[1]*Department of Statistics, The University of Burdwan, Burdwan, WB, India. E-mail: rabin.bwn@gmail.com*
[2]*Department of Mathematics, NSHM Knowledge Campus, Durgapur, WB, India.*

Article History
Received : 9 March 2021
Revised : 12 March 2021
Accepted : 24 March 2021
Published : 2 September 2021

To cite this paper
Das, R. N., & Mukherjee, S. (2021). Robust Regression Analysis under a Special Compound Symmetry Structure. *Journal of Econometrics and Statistics.* 1(1), 1-16.

Abstract: Most real data sets from many sources such as medical sciences, quality engineering, environmental, econometrics etc. are correlated in nature. The present article aims to derive the necessary regression analysis techniques for a correlated data set with a special form of compound symmetry error structure with two sets of observations such that the first set contains only the first observation, and the other set contains the remaining $(N - 1)$ observations, where N is the total number of observations. A constant correlation (ρ_1) is assumed between the first and anyone of the remaining observation, and for the second set, a constant correlation (ρ) is assumed between any two observations within themselves. The variance is assumed constant for all the observations. Correlation structural form is known, but the parameters involved in it are always unknown. In the article, we have derived a robust estimating method for the best linear unbiased estimators (BLUE) of all the regression parameters except the intercept, which is often unimportant. In addition, we have developed a robust testing procedure for any set of linear hypotheses regarding the unknown regression coefficients, and along with a confidence ellipsoid for a set of estimable functions of regression coefficients. Index of fit for the fitted regression equation has also been developed. An example with simulated data illustrates all the developed theories in the article.

Keywords: Confidence ellipsoid; Correlated error; Index of fit; Linear hypothesis; Regression analysis; Robust estimation.

1. Introduction

Regression analysis is conceptually a simple statistical method for establishing the functional relationships among variables. The relationship is expressed in the form of an equation, or a mathematical function connecting the response (dependent variable) with a set of explanatory (independent) variables. Therefore, it can be said that regression analysis is a package full of data analytic techniques which are used to help for understanding the interrelationship among variables in a certain environment. For a detailed regression analysis discussion, readers are suggested to go through the books by Draper and Smith (1998), Chatterjee and Price (2000), Palta (2003), Box and Draper (2007) etc. The data source may be either from environment (environmental data), or may be collected from a controlled experiment (experimental data).

Regression analysis theories are generally derived always with some basic standard assumptions such as the errors are independent and identically distributed (IID) with equal variance. Due to these above assumptions, the ordinary least squares (OLS) method is allowed for estimating the regression parameters. If the errors are correlated with a known dispersion matrix, while the equal variance is unknown, the generalized least squares (GLS) method is allowed for estimating regression parameters. Generally, the dispersion matrix structure can be realized from the data nature, while the correlation parameters that are involved in it are always unknown. There are many sources and causes of arising correlation in the errors which are clearly illustrated in these books by Chatterjee and Price (2000), Palta (2003), Das (2014), Lee *et al.* (2017).

Correlated regression designs are well described in the book by Das (2014), which has been introduced by Panda and Das (1994). There are many research articles on the correlated regression designs by Das (1997, 2003, 2004), and Das and Park (2006, 2007, 2008). For the correlated model, Bischoff (1996) suggested the estimation of regression parameters by OLS method, which is not appropriate. Das (2010, 2014) has developed regression analysis techniques for the compound symmetric, autocorrelated, tri-diagonal correlated error structures. Optimal designs for tri-diagonal and autocorrelated error structures are studied by different authors such as Kiefer and Wynn (1981, 1984), Bischoff (1992, 1995), Box and Draper (2007) etc.

For the correlated regression analysis with unknown error dispersion matrix, GLS method is not applicable for estimating the unknown regression parameters, while the maximum likelihood estimation (MLE) method is used frequently. Mukherjee (1981) has initiated an explicit solution of the ML equations for estimating the unknown correlation parameters for a positive definite variance-covariance matrix, or its inverse through spectral decomposition. Different iterative ML equations solution methods are given in Rubin and Szatrowski (1982), Rogers and Young (1977), Szatrowski (1978), Palta (2003) and Lee *et al.* (2017). Many authors have studied iterative regression coefficients estimation and asymptotic statistical inference methods for the correlated observations with compound symmetry, tri-diagonal, inter-class, intra-class, compound autocorrelated error structures, but there is no study of regression analysis with a special form of compound symmetry correlated error structure as stated in the Abstract.

The rest of the paper is organized as follows. Section 2 presents a correlated regression model and estimation method. Regression parameters interpretation and index of fit, along with their illustrations are presented in Section 3, and concluding remarks are given in Section 4.

2. Correlated Regression Model and Estimations

2.1. Model

Suppose there are p factors $x_1, x_2, \ldots, x_p$ and their u-th observation $(x_{u1}, x_{u2}, \ldots x_{up})$, $1 \le u \le N$, yields a response of y_u on the study variable y. Assuming that the response surface is of first-order, or linear,we adopt the model

$$y_u = \beta_0 + \sum_{k=1}^{p} \beta_k x_{uk} + e_u;\ 1 < u < N$$

or,

$$y = X\beta + e \tag{1}$$

where $y = (y_1, y_2, \ldots, y_N)'$ is the vector of recorded observations on the study variable y, $\beta = (\beta_0, \ldots, \beta_p)'$ is the vector of regression coefficients, $X = (1{:}(x_{uk});\ 1 \le k \le p,\ 1 \le u \le N)$ is the model matrix. Further, e is the vector of errors which are assumed to be normally distributed with $E(e) = 0$ and $D(e) = \sigma^2 W$ with rank $(W) = N$. Therefore 'e' follows a multivariate normal distribution $MN(0, \sigma^2 W)$. The matrix W may represent any correlated error structure. In general, the matrix W is unknown but for all the calculations as usual, W is assumed to be known. In practice, however, W includes a number of parameters unknown, and in the calculations which follow, the expressions for W and W^{-1} are replaced by those obtained by replacing the unknown parameters by their suitable estimates or some assumed values. If there is a curvature in the system, then a polynomial of higher degree, such as the second-order model can be used as given below:

$$y_u = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ui} + \sum\sum_{i \le j=1}^{p} \beta_{ij} x_{ui} x_{uj} + e_u; 1 \le u \le N,$$

or,

$$y = X_1\beta^* + e, \tag{2}$$

where $y = (y_1, y_2, \ldots, y_N)'$ is the vector of recorded observations on the study variable y, $\beta^* = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{11}, \ldots, \beta_{12}, \ldots, \beta_{1p}, \beta_{23}, \ldots, \beta_{2p}, \ldots, \beta_{(p-1)p})'$ is the vector of regression coefficients of order

$$\binom{p+2}{2} \times 1$$

and $X_1 = (1{:}Z^*)$ is the model matrix, where Z^* is given below by using the Hadamard product ($\circ$) as

$$Z^* = (x_1, \ldots, x_p, x_1 \circ x_1, \ldots, x_p, \circ x_2, x_1 \circ x_{p-1} \circ x_p),$$

where $x_i = (x_{1i}, x_{2i}, \ldots, x_{Ni})'$ and $x_i \circ x_j = (x_{1i}x_{1j}, x_{2j}, \ldots, x_{Ni}x_{Nj})'$.

Correlated regression models are well illustrated in the book by Das (2014). These models are used in many fields such as health research (book by Palta, 2003), quality engineering (book by Myers *et al.* 2002; Lee *et al.* 2017) etc. Myers *et al.* (2002, p.128) illustrated that in industrial production processes experimental units are not independent at times *by* design, which incorporates correlation among observations via a repeated measuresscenario as in split plot design.

Intra-class, inter-class, compound symmetry, tri-diagonal and autocorrelated error structures are well described in the book by Das (2014). Intra-class structure is the simplest structure with constant correlation, which is known as uniform structure. Inter-class is an extension of intra-class structure, where within groups there is constant correlation, and in between groups there is no correlation. Compound symmetry structure is an extension of inter-class structure, where within groups there is a constant correlation, and between groups there is another correlation (Das, 2014).

The present article considers a special form of compound symmetry error structure with two sets of observations such that the first set contains only the first observation, and the other set contains the remaining $(N-1)$ observations, where N is the total number of observations. A constant correlation (ρ_1) is assumed between the first and anyof the remaining observations, and for the second set, a constant correlation (ρ) is assumed between any two observations within themselves. The variance is assumed constant (σ^2) for all the observations. This situation is commonly observed when the machine is started initially, the first observation may be recorded with little more disturbance than the remaining others. As a result, the correlation between the first observation with the remaining is little different than the correlation between any two observations of the rest, excluding the first one. This is observed in practice in any production process, or in the measuring units with some instruments, etc. The first group may contain one or more observations. In the very sensitive cases, it may be only the first observation as the first group, and the rest others as the second group.The special form of compound symmetry structure as stated above can be expressed as

$$D(e) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho & \cdots & \rho \\ \rho_1 & \rho & 1 & \cdots & \rho \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho & \rho & \rho & \cdots & \rho_1 \end{pmatrix}$$

It is simply represented by

$$D(e) = w\sigma^2 W \tag{3}$$

2.2. Regression Paramter Estimation

The present section focuses on the derivation of the regression parameters, correlation coefficients and error variance estimation methods. The first-order linear model as given in equation (1) is considered herein, and the similar method can be used for the second-order model as in equation (2).

Suppose there are p-factor $(x_1, x_2, \ldots, x_p)$ and their u-th run is $(x_{u1}, x_{u2}, \ldots, x_{up})$; $1 \le u \le N$ yields a response y_u on the study variable y. Assuming that there are two groups of observations. First group contains only one observation y_1 and the second group contains the other $(N-1)$ observations $(y_2, y_3, \ldots, y_N)$. For first-order linear model we have

$$y_u = \beta_0 + \sum_{k=1}^{p} \beta_k x_{uk} + e_u; 1 \le u \le N$$

or, $$y_u = \beta_1 x_{u1} + \beta_1 x_{u1} + \beta_2 x_{u2} + \ldots + \beta_p x_{up} + e_u;\ 1 \le u \le N$$

or, $$\boldsymbol{y} = X\boldsymbol{\beta} + \mathbf{e} \tag{4}$$

where $y = (y_1, y_2, \ldots, y_N)'$. e_u is the corresponding error of y_u, $\boldsymbol{\beta}$ and x_{uk} are as in equation (1). Note that $E(\boldsymbol{e}) = \mathbf{0}$, $D(e) = \sigma^2 W$, $e \sim MN(0, \sigma^2 W)$ where W as in equation (3).

Let us define

$$Z_u = y_u - y_N;\ u = 1, 2, \ldots, (N-1)$$

or $$Z_u = \beta_1(x_{u1} - x_{N1}) + \beta_2(x_{u2} - x_{N2}) + \ldots + \beta_p(x_{up} - x_{Np}) + (e_u - e_N);\ u = 1, 2, \ldots, (N-1)$$

or $$Z_u = \beta_1 s_{u1} + \beta_2 s_{u2} + \ldots + \beta_p s_{up} + \in_u;\ u = 1, 2, \ldots, (N-1)$$

where,

or $$\boldsymbol{Z} = \boldsymbol{S}\eta + \in \tag{5}$$

Note that, $E(\in_u) = 0$;

$$V(\in_u) = \begin{cases} 2\sigma^2(1-\rho_1);\ u = 1 \\ 2\sigma^2(2-\rho);\ u = 2, 3, \ldots, (N-1) \end{cases}$$

$$Cov(\in_u, \in_{u'}) = \sigma^2(1-\rho);\ 1 \le u \ne u' \le (N-1)$$

Therefore, $E(\in) = 0$, $D(\in) = \in_1^2 W_1$ (say), where $\sigma_1^2 = 2\sigma^2(1-\rho)$ and W_1 is defined as follows :

$$W_1 = \begin{pmatrix} \frac{1-\rho_1}{1-\rho} & \frac{1}{2} & \ldots & \frac{1}{2} \\ \frac{1}{2} & 1 & \ldots & \frac{1}{2} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \frac{1}{2} & \frac{1}{2} & \ldots & 1 \end{pmatrix}$$

The first cell element of W_1 is $\frac{1-\rho_1}{1-\rho}$ which contains two unknown parameters ρ_1 and ρ. We partitioned W_1 into four sub matrix namely $W_{1_{11}}$, $W_{1_{12}}$, $W_{1_{21}}$ and $W_{1_{22}}$ define as follows :

$$W_1 = \begin{bmatrix} W_{1_{11}} & W_{1_{12}} \\ W_{1_{21}} & W_{1_{22}} \end{bmatrix}$$

where the first partitioned matrix $W_{1_{11}}$ contains only the first cell element is $\frac{1-\rho_1}{1-\rho}$, $W_{1_{12}}$ and $W_{1_{21}}$ are the transpose to each other having the elements $\left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$ with dimension $(\overline{N-1} \times 1)$.

For estimation purposes, we have taken the marginal distribution of $Z_0 = (Z_2, Z_3, \ldots, Z_{(N-1)})'$, excluding Z_1 from the Z. From the model as in equation (5), it can be considered for Z_0 as follows:

$$Z_u = \beta_1 s_{u1} + \beta_2 s_{u2} + \beta_p s_{up} + \in_u;\ u = 2, 3, \ldots, \overline{N-1}$$

or

$$Z_0 = S_0 \eta + \in_0 \tag{6}$$

where S_0 is the new model matrix with $s_{uj} = (x_{uj} - x_{Nj})$; $u = 2, 3, \ldots, \overline{N-1}$; $j = 1, 2, \ldots, p$; $\in_0 = (\in_2, \in_3, \ldots, \in_{N-1})'$, $E(\in_0) = 0$; $D(\epsilon_0) = \alpha_1^2\ W_{1_{22}}$, $\sigma_1^{\ 2} = 2\sigma^2(1-\rho)$; and $\in_0 \sim MN(0,\ \sigma_1^2\ W_{1_{22}})$. The related dispersion matrix is $W_{1_{22}}$ (with order $\overline{N-2} \times \overline{N-2}$), and it can be shortly written as $W_{1_{22}} = \left(\frac{1}{2} I_{N-2} + \frac{1}{2} E_{N-2}\right) = W_2$ (say), where I is an identity matrix, E is a matrix of all elements unity, and $W_{1_{22}}$ is explicitly expressed as follows:

$$W_{1_{22}} = \begin{pmatrix} 1 & \frac{1}{2} & \ldots & \frac{1}{2} \\ \frac{1}{2} & 1 & \ldots & \frac{1}{2} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \frac{1}{2} & \frac{1}{2} & \ldots & 1 \end{pmatrix}$$

The model in equation (6) is a generalized linear least squares model (with known $W_{1_{22}} = W_2$, say). Therefore, we have the following results for the reduced model (6).

Theorm 1. Under the model (6), the best linear unbiased estimator (BLUE) of η is

$$\hat{\eta} = (S_0' W_2^{-1} S_0)^{-1} (S_0' W_2^{-1} Z_0) \tag{7}$$

where $W_2^{-1} = \left(2 * I_{N-2} - \frac{2}{N-1} E_{N-2}\right)$.

Theorm 2. An unbiased estimator (UE) of $\sigma_1^{\ 2} = 2\sigma^2(1-\rho)$ is

$$\hat{\sigma}_1^2 = \frac{z_0' W_2^{-1} z_0 - \hat{\eta}' (S_0' W_2^{-1} S_0) \hat{\eta}}{(N-2) - p} \tag{8}$$

Note that $\hat{\eta} \sim MN(\eta, \sigma_1^2(S_0' W_2^{-1} S_0)^{-1})$, and $\hat{\eta}$ does not depend on the other unknown parameters ρ, ρ_1 and ρ_1^2.

The scheme for calculations of other unknown parameters is given hereunder. From equation (4), one can find the estimate of β_0 as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \ldots - \hat{\beta}_p \bar{x}_p \tag{9}$$

where $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ are as in equation (7) because $\hat{\eta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)'$, $\bar{y}, \bar{x}_1, \ldots, \bar{x}_p$ (respective means) are known.

Theorem 3. An estimate of ρ_0 for known is σ^2 is

$$\hat{\rho} = 1 - \frac{1}{2\sigma^2} \frac{Z_0' W_2^{-1} Z_0 - \hat{\eta}(S_0' W_2^{-1} S_0)\hat{\eta}}{(N-2) - p} \tag{10}$$

To estimate ρ for unknown σ^2, an estimate of σ^2 is required.

Theorem 4. An estimate of σ^2 (from the full model (4)) is

$$\hat{\sigma}^2 = \frac{\widehat{e_0'}\widehat{e_0}}{N - p - 1} \tag{11}$$

where $\widehat{e_0} = W^{-\frac{1}{2}}(Y - X\hat{\beta}), \hat{\beta} = (\hat{\beta}_0, \hat{\eta}')'$ and W is obtained from the scheme given below.

The scheme for calculations of W (*i.e.*, ρ, ρ_1 and σ^2) is given below.

1. Assume some value of $\rho_1 \in (-1, 1)$.
2. Compute $\hat{\rho}$ using Equation (10) (taking $\sigma^2 = 1$ in the first iteration, and for any other iteration, $\hat{\sigma}^2$ by plugging for σ^2 obtained in step 5 just in the previous iteration).
3. With the assumed value of ρ_1, say $\hat{\rho}_1$, and the estimate of ρ, say $\hat{\rho}$ in step 2, compute W (examining W is non-singular) and W^{-1} as in Equation (3).
4. Calculate $\hat{\beta} = (X'W^{-1}X)^{-1}(X'W^{-1}Y)$, assuming X has full column rank.
5. Compute $\hat{\sigma}^2$ using Equation (11).
6. Calculate $S_0(\hat{\rho}_1, \hat{\beta}) = (Y - X\hat{\beta})' W^{-1}(Y - X\hat{\beta})$, where W^{-1} as in step 3 and $\hat{\beta}$ is as in step.

The same routine of calculations 1 through 6 is to be followed for different permissible values of ρ_1 in its range. We select that value of ρ_1 as the final estimate of ρ_1 for which $S_0(\hat{\rho}_1, \hat{\beta})$ is minimum. For the final estimate of ρ_1, we get the final estimate of ρ in step 2. Thus, for the final estimates of ρ and ρ_1, one can compute W (an estimate of W) in step 3. Note that the estimates of all the regression parameters β_0 (in (9)) and η in (7) are free of and , so the above derived estimation procedure of regression parameters is a robust method.

3. Inference of Regression Parameters and Index of Fit

Testing of hypothesis regarding the unknown regression parameters is an important problem under regression analysis. The present section focuses on the necessary results for testing a set of linear hypotheses based on the model (6), where $\epsilon_0 \sim MN(0, \sigma_1^2 W_2)$, σ_1^2 is unknown but W_2 is known.

3.1. Testing of Hypothesis

A set of m linear independent hypothesis regarding the unknown regression parameters are stated as follows:

$$H_0 : \begin{cases} l_{11}\beta_1 + \ldots + l_{1p}\beta_p = l_{10}(known) \\ l_{21}\beta_1 + \ldots + l_{2p}\beta_p = l_{20}(known) \\ \vdots \\ l_{m1}\beta_1 + \ldots + l_{mp}\beta_p = l_{m0}(known) \end{cases}$$

or $H_0 : R\eta = l_0$ (known) against $H_A : R\eta \neq l_0$. Here rank $(R) = m$, where

$$R = \begin{pmatrix} l_{11} & \cdots & l_{1p} \\ l_{21} & \cdots & l_{2p} \\ \cdots & \cdots & \cdots \\ l_{m1} & \cdots & l_{mp} \end{pmatrix} \text{ and } l_0 = (l_{10}, \ldots, l_{m0})'.$$

Note that $\hat{\eta} \sim MN(\eta, \sigma_1^2(S_0' W_2^{-1} S_0)^{-1})$ as given in Section 2, therefore,

$R\hat{\eta} \sim MN(R\eta, \sigma_1^2 R(S_0' W_2^{-1} S_0)^{-1} R')$ and under H_0, $R\hat{\eta} \sim MN(l_0, \sigma_1^2 R(S_0' W_2^{-1} S_0)^{-1} R')$.

Therefore, under H_0

$$(R\hat{\eta} \sim l_0)' [\sigma_1^2 R(S_0' W_2^{-1} S_0)^{-1} R']^{-1} (R\hat{\eta} \sim l_0) \sim \chi_m^2,$$

where the degree of freedom m is given by the number of independent linear hypotheses in the $R\eta$ vector. Also for the model in Equation (6), $\hat{\epsilon}_{0R} = W_2^{-\frac{1}{2}}(Z_0 - S_0\hat{\eta})$, $\hat{\epsilon}_{0R}\hat{\epsilon}_{0R} / \sigma_1^2 \sim \chi^2_{(N-2)-p}$, and it is independent of $R\hat{\eta}$. Thus we have the following result.

Theorem 5. If $R\eta = l_0$ is true, the basic result is

$$F = \frac{(R\hat{\eta} - l_0)'[R(S_0' W_2^{-1} S_0)^{-1} R']^{-1}(R\hat{\eta} - l_0)/m}{\hat{\epsilon}_{0R}' \hat{\epsilon}_{0R} / \{(N-2) - p\}} \sim F_{m,(N-2)-p}. \tag{12}$$

H_0 is rejected at 100α% level of significance, if observed $F > F_{\alpha;\, m,(N-2)-p}$, and accepted otherwise. Note that the test statistic in (12) is free of ρ_1 and ρ, so the above test procedure is robust.

3.2. Confidence ellipsoids of regression parameters

In the present subsection, the confidence ellipsoids for a set of independent linear functions of regression parameters and the confidence interval for a linear function of regression parameters are developed. Note that S_0 has full rank (assumed), so every linear function of β_1, β_1, ..., β_p are estimable for the linear model in Equation (6).

Suppose ψ_1, ψ_2, ..., ψ_v, are v independent linear functions of β_1, β_1, ..., β_p. Let $\psi_{\text{v}} = (\psi_1, \psi_2, ..., \psi_v)'$ be a vector of order $v \times 1$. Then

$$\psi = C\eta,$$

where $\eta = (\beta_1, ..., \beta_p)'$ and C(known) is a $v \times p$ matrix whose rows are linearly independent. Then $\hat{\psi} = AZ_0$ is the GLS estimate of ψ (for the model in Equation (6)), and $A = ((a_{ij}))$ (known matrix depending on ψ_{is}). Thevariance-covariance matrix of $\hat{\psi}$ is then

$$Dis(\hat{\psi}) = ADis(Z_0)A' = \sigma_1^2(AW_2A'),$$

where $W_2 = \left(\frac{1}{2}I_{N-2} + \frac{1}{2}E_{N-2}\right)$. Note that $\hat{\psi} \sim MN(\psi, \sigma_1^2(AW_2A'))$ and is independent of

$$\frac{\hat{\in}'_{0R}\hat{\in}_{0R}}{\sigma_1^2} = \frac{(Z_0 - S_0\hat{\eta})'W_2^{-1}(Z_0 - S_0\hat{\eta})}{\sigma_1^2} \sim \chi^2_{(N-2)-p}$$

Again

$$(\hat{\psi} - \psi)'\{\sigma_1^2(AW_2A')\}^{-1}(\hat{\psi} - \psi) \sim \chi_v^2,$$

and independent of $\hat{\in}_{0R}\hat{\in}_{0R}/\sigma_1^2$ (for the model in Equation (6)). Therefore, we have the following result.

Theorem 6. The distribution of the test statistic is

$$F = \frac{(\hat{\psi} - \psi)'(AW_2A')^{-1}(\hat{\psi} - \psi)/v}{(\hat{\in}_{0R}\hat{\in}_{0R})/((N-2) - p)} \sim F_{v,(N-2)-p} \tag{13}$$

Therefore,

$$(\hat{\psi} - \psi)'(AW_2A')^{-1}(\hat{\psi} - \psi) \leq vs^2 F_{\alpha;v,(N-2)-p} \tag{14}$$

where $S^2 = \frac{\hat{\in}'_{0R}\hat{\in}_{oR}}{(N-2) - p}$ which is an UE of σ_1^2 in Equation (8).

Inequality (14) determines an *ellipsoid* in the v-dimensional ψ-space with center $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2, ..., \hat{\psi}_v)'$, and the probability that this *random ellipsoid* covers the true parameter ψ is $(1 - \alpha)$, no matter whatever be the values of ρ_1 and ρ unknown parameters.

We may obtain a confidence interval for a single linear function $\psi = c'\eta (c \neq 0)$ by specializing the above calculation to $v = 1$. The resulting confidence interval is given by

$$(a'W_2 a)^{-1}(\hat{\psi} - \psi)^2 \leq s^2 F_{\alpha;1,(N-2)-p} \tag{15}$$

where $\hat{\psi} = a'Z_0$ is the GLS estimate of ψ. Note that $Var(\hat{\psi}) = a'Dis(Z_0)\alpha = \alpha_1^2(a'W_2 a)$, and its unbiased estimate $\hat{\sigma}^2_{\hat{\psi}} = s^2(a'W_2 a)$. We may write the Inequality (15) as

$$\hat{\psi} - t_{\frac{\alpha}{2},(N-2)-p}\hat{\sigma}_{\hat{\psi}} \leq \psi + \hat{\psi} - t_{\frac{\alpha}{2},(N-2)-p}\hat{\sigma}_{\hat{\psi}}, \tag{16}$$

the probability that this random interval covers the unknown ψ is $(1 - \alpha)$. The interval (16) could also be derived from the fact that $(\hat{\psi} - \psi)/\hat{\sigma}_{\hat{\psi}} \sim t_{(N-2)-p}$. Note that the above inference procedures are free of the values of ρ_1 and ρ unknown parameters. Therefore, all the derived inference procedures are robust.

3.3. Index of fit

The original model of the present study is given in Equation (4), and Equation (6) is the transformed (or reduced) model. In the present section, the index of fit is suggested for the models in Equations (4) and (6). Analogous to uncorrelated errors, two criteria of judging the best fit are described for the models in Equations (4) and (6), under a special form of compound symmetry error structure in Equation (3).

For multiple regression analysis with uncorrelated and homoscedastic errors, the index of fit is measured by the multiple correlation coefficient (R^2), and adjusted multiple correlation coefficient (R^2_{adj}) of the fitted regression model. Analogous to uncorrelated case, we define the multiple correlation coefficients $R^2(Y)$, $R^2(Z_0)$, and adjusted multiple correlation coefficients $R^2_{adj}(Y)$, $R^2_{adj}(Z_0)$ for the fitted models in Equations (4) and (6), respectively, as follow:

$$R^2(Y) = Corr.^2(Y, \hat{Y}) \text{ and } R^2(Z_0) = 1 - \frac{\hat{\epsilon}_{0R}\hat{\epsilon}_{0R}}{TSS_{Z_0}}, \tag{17}$$

$$R^2_{adj}(Y) = 1 - \frac{N-1}{N-p-1}(1 - R^2(Y)) \text{ and } R^2_{adj}(Z_0) = 1 - \frac{(N-2)-1}{(N-2)-p}(1 - R^2(Z_0)), \tag{18}$$

where $\hat{\epsilon}_{0R} = W_2^{-\frac{1}{2}}(Z_0 - S_0\hat{\eta})$, $TSS_{Z_{0'}} = (Z_0 - \bar{Z}_0)'W_2^{-1}(Z_0 - \bar{Z}_0)$, $\bar{Z}_0 = \sum_{j=2}^{N-1} z_j/(N-2)$ (for the model in Equation (6)). Generally, R^2 and R^2_{adj} as in Equations (17) and (18) are both close to unity for a good fitted model (Illustration Section 3.4).

3.4. Illustration

In the above, all the necessary regression analysis results are derived based on theory. It is noted that the considered special compound symmetry correlated error structure contains two unknown correlation coefficients ρ_1 and ρ, but the derived results are free of both these two correlation coefficients. In this section, a simulated data example illustrates all the above derived results. Let 'y' be the response variable with sixty observations in total. According to the defined structure there are two groups, the first group contains only one observation and the other has the rest fifty nine observations.

The model matrix X is formed using three factors (exploratory variables) x_1, x_2 and x_3. The appropriate changes of origin and scale are used for the exploratory variables such that the values lie between –1 to 1 (the range within which the experimentation is conducted). Considering three factors as explanatory variables the assumed model is

Table 1 : Responses under the simulation setting of ($\sigma^2 = 2$, $\rho = 0.8$, $\rho_1 = 0.1$)

Observation	*Value*	*Observation*	*Value*	*Observation*	*Value*
1	3.358	21	3.265	41	6.829
2	3.873	22	4.320	42	–0.393
3	–0.441	23	3.233	43	2.532
4	7.062	24	–0.419	44	4.559
5	2.014	25	6.985	45	2.295
6	3.009	26	7.045	46	3.168
7	5.078	27	–0.381	47	4.566
8	2.695	28	3.221	48	–0.959
9	–0.965	29	4.663	49	16.722
10	7.458	30	3.374	50	2.608
11	7.044	31	3.263	51	3.154
12	–1.533	32	4.580	52	4.773
13	3.060	33	–0.572	53	2.433
14	3.759	34	7.207	54	–0.486
15	3.590	35	3.359	55	7.216
16	2.924	36	3.093	56	7.194
17	5.147	37	5.248	57	0.045
18	–0.854	38	3.246	58	3.165
19	6.499	39	–0.777	59	4.104
20	2.746	40	7.341	60	3.498

$$y_u = \beta_0 + \beta_1 x_{u1} + \beta_2 x_{u2} + \beta_3 x_{u3};\ u = 1, 2, ..., 60 \tag{19}$$

The generated values are displayed in Table 1, using the fixed model matix X, which is defined as

$$X = \begin{pmatrix} D \\ D \\ D \\ D \end{pmatrix}, \text{ where } D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

In the absence of real data we generate observations according to the formula (19) with β_0 = 3.5, β_1 = 2.5, β_2 = –1.5, β_3 = 0.05, σ^2 = 2, ρ = 0.8 and ρ_1 = 0.1 using the above model matrix "X" and $e \sim MN(0, \sigma^2 W)$ where W is given in equation (3). The observations obtained are given in Table 1. For this present simulation study, the article considers the following eight combinations of parameters (σ^2, ρ, ρ_1), with fixed β_0 = 3.5, β_1 = 2.5, β_2 = –1.5, β_3 = 0.05. The combinations are σ^2 = 1, 2; ρ = 0.4, 0.8; ρ_1 = 1.1, 0.6.

Using these combinations, we take each simulation setting and repeat the entire calculation 100 times. The sample bias, sample variance for every estimate, where sample bias and sample variance for the parameter θ are defined by

$$Bias(\hat{\theta}) = \left|\bar{\hat{\theta}} - \theta\right|, \bar{\hat{\theta}} = \frac{\Sigma\hat{\theta}}{100} \text{ and } Var(\hat{\theta}) = \frac{\Sigma(\hat{\theta} - \theta)^2}{100}$$

Summarized simulation results are given in Table 2.

We consider the following four linear hypotheses (given in Table 3) for testing of hypotheses. Each hypothesis is tested 100 times using the equation (12) and the results are reported in Table 3.

The average values of 200 replicates for two index of fit measures $R_2(Y)$, $R^2(Z_0)$, $R^2_{adj}(Y)$ and $R^2_{adj}(Z_0)$ using the equations (17 and 18) are reported in Table 4. These two index of fit measures are

Table 2: Simulation results: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\rho}, \hat{\rho}_1, \hat{\sigma}^2$ and $\hat{\sigma}_1^2$

	$\beta_0 = 3.5$ $\hat{\beta}_0$	$\beta_1 = 2.5$ $\hat{\beta}_1$	$\beta_2 = -1.5$ $\hat{\beta}_2$	$\beta_3 = 0.05$ $\hat{\beta}_3$	$\hat{\rho}$	$\hat{\rho}_1$	$\hat{\sigma}^2$	$\hat{\sigma}_1^2$
				$\rho = 0.8, \rho_1 = 0.1, \sigma^2 = 1$				
Mean	3.441	2.505	–1.500	0.040	0.759	0.931	0.891	0.405
Bias $(\hat{\theta})$	0.059	0.005	0.0002	0.009	0.041	0.831	0.109	0.005
Var $(\hat{\theta})$	0.863	0.004	0.004	0.003	0.015	0.005	0.018	0.006
				$\rho = 0.8, \rho_1 = 0.1, \sigma^2 = 2$				
Mean	3.593	2.508	–1.502	0.048	0.552	0.829	0.920	0.811
Bias $(\hat{\theta})$	0.093	0.008	0.002	0.001	0.248	0.729	1.080	0.011
Var $(\hat{\theta})$	1.453	0.007	0.010	0.008	0.010	0.003	0.011	0.024
				$\rho = 0.4, \rho_1 = 0.1, \sigma^2 = 1$				
Mean	3.473	2.501	–1.502	0.045	0.375	0.743	0.983	1.230
Bias $(\hat{\theta})$	0.026	0.001	0.002	0.004	0.025	0.643	0.016	0.030
Var $(\hat{\theta})$	0.417	0.013	0.013	0.014	0.011	0.003	0.0003	0.050
				$\rho = 0.4, \rho_1 = 0.1, \sigma^2 = 2$				
Mean	3.460	2.532	–1.504	0.044	–0.232	0.443	1.009	2.490
Bias $(\hat{\theta})$	0.039	0.032	0.004	0.005	0.632	0.343	0.990	0.090
Var $(\hat{\theta})$	0.794	0.023	0.023	0.028	0.046	0.012	0.0003	0.202
				$\rho = 0.4, \rho_1 = 0.6, \sigma^2 = 1$				
Mean	3.542	2.505	–1.507	0.043	0.406	0.759	0.996	1.184
Bias $(\hat{\theta})$	0.042	0.005	0.007	0.006	0.006	0.159	0.003	0.015
Var $(\hat{\theta})$	0.505	0.010	0.011	0.013	0.011	0.003	0.0001	0.045
				$\rho = 0.4, \rho_1 = 0.6, \sigma^2 = 2$				
Mean	3.551	2.507	–1.514	0.040	–0.198	0.460	1.0009	2.399
Bias $(\hat{\theta})$	0.051	0.007	0.014	0.009	0.598	0.140	0.999	0.0003
Var $(\hat{\theta})$	0.728	0.024	0.019	0.025	0.043	0.011	0.000	0.177
				$\rho = 0.8, \rho_1 = 0.6, \sigma^2 = 1$				
Mean	3.484	2.505	–1.495	0.054	0.789	0.945	0.956	0.401
Bias $(\hat{\theta})$	0.015	0.005	0.004	0.004	0.010	0.345	0.043	0.001
Var $(\hat{\theta})$	0.795	0.003	0.004	0.003	0.001	0.002	0.004	0.004
				$\rho = 0.8, \rho_1 = 0.6, \sigma^2 = 2$				
Mean	3.467	2.495	–1.497	0.058	0.585	0.848	0.962	0.794
Bias $(\hat{\theta})$	0.032	0.004	0.002	0.008	0.214	0.248	1.037	0.005
Var $(\hat{\theta})$	1.349	0.007	0.008	0.007	0.007	0.003	0.004	0.026

Table 3 : Test result from 100 replications with α = 0.05

Null Hypothesis	*Degrees of freedom*	*Accepted cases*	*Rejected cases*
$H_{01}: \beta_1 = \beta_2 = \beta_3 = 0$	(3, 54)	0	100
$H_{02}: \beta_1 = 0$	(1, 54)	0	100
$H_{03}: \beta_2 = 0$	(1, 54)	0	100
$H_{04}: \beta_3 = 0$	(1, 54)	93	7

considered for the following four models (replicates generated with $\beta_0 = 3.5$, $\beta_1 = 2.5$, $\beta_2 = -1.5$, $\beta_3 = 0.05$, $\sigma^2 = 2$, $\rho = 0.8$ and $\rho_1 = 0.1$ values):

$$M_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e,$$
$$M_2: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e,$$
$$M_3: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e,$$
$$M_4: y = \beta_0 + \beta_1 x_1 + e.$$

Table 4: Average index of fit measures from 200 replicates for the four models M_1 to M_4

Model	$R^2(Y)$	$R^2(Z_0)$	$R^2_{adj}(Y)$	$R^2_{adj}(Z_0)$
M_1	0.9251	0.9255	0.9211	0.9227
M_2	0.9222	0.9224	0.9181	0.9196
M_3	0.6971	0.3975	0.6916	0.3938
M_4	0.6954	0.3959	0.6898	0.3921

4. Conclusions

The present article derives the regression analysis with correlated observations under a special type of compound symmetry correlated error structure. The derived estimation method gives the best linear unbiased estimator (BLUE) for all the regression parameters except the intercept. Analytically, the estimates of all regression coefficients β's, σ^2, ρ and ρ_1 are derived herein. The simulation study clearly shows that each estimated value is very close to its imputed value. Table 3 reflects the true values of regression parameters. The true model can be selected from Table 4, which expresses that M_1 and M_2 models are equivalent, while models M_3 and M_4 are incorrect. The values of $R^2(Y)$ and $R^2_{adj}(Y)$ are the real index of fit measures for the original model, whereras $R^2(Z_0)$ and $R^2_{adj}(Z_0)$ are the measures for the reduced model. It can be observed that the value of $R^2(Z_0)$ (or, $R^2_{adj}(Z_0)$) is more than $R^2(Y)$ (or, $R^2_{adj}(Y)$) for the correct models and these values are less for incorrect models (Table 4), as $R^2(Z_0)$ is based on the BLUEs. It has been observed that for a regression model with a special type of compound symmetric error structure (considered herein), the estimates of regression parameters $(\hat{\eta})$ are generally used for deriving all the results, while the estimates $\hat{\beta}_0$, $\hat{\rho}_1$, and $\hat{\rho}$ are not used in any derived results. The estimates $\hat{\beta}_0$, $\hat{\rho}_1$ and are used in case of the index of fit measure for the full model only, which is not important in the study. All the derived results are free of all the values of the two correlation coefficients, so the present study is a robust regression analysis.

Acknowledgements

The authors are very much indebted to the referees and the Editor-In-Chief who have provided valuable comments to improve this paper.

References

Bischoff, W. (1992). On exact D-optimal designs for regression models with correlated observations, *Annals of the Institute Statistical Mathematics*, 44, 229–238.

Bischoff, W. (1995). Determinant formulas with applications to designing when the observations are correlated, *Annals of the Institute Statistical Mathematics*, 47, 385–399.

Bischoff, W. (1996). On maximin designs for correlated observations, *Statistics and Probability Letters*, 26, 357–363.

Box, G.E.P. and Draper, R.N. (2007). *Response surfaces, Mixtures, and Ridge Analyses*, 2nd ed., Wiley & Sons, New York.

Chatterjee, S. and Price, B. (2000). *Regression Analysis by Example*, 3rd ed., Wiley Sons, New York.

Das, R.N. (1997). Robust Second Order Rotatable Designs: Part-I, *Cal. Statist. Assoc. Bull.*, 47, 199–214.

Das, R.N. (2003). Robust Second Order Rotatable Designs: Part-III, *Journal of the Indian Society of Agricultural Statistics*, 56, 117–130.

Das, R.N. (2004). Construction and Analysis of Robust Second Order Rotatable Designs, *Journal of Statistical theory and Applications*, 3, 325–343.

Das, R.N. (2010). Regression Analysis for Correlated Data, *J. of Quality Technology and Quality Management*, 7(3), 263–277.

Das, R.N. (2014). *Robust Response Surfaces, Regression, and Positive Data Analyses*, Chapman Hall, London.

Das, R.N. and Park, S.H. (2006). Slope rotatability over all directions with correlated errors, *Appl. Stochastic Models Bus. Ind.*, 22, 445–457.

Das, R. N. Park, S. H. (2007). A measure of robust rotatability for second order response surface designs, *Journal of the Korean Statistical Society*, 36(4), 557–578.

Das, R. N. and Park, S. H. (2008). Analysis and multiple comparison of treatments of an extended randomized block design with correlated errors, *Journal of Statistical Theory and Application* , 7(3), 245–262.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, Wiley Sons, New York.

Kiefer, J. and Wynn, H. P. (1981). Optimum balanced block and latin square designs for correlated observations, *Annals of Statistical*, 9, 737–757.

Kiefer, J. and Wynn, H. P. (1984). Optimum and minimax exact treatment designs for one-dimensional autoregressive error processes, *Annals of Statistical*, 12, 431–449.

Lee, Y., Ronnegard, L., and Noh, M. (2017). Data Analysis Using Hierarchical Generalized Linear Models with R (1st ed.). Chapman and Hall/CRC.

Mukherjee, B.N. (1981). A simple approach to testing of hypotheses regarding a class of covariance structures, *Indian Statistical Institute Golden Jubilee Inter. Conf. on Statistics: Applications and new directions, Kolkata 16–19 Dec., 1981*, 442–465.

Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley Sons, New York.

Palta, M. (2003). *Quantitative Methods in Population Health: Extensions of Ordinary Regression*, Wiley Sons, New York.

Panda, R.N. and Das, R.N. (1994). First Order Rotatable Designs With Correlated Errors, *Cal. Statist. Assoc. Bull.*, 44, 83–101.

Rogers, G.S. and Young, D.L. (1977). Explicit maximum likelihood estimators for certain patterned covariance matrices, *Communications in Statistics*, 6, 121–133.

Rubin, D. and Szatrowski, T. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the (EM) algorithm, *Biometrika*, 69(3), 657–660.

Szatrowski, T. (1978). Explicit solutions, one iteration convergence and averaging in the multivariate normal estimation problem for patterned means and covariances, *Ann. Inst. Stat. Math.*, 30, 81–88.