Journal of the Nigerian Association of Mathematical Physics
Volume 65, (October 2022– August 2023 Issue), pp217 – 224

ALTERNATIVE ESTIMATOR IN THE ANALYSIS OF VARIANCE TECHNIQUE IN THE PRESENCE OF OUTLIERS

ODIOR A.K.[1] and OYEYEMI G. M.[2]

[1]Department of Statistics, Delta-State Polytechnic Otefe Oghara, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria

Abstract

The estimation of ANOVA model parameters in the presence of outliers is one of the most pervasive problems in data analysis, statistical applications and inferences. The regularity of heavy tailed error distributions due to the presence of outliers in both experimental and observational data is of keen interest to researchers due its negative impact on most useful classical techniques in the field of statistical inferences. Authors at various times have examined empirically the problems of outliers in data analysis and inference from different considerations and various estimators including the classes of M estimators with fixed cut off point have been suggested in literature to address the limitations of the classical methods. Consequently, this paper examines the efficiency of the proposed alternative estimator: Adaptive Robust M Estimator (ARME) with data dependent (flexible) cut off point. The efficiency (robustness) of the proposed method and the other existing methods: Huber M Fixed Cut off (HMFC), Bisquare M Fixed Cut off (BMFC) and Least Square Estimator (LSE) was compared using Monte Carlos simulated data for One-Way ANOVA with varying percentages of outliers on the response variable crossed with different sample size. The performance of the estimators was assess using Root Mean Square Error (RMSE). The results of the study revealed that the performance of the proposed estimator (ARME) is substantially better when compare with the existing methods using RMSE as measure of efficiency and goodness of fit at different degree of outliers.

Keywords: Outliers, Adaptive, Robust, Estimators, RMSE, Data Dependent, Cut Off Point.

1. Introduction

Analysis of Variance (ANOVA) is one major established classical parametric statistical method that is frequently engaged in diverse field of studies particularly in design and analysis of experiment. Montgomery (2010) asserted that ANOVA is probably the most useful technique in the field of statistical inference. The classical ANOVA is a standard procedure used to generate confident statistical inferences about systematic differences between group means of normally distributed outcome measures in randomized experiment Kevin (2004). A statistical procedure focuses on the theory of statistical data produced by an experiment of various dimension. ANOVA methodology is aptly and sufficiently described as one of the most flexible and feasible techniques for comparing several population means. ANOVA tool provides the methodology for partitioning the total variation computed from the dataset into components, each of which represents the amount of the total variances that can be attributed to a specific source of variation Wayne (2013). The attractiveness and usefulness of this statistical tool is hinged on its strength and capacity to separate the total variability found within the data set into several components.

However, the methodology of ANOVA framework is most powerful and resilient when the classical assumptions of homogeneity of variance, independent and normality of error distribution are sufficiently fulfilled. Thus, empirical evidences involving real life data extracted from review of several scientific journal indicates that these assumptions are not usually met Blanca et al. (2017). Dinesh and Padmini (2015) averred that existing statistical literature and empirical studies revealed that many of the past experiment conducted in different parts of the globe suffers from the problem non-normality and heterogeneous error distribution variances due to the presence of outliers.

Corresponding Author**:** Odior A.K., Email: odifullness@gmail.com, Tel: +234

1.1 **Outliers**

One common feature of many real life data particularly in designed experiments is that they contain observations, which are inconsistent and significantly far away from the remaining dataset Bruno and Roberta (2007). These observations are technically called outliers. Outliers are observations that do not follow the pattern of other observations in a given data set. Barnett and Lewis (1993), Staudte and Sheather (1990) submitted that in the context of designed experiment, outliers are observations, which may responsible for the disruption of the usual pattern of the data. According to Dinesh and Padmini (2015) in designed experiment outliers in the dataset control the significance of the treatment effects, thus conclusion drawn from such experiment will be wrong and misleading. Therefore, the presence of outliers in designed experiment significantly influences the parameter estimates of the ANOVA model resulting in wrong statistical inferences. The presence of outliers is an indication of weakness in the model, the data or both. Unfortunately, in real life situation off center (skewed) data, heavy-tailed distributions and outliers are common problems encounter in data collection and analysis considering their regularity in many scientific and methodical studies. Avi (2006) reported that most real life data analysis these assumptions are not met due to distributional problems occasion by the presence of outliers.

Markus (2016) reported in his simulated study that parameter estimates in ANOVA model are sensitive to outliers using the classical procedures. Dinesh and Padmini (2015) examine the efficacy of robust ANOVA methods in the analysis of horticultural field experimental data in the presence of outliers. The results obtained fortify the use of robust ANOVA methods, as there was substantial reduction in error mean square when compared with the classical approach.

Consequently, robust regression methodology including the class of M estimators was introduced to reduce the effect of outliers and enhance accuracy of model parameter estimates using the tuning constant (fixed cutoff point). The methodology advocated mitigates the effect of outliers by minimizing the residual function through appropriate weighting method. However, the fixed cutoff applicable in the existing estimators is major drawback.

2 Methodology

Several robust estimation methods have been developed from diverse viewpoint, each of which has various justifications. In this research, our aim is to develop an alternative robust estimator called Adaptive Robust M Estimator (ARME) with data dependent cutoff point. However, in the presence of outliers in the dataset the LSE is easily affected because all the observations including outlying observations are assigned the weight of one. In response to the limitations of the LSE, Huber (1964) introduced the class of M estimation method. The method minimizes the sum of the objective function of the residuals to obtain the parameter estimates through appropriate weights function.

2.1 Huber M FIXED CUTOFF (HMFC)

Huber weighted function is defines as:

$$\rho(r) = \begin{cases} \frac{r^2}{2} & if\ |r| \le C_H \\ C_H|r| - \frac{1}{2}C_H^2 & if\ |r| > C_H \end{cases}$$

Where C_H is 1.345 (Huber Tuning constant)

$$\varphi(r) = \rho^I(r) = \begin{cases} 1 & if\ |r| \le C_H \\ \frac{C_H}{r|} & if\ |r| > C_H \end{cases}$$

$$W(r) = \frac{\varphi(r)}{r} = \begin{cases} if\ |r| \le C_H \\ f\ |r| > C_H \end{cases} \qquad (2.1)$$

2.3 **BISQUARE M FIXED CUTOFFT (BMFC)**

$$\rho(r) = \begin{cases} \frac{r^2}{2} - \frac{r^4}{2C_B^2} + \frac{r^6}{6C_B^2}\ for\ |r| \le C_B \\ \frac{C_B^2}{6}\ for\ |r| > C_B,\ C_B = 4.685 \end{cases}$$

$$\psi(r) = \rho^1(r) = \begin{cases} r - \frac{2r^3}{C_B^2} + \frac{r^5}{C_B^4}\ for|r| \le C_B \\ 0\ for|r| > C_B \end{cases}$$

Where C_B is the tuning constant

$$W(r) = \frac{\varphi(r)}{r} = \begin{cases} 1 - \frac{2r^2}{C_B^2} + \frac{r^4}{C_B^4} \; for \; |r| \le C_B \\ 0 \quad for \; |r| > \; C_B \end{cases}$$

$$W(r) = \begin{cases} [1 - \left(\frac{r}{C_B}\right)^2]^2 \; for \; |r| \le C_B \\ 0 \quad for \; |r| > \; C_B \end{cases} \quad (2.2)$$

2.3 Derivation of M Estimator for the Proposed Procedure

$y_i = X_i^t\beta + e_i$ (3.21)

$e_i = y_i - X_i^t\beta$

The standardized residual is defines as:

$\frac{e_i}{\hat{\sigma}} = (\frac{y_i - X_i^t\beta}{\hat{\sigma}})$

M estimator minimized

$Min_\beta = \left\{\sum_{i=1}^{n} \rho(\frac{y_i - X_i^t\beta}{\hat{\sigma}})\right\}$ (2.3)

To minimize, differentiate (3.22) with respect to β and equate to zero

Let $U = (\frac{y_i - X_i^t\beta}{\hat{\sigma}})$

$\frac{du}{d\beta} = (\frac{-X_i}{\hat{\sigma}})$

For $\rho(U) \rightarrow \frac{d[\rho(u)]}{du} = \varphi(u)$ (2.4)

Using function of function

$\frac{d[\rho(u)]}{du} \times \frac{du}{d\beta} \rightarrow \frac{-1}{\hat{\sigma}} \sum \varphi\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right) X_i = 0$ (3.24)

$\sum \varphi\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right) X_i = 0$

$\sum \varphi\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right) \frac{\left(y_i - X_i^t\hat{\beta}\right)}{\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right)} X_i = 0$

$\sum \varphi \frac{\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right)}{\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right)} \left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right) X_i = 0$ (2.5)

If

$W_i = \frac{\varphi\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right)}{\left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right)}$

$\sum W_i \left(\frac{y_i - X_i^t\hat{\beta}}{\hat{\sigma}}\right) X_i = 0$

$\frac{1}{\hat{\sigma}} \sum W_i \left(y_i - X_i^t\hat{\beta}\right) X_i = 0$

$\sum W_i \left(y - X_i^t\hat{\beta}\right) X_i = 0$

$\sum X_i W_i y_i - X_i^t W_i X_i \hat{\beta} = 0$

Hence by matrix notation:

$XWy - XWX\hat{\beta} = 0$

$XWX\hat{\beta} = XWy$

$\hat{\beta} = (XWy)^{-1}XWy$ (2.6)

2.4 Data Dependent Cut-Off Point Using Robust Confidence Intervals

In order to achieve high efficiency in relation to the M estimator, we propose the use of robust confidence interval to determine the cut-off point that is data dependent. Carling (2000) recommended the use of median rule (box plot) given as:

$$C^L = q_2 - k_2(q_3 - q_1)$$
$$C^U = q_2 + k_2(q_3 - q_1) \quad (2.7)$$

to determine the robust confidence interval. Where C^L is the lower cut-off point while C^U, $q_{2,}$ is the sample median, $q_{1,}$ is the lower quartile and $q_{3,}$ is the upper quartile. The default value of $k_{2,}$ is 2.3.

2.5 Algorithm for obtaining proposed alternative estimator

Given a vector of responses Y and vectors of treatment codes for the appropriate treatment contrasts, to find a robust ANOVA using M-estimator with Huber weight:

1. Obtain the initial estimate of the parameters of the model using the Least Squares Estimator (LSE).
2. Determine the residuals of the LSE model and denote it by r.
3. Calculate the robust estimate $(\hat{\sigma})$ of the standard deviation of the residuals r using the median of the absolute deviation.
4. Determine the standardized residuals $u = \frac{r}{\hat{\sigma}}$.
5. Find the robust Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL) of the standardized residuals using (2.7).
6. Set $k = 0$ and perform the following iterations:
   - (i) Use Huber weight function given in (2.1) with data dependent cut-off point UCL to determine the appropriate weight (W) for each standardized residual.
   - (ii) Use Weighted Least Squares Estimator (WLSE) with the weight W to obtain a new estimate of the parameters of the model.
   - (iii) Determine the residuals r, the robust estimate $(\hat{\sigma})$ of the standard deviation of the residuals and the standardized residuals $u = \frac{r}{\hat{\sigma}}$ of the new model.
   - (iv) Calculate the robust LCL and UCL for each of the new standardized residual of the WLSE.
   - (v) Determine the maximum absolute difference (D) between the pairs of newly estimated parameters and the estimated parameters of the immediate preceding model.
   - (vi) If either $k = maxiter$ (maximum number of iterations) or $D < \varepsilon$, where ε is the tolerance limit, then set the new estimate as the final estimate of the model and used the model to obtain robust ANOVA estimates.
   - (vii) Else, set $k = k + 1$

3.0 DATA ANALYSIS

3.1 Dataset Simulation Design

Monte Carlo simulation was used to examine the robustness and efficiency of the proposed estimator. We performed experiment involving one-way ANOVA model. In the one-way experiment, four treatment levels were crossed with 4-sample size (20, 60,100 and 200). To examine the effect of outliers, five levels of data contaminations (10%, 20%, 30% and 40%) were applied. The assumed models for one-way ANOVA is y_{ij}=μ+$\tau_i + e_{ij}$, i= *1,2,3,4* and *j=1,2,...,n*, where n is the sample size, τ_j is the treatment effect *e* is the error term. The error terms for non-outlying observations were generated from Standard Normal distributions, while the error terms for outlying observation were generated from Cauchy distribution with location and scale parameter 0 and 5 respectively. All simulation programs were developed using R statistical programming language (R core Team, 2018). The function *lm* in the base package was used to obtain the estimates of regression parameters for the OLS estimator of the classical ANOVA method while the algorithm development for robust ANOVA M-estimator was used for proposed estimation method. Each simulation case was replicated M=1000 times. The estimates of each estimator were calculated in each iteration and the goodness of fit of each estimator was determined using Root Mean Square Error (RMSE) defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}(y_i - \hat{y})^2}{M}}$$

Where y_i is the actual observation and $\hat{Y}$ is the fitted (estimated) value by the estimator

The mean and standard deviation of the 1000 replicated RMSEs were used as the point estimate and the measure of deviation respectively.

The estimator with the lowest mean RMSE is the most efficient, the smaller the RMSE the more efficient is the estimator. Also, the smaller the standard deviation of the RMSEs the more consistence is the estimator.

3.2 Simulation Setup

3.2.1 One-way ANOVA

The model for the analysis of One-way ANOVA can be specified as:

$$Y_{ij=\,\mu+\,\tau_{i\,+\,\varepsilon_{ij}}} \quad (3.1)$$

μ is a parameter common to all treatments called the overall mean

τ_i is the parameter unique to treatment called the ith treatment effect,

ε_{ij} is the random error component that is incorporates all other source of variability.

Thus, the parameter of interest and to be estimated is the common mean (μ) and the treatment effect (τ_i). The reasonable concern in this study is the effect of outliers on the parameter estimates, on which the inferences is based.

3.2.2 without Outliers

The observations were randomly generated from the standard normal distribution $\varepsilon \sim N(0,1)$. The data generated in the simulation without outliers will serve as a reference data (clean) in the study.

3.2.3 With Outliers

To generate a certain percentages of outliers using the contaminated normal mixture with $\varepsilon_{ij} \sim N(0,1) + Cauchy(0,5)$. The clean or reference observation is substituted with 10, 20, 30, and 40 percentages of outliers respectively.

3.3 Criteria for Evaluating the Estimator Performance

a. Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}(y_i - \hat{y})^2}{M}}$$

Where y_i is the actual observation and $\hat{Y}$ is the fitted (estimated) value by the estimator

One-Way ANOVA

Table 4.1: The Mean RMSE and Standard deviation values for different estimation method

Sample size	Outliers	Criteria	Estimators				
			LSE	*HMFC*	*BMFC*	*HMDD*	*BMDD*
n=20	0%	*Mean RMSE*	**0.1534**	0.1575	0.1719	0.1534	0.1711
		Std Dev RMSE	**0.8862**	0.9049	0.9486	0.8868	0.8608
n=60	0%	*Mean RMSE*	**0.0915**	0.0921	0.0937	0.0915	0.0964
		Std Dev RSME	**0.9637**	0.9686	0.9763	0.9638	0.9863
n=100	0%	*Mean RMSE*	**0.0732**	0.0736	0.0743	0.0732	0.0753
		Std Dev RMSE	**0.9765**	0.9792	0.9827	0.9765	0.9881
n=200	0%	*Mean RMSE*	**0.0520**	0.0520	0.0522	0.0520	0.0523
		Std Dev RMSE	**0.9900**	0.9914	0.9928	0.9901	0.9956

The performance of LSE is similar to that of HMFC, BMFC, HMDD and BMDD when the data is free from outliers. The separate results obtained for each estimator, the RMSE therefore revealed an infinitesimal difference in their estimates. This implies that all the estimators are suitable for the data.

Table 4.2: The mean RMSE and Standard deviation values for different estimation Method

Sample size	Outliers	Criteria	Estimators				
			LSE	*HMFC*	*BMFC*	*HMDD*	*BMDD*
n=20	10%	*Mean RMSE*	2.0239	0.1853	0.2009	0.1111	**0.1039**
		Std Dev RMSE	2.1712	0.9583	0.9726	0.0471	**0.0765**
n=60	10%	*Mean RMSE*	2.0930	0.1032	0.1003	0.1512	**0.1013**
		Std Dev RSME	2.0025	0.9852	0.9798	0.0086	**0.8847**
n=100	10%	*Mean RMSE*	2.0741	6.0764	0.0756	0.0861	**0.0762**
		Std Dev RMSE	2.6495	0.9901	0.9878	0.0043	**0.0908**
n=200	10%	*Mean RMSE*	2.5266	0.0516	0.0516	0.0543	**0.0517**
		Std Dev RMSE	2.8157	0.9929	0.9915	0.9095	**0.2931**

The performamce of LSE at 10% injection of outliers represent a breakdown of the estimator with a significant increase in the value of RMSE. However, the RMSE estimates of HMFC, BMFC, HMDD and BMDD show that BMDD outperformed the other estimators with the least value of RMSE.

Table 4.3: The mean RMSE and Standard deviation values for different estimation Method

Sample size	Outliers	Criteria	Estimators				
			LSE	*HMFC*	*BMFC*	*HMDD*	*BMDD*
n=20	20%	*Mean RMSE*	2.7539	0.8458	0.8617	0.4074	**0.3946**
		Std Dev RMSE	2.7339	0.5160	0.5181	0.6527	**0.5195**
n=60	20%	*Mean RMSE*	2.8363	0.1686	0.1879	0.2346	**0.1325**
		Std Dev RSME	3.8614	1.0405	1.0168	1.1302	**1.0121**
n=100	20%	*Mean RMSE*	2.9620	0.0959	0.0879	0.1403	**0.0821**
		Std Dev RMSE	2.8213	1.0159	1.0023	1.0667	**1.0014**
n=200	20%	*Mean RMSE*	2.7560	0.0559	0.0537	0.0747	**0.0537**
		Std Dev RMSE	3.3151	1.0081	0.9987	1.0326	**0.9989**

At 20% contamination of the data, the RMSE value of LSE increased an indication that the estimator is easily affected by outlying observations. Similarly, the RMSE values of HMFC, BMFC, HMDD and BMDD suggest that the estimators are applicable to the data with reasonably low values of RMSE. However, BMDD display superior efficiency with the smallest value of RMSE.

Table 4.4: The mean RMSE and Standard deviation values for different estimation Method

Sample size	Outliers	Criteria	Estimators				
			LSE	*HMFC*	*BMFC*	*HMDD*	*BMDD*
n=20	30%	*Mean RMSE*	2.9017	1.1594	1.2100	1.0272	**0.1122**
		Std Dev RMSE	3.2428	1.5080	1.4832	2.2267	**0.4323**
n=60	30%	*Mean RMSE*	2.9217	0.1634	0.1641	0.3665	**0.1072**
		Std Dev RSME	3.0523	1.0685	1.0320	1.2917	**1.0021**
n=100	30%	*Mean RMSE*	2.9576	0.1080	0.0.845	0.2431	**0.0843**
		Std Dev RMSE	2.6942	1.0433	1.0146	1.1836	**1.0161**
n=200	30%	*Mean RMSE*	2.9597	0.0595	0.0546	0.1068	**0.0547**
		Std Dev RMSE	3.2630	1.0191	1.0071	1.0135	**1.0078**

At 30% injection of outliers, LSE experienced a steady increase in the value of RMSE, an evidence and strong proof of the impact of outliers on the estimator. The results revealed dwindling performance of LSE with an increase in the percentage of outliers. In addition, the RMSE values of HMFC, BMFC, HMDD and BMDD indicates that BMDD performed better than other estimators with the least value of RMSE. This is because BMDD offered more protection against the effect of outliers, hence higher accuracy in parameter estimation was achieved

Table 4.5: The mean RMSE and Standard deviation values for different estimation Method

Sample size	Outliers	Criteria	Estimators				
			LSE	*HMFC*	*BMFC*	*HMDD*	*BMDD*
n=20	40%	*Mean RMSE*	3.1600	1.1600	1.1278	1.0812	**1.0153**
		Std Dev RMSE	3.9883	0.6280	0.6091	0.0598	**0.6263**
n=60	40%	*Mean RMSE*	3.3329	0.2141	0.2511	0.1944	**0.1247**
		Std Dev RSME	3.8586	1.1212	1.0758	0.5569	**0.1111**
n=100	40%	*Mean RMSE*	3.5607	0.1214	0.0925	0.0277	**0.1078**
		Std Dev RMSE	3.9035	1.0675	1.0289	1.0021	**1.0049**
n=200	40%	*Mean RMSE*	3.9290	0.0659	0.0579	0.0175	**0.0602**
		Std Dev RMSE	3.1463	1.0326	1.0137	0.1894	**0.0254**

At 40% injection of outliers on the response variable, the performance of LSE was very poor with a continuous increase in the value of RMSE. This suggests that LSE is sensitive to outliers. Further, RMSE values of HMFC, BMFC, HMDD and BMDD suggest that BMDD yields smaller value of RMSE than other estimators' do.Thus, it is evident that BMDD yields superior efficiency.

4 Comparison of Methods for One-Way ANOVA

The simulation results presented in Tables 4.1, 4.2, 4.3, 4.4 and 4.5 revealed the mean RMSE and standard deviation values for different estimation methods respectively with sample size n=20, 60, 100, and 200 crossed with different percentage of outliers 10, 20, 30 and40. The mean RMSE is the point estimate of the 1000 replicated RMSE and Standard deviation measures the deviation of the 1000 replicated estimates.

From Table 4.1, it was evident that LSE retained its optimal property of minimum variance in a clean data (outlier free) with smallest value of RMSE and standard deviation. Table 4.1 also indicated that, in general small values of RMSE and standard deviation for the existing robust methods and proposed robust method (ARME) are consistent with existing statistical literature on the adequacy of robust estimator in both outliers-filled data and when the data is free from outliers. Thus, LSE, existing robust methods and the proposed method were computationally satisfactory and equivalent in a clean data.

Also, Tables 4.2, 4.3, 4.4 and 4.5 give separate results for each selected estimation methods using different percentage of outliers for all sample size. The result revealed that in the presence of outliers, LSE becomes increasingly inefficient and unfit as the percentage outliers increases in the simulation. This suggests that LSE is sensitive and easily affected by outliers and not of good fit for data that are contaminated with outliers. Thus, it failed to retain its optimal property of minimum variance when face with dataset that are contaminated with outliers.

Furthermore, considering Table 4.1, 4.2, 4.3, 4.4 and 4.5 for the one-way ANOVA, comparing the performance of HMFC, BMFC, HMDD and BMDD, the proposed method BMDD has the least RMSE acrossed all the sample size used hence offered superior efficiency. However, HMFC, BMFC and HMDD have similar low RMSE but they are not as efficient as BMDD. The results presented indicate that increasing the sample size consistently improves the performance of HMDD and BMDD estimators in terms of substantial reduction in the value of RMSE. The results indicate (HMDD) and (BMDD) offer substantial improvement in efficiency and goodness of fit when compared with the existing method judging from the estimated values of RMSE and standard deviation.

It was also observed from the results that when the proportion of outliers is increased to 10, 20, 30 and 40 percentage the proposed method performed better than the existing methods considered, as it offer smallest values of RMSE and standard deviation. In general, for y direction outliers in the estimation of ANOVA parameters the proposed BMDD surpassed the performance of the existing methods (HMFC, BMFC), using efficiency as a measure of robustness.

5. Conclusion

Comparing the five estimators in this study from different sample size and percentages of outliers using RMSE as a measure of efficiency, the alternative estimator (ARME) is more insensitive to both the percentages and magnitude of outliers and the sample size. Generally, the proposed alternative method performed better than LSE, HMFC, BMFC methods. Thus, BMDD performance is considerably better than the HMDD.

Appendix A

ROBUST ANOVA1

```
############################################################
# RobustANOVA1 One-way, Case 1.0, n = 20, Outlier = 0%   #
############################################################
source("C:/Odior/Robust ANOVA/Odior/Outbox function.R")
sink("RobustANOVA1Case1.0.txt")
library(broom)
library(robustbase)
set.seed(13)
M <- 1000 # Monte Carlo Sample. Should be 1000
ng <- 5 # Sample size per group. Total sample size (N) = ng * Number of groups
Miu_A <- 10 # Factor A effect
Miu_B <- 10 # Factor B effect
Miu_C <- 10 # Factor C effect
Miu_D <- 10 # Factor D effect
Sigma_A <- 1
Sigma_B <- 1
Sigma_C <- 1
Sigma_D <- 1

RMSELSE <- numeric(M)
RMSEHuberMFC <- numeric(M)
RMSEBisquareMFC <- numeric(M)
RMSEHuberMDD <- numeric(M)
RMSEBisquareMDD <- numeric(M)
PValueLSE <- numeric(M)
PValueHuberMFC <- numeric(M)
PValueBisquareMFC <- numeric(M)
PValueHuberMDD <- numeric(M)
PValueBisquareMDD <- numeric(M)

for(m in 1:M){                                  # Start the loop

  sim_data = data.frame(
    response = c(rnorm(n = ng, mean = Miu_A, sd = Sigma_A),
                 rnorm(n = ng, mean = Miu_B, sd = Sigma_B),
                 rnorm(n = ng, mean = Miu_C, sd = Sigma_C),
                 rnorm(n = ng, mean = Miu_D, sd = Sigma_D)),
    group = c(rep("A", times = ng), rep("B", times = ng),
              rep("C", times = ng), rep("D", times = ng))
  )
  Y <- sim_data$response
  X2 <- c(rep(0,ng),rep(1,ng),rep(0,ng),rep(0,ng))
  X3 <- c(rep(0,ng),rep(0,ng),rep(1,ng),rep(0,ng))
  X4 <- c(rep(0,ng),rep(0,ng),rep(0,ng),rep(1,ng))

  # Least Squares Method Using Cell Reference Method
  ModLSEMod <- lm(Y ~ X2 + X3 + X4)
  ResLSE <- residuals(ModLSEMod)
  RMSELSE[m] <- sqrt(mean(ResLSE^2))
  PValueLSE[m] <- glance(ModLSEMod)$p.value
print(ResMat)

LSEPower.10 <- round((length(PValueLSE[PValueLSE > 0.10])/M),4)
LSEPower.05 <- round((length(PValueLSE[PValueLSE > 0.05])/M),4)
LSEPower.01 <- round((length(PValueLSE[PValueLSE > 0.01])/M),4)
```

```
HuberMFCPower.10 <- round((length(PValueHuberMFC[PValueHuberMFC > 0.10])/M),4)
HuberMFCPower.05 <- round((length(PValueHuberMFC[PValueHuberMFC > 0.05])/M),4)
HuberMFCPower.01 <- round((length(PValueHuberMFC[PValueHuberMFC > 0.01])/M),4)
BisquareMFCPower.10 <- round((length(PValueBisquareMFC[PValueBisquareMFC > 0.10])/M),4)
BisquareMFCPower.05 <- round((length(PValueBisquareMFC[PValueBisquareMFC > 0.05])/M),4)
BisquareMFCPower.01 <- round((length(PValueBisquareMFC[PValueBisquareMFC > 0.01])/M),4)
HuberMDDPower.10 <- round((length(PValueHuberMDD[PValueHuberMDD > 0.10])/M),4)
HuberMDDPower.05 <- round((length(PValueHuberMDD[PValueHuberMDD > 0.05])/M),4)
HuberMDDPower.01 <- round((length(PValueHuberMDD[PValueHuberMDD > 0.01])/M),4)
BisquareMDDPower.10 <- round((length(PValueBisquareMDD[PValueBisquareMDD > 0.10])/M),4)
BisquareMDDPower.05 <- round((length(PValueBisquareMDD[PValueBisquareMDD > 0.05])/M),4)
BisquareMDDPower.01 <- round((length(PValueBisquareMDD[PValueBisquareMDD > 0.01])/M),4)

MatPower <- cbind(PValueLSE,PValueHuberMFC,PValueBisquareMFC,PValueHuberMDD,PValueBisquareMDD)
print("")
print(MatPower)
VecAlpha10 <- c(LSEPower.10,HuberMFCPower.10,BisquareMFCPower.10,HuberMDDPower.10,BisquareMDDPower.10)
VecAlpha5 <- c(LSEPower.05,HuberMFCPower.05,BisquareMFCPower.05,HuberMDDPower.05,BisquareMDDPower.05)
VecAlpha1 <- c(LSEPower.01,HuberMFCPower.01,BisquareMFCPower.01,HuberMDDPower.01,BisquareMDDPower.01)
Vec2 <- c(VecAlpha10,VecAlpha5,VecAlpha1)
ResTab2 <- as.table(matrix(Vec2, nrow=3, byrow=TRUE, dimnames=list(Alpha= c("10%","5%","1%"),Estimator = EstVec1)))
print("")
print(ResTab2)print(m)sink()
```

References

[1] Avi, G. (2006). Robust analysis of variance. Process design and quality improvement. *International Journal of Productivity and Quality Management, 1(3) 1-14.*

[2] Barnett, V., and Lewis, T. (1993). Outliers in statistical data, third ed. Wiley, New -York.

[3] Blanca, M. J., Arnau, J., and Rebecca, B. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema,* 29(4) 552-557.

[4] Bruno, B. and Roberta, V. (2007). Robust analysis of variance: An approach based on the forward search. *Computational Statistics and Data Analysis, 51(7), 5172 – 5183.*

[5] Carling, K. (2000). Resistant outlier rules and non –Gaussian case. *Computational Statistics and Data Analysis*, 33, 249-258

[6] Dinesh, I. R. and Padmini, V. P. (2015). Robust ANOVA: An illustrative study in horticultural crop research, *International Journal of Mathematics and Computational Sciences,* 9(2), 85-89

[7] Huber, J. P. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 73-101.

[8] Kevin, D. B. (2004). Analysis of Variance via Confidence Intervals. Sage Publications, London.

[9] Markus, H. (2016). ANOVA-The effect of outliers. A thesis submitted to the department of Statistics Uppsala University, Uppsala, Sweden.

[10] Montgomery, (2010). Design and analysis of experiments. John Wiley and sons, New -York.

[11] Staudtle, R. G. and Sheather, S. J. (1990). Robust Estimation and Testing. Wiley, New- York.

[12] Wayne, W. D & Chad, L. C. (2013). Biostatistics: A Foundation for Analysis in the Health Sciences. Wiley, New-York.