AP STATISTIC FOR IDENTIFICATION OF OUTLIERS IN MULTI-RESPONSE EXPERIMENTS WITH CORRELATED ERRORS

ABSTRACT

In multi response experiments where correlated errors are present, identification of outliers becomes a complex yet crucial task. Outliers not only distort the estimation of model parameters but also jeopardize the validity of statistical inferences drawn from the data. In the present study the statistic given by Andrews and Pregibon (AP) (1978) for detection of influential observations in linear regression is suitably modified for detection of outliers in multi – response experiments with correlated errors. The statistic was developed considering the data structure of auto – regressive order 1. The developed statistic is successful in detection of outliers when tested upon simulated datasets.

Keywords: Multi – response experiments, Andrews and Pregibon statistic, Correlated errors, Outliers, Influential observations.

1. INTRODUCTION

Identifying outliers, or data points that significantly deviate from others in a dataset, is crucial in ensuring the accuracy and reliability of conclusions drawn from multi-response experiments. These experiments, often conducted in block design setups, can include observations across various response variables, where a single outlier can potentially skew results dramatically.In the realm of multi-response experiments, outliers are not merely anomalies but pivotal points that can significantly influence the outcome and interpretation of data. In the intricate realm of multi-response experiments, the interplay of correlated errors emerges as a pivotal factor influencing the reliability and interpretation of experimental outcomes. These errors, inherently intertwined, demand a nuanced approach for accurate analysis.By delving into the intricacies of correlated errors, reveals their critical role in shaping the accuracy and reliability of multi-response experiments.

In the convolutedlandscape of multi-response experiments, outliers play a pivotal role that extends beyond mere anomalies within the dataset. Their presence and handling can significantly influence the outcome of statistical analyses, making their identification and treatment a critical aspect of research methodology.

a. **Impact on Parameter Estimation**: Outliers can cause departures from the assumptions of parameter estimation, leading to misleading results. This deviation is not trivial, as parameter estimation forms the backbone of many statistical analyses, influencing everything from hypothesis testing to predictive modeling .
b. **Influence on Statistical Analyses**: The presence of outliers can skew the results of hypothesis tests, potentially leading to incorrect conclusions. This skewness is attributed to the extreme values of outliers, which can disproportionately affect the mean, variance, and other statistical measures of a dataset .

When it comes to dealing with outliers in multi-response experiments, the approach is nuanced, emphasizing the retention of outliers as much as possible. This strategy is underpinned by the recognition that outliers, unless clearly erroneous, hold valuable information about the dataset and the phenomena under study.

1.1 Dealing with Outliers:

i. **Retention Over Removal**: It's recommended to retain outliers in the dataset unless there is clear evidence that they represent errors. This approach acknowledges the potential value of outliers in uncovering unexpected patterns or errors in the experimental setup .
ii. **Thorough Investigation**: Before deciding on the treatment of an outlier, a thorough investigation is necessary to determine its cause. This might involve examining the experimental conditions, data collection processes, or even the possibility of transcription errors.
iii. **Contextual Decision-Making**: The decision to retain or remove an outlier should be made in the context of the specific experiment and its objectives. The impact of outliers on the analysis should be weighed against the potential loss of valuable information if they were to be excluded .

This nuanced approach to handling outliers underscores the complexity of outlier management in multi-response experiments. By prioritizing retention and careful examination, researchers have to navigate through the challenges posed by outliers, ensuring that their analyses remain robust and their conclusions valid.

2. REVIEW OF LITERATURE:

L M Bhar (1997) worked on identification of outliers in design of experiments. The distances considered were Cook's distance, AP statistic and Q_k statistic for estimation of outliers [**Error! Reference source not found.**]. He has appropriately modified the above distances

for its applicability to design of experiments where interest of experimenter lies in estimation of treatment effects only. The designs considered by him were single response experiments only.

Later, S Ojha (2015) extended the work of LM Bhar to design of experiments with correlated errors case [**Error! Reference source not found.**]. They studied the correlated errors case with two different correlation structures - autocorrelated errors case and equi-correlation structure and suitably modified AP statistic and Cook's statistic for identification of outliers with correlated errors.

For the present paper we will be considering AP statistic for extension to multi response experiments with correlated errors. In order to identify influential observations, Andrews and Pregibon (1978) have developed a statistic that utilizes the concept of data point remoteness in factor space [0]. This measure can effectively detect influential observations as a highly remote point can significantly impact parameter estimation. The extension of AP statistic to design of experiments with correlated errors case covers the subjects like comprehensive analysis of the data by providing a more nuanced understanding of experimental treatment effects and detect influential observations while appropriately considering the correlation structure.

Berenblut (1974) considered the problem of possible autocorrelations between the observations in change over designs which involves minimization of generalized variance of parameter estimates [0]. He produced a design for optimum settings of quantitative factors under mild conditions in block designs and time sequences.

3. AP STATISTIC

Consider a general linear model,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \#(1)$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observations, $\mathbf{X}$ is $n \times m$ full rank design matrix, $\boldsymbol{\beta}$is$m \times 1$ vector of unknown parameters and $\boldsymbol{\varepsilon}$is$n \times 1$ vector of independent random variables.

Assuming that *t* (known) observations being outliers, indicating that their expected values are shifted from the expected value of other observations. Now mean – shift model for *t* outlying observations can be written as,

$$\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \quad \#(2)$$

where $\boldsymbol{Z} = (\boldsymbol{X} \quad \boldsymbol{U}), \boldsymbol{U} = (\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \boldsymbol{u}_3 \quad \dots \quad \boldsymbol{u}_t)$and $\boldsymbol{u}_i = (0 \quad 0 \quad 0 \quad \dots \quad 1(i^{th}) \quad \dots \quad 0)'$ is a $n \times 1$ vector where i^{th}observation is 1 indicating an outlier and $\boldsymbol{\gamma} = (\boldsymbol{\theta}' \quad \boldsymbol{\delta}')'$where $\boldsymbol{\delta} = (\boldsymbol{\delta}_1 \quad \boldsymbol{\delta}_2 \quad \dots \quad \boldsymbol{\delta}_t)'$.

The AP statistic is now defined as

$$\boldsymbol{AP} = \frac{\left|\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*}\right|}{\left|\boldsymbol{A}'\boldsymbol{A}\right|} \#(3)$$

where $\boldsymbol{A} = (\boldsymbol{X} \quad \boldsymbol{y})\ and\ \boldsymbol{Z}^{*} = (\boldsymbol{X} \quad \boldsymbol{U} \quad \boldsymbol{y})$and $|\boldsymbol{A}|$ denotes the determinant value of $\boldsymbol{A}$.

It's been suggested that the value of $(1 - AP_t)$ reflects the volume proportion associated with $\boldsymbol{A}$attributes to the *t* outlying observations. When these particular data points are situated far from the core in factor space, they significantly contribute to the overall volume, thereby providing a more meaningful interpretation of outliers. Hence, small values of AP statistic are associated with influential observations [**Error! Reference source not found.**].

This statistic cannot be directly applied to design matrices because of the rank deficiency problem. Hence the statistic has to be modified for the application to the design matrices.

3.1 AP Statistic for Design of Experiments

Consider a design model

$$\boldsymbol{y}^{*} = \boldsymbol{X}^{*}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^{*} \#(4)$$

where $\boldsymbol{y}^{*}$ is $n \times 1$, vector of observations, $\boldsymbol{X}^{*}(n \times m)$ is not of full rank. The rank of $\boldsymbol{X}^{*}$ is *w* (which is < m) is known. So, we reparametrize the model in the following way.

Since the rank of $\boldsymbol{X}^{*}$ is *w* there are almost *w* linearly independent estimable parametric functions and we let $\boldsymbol{P}^{*}\boldsymbol{\theta}$be such independent estimable parametric functions. Hence every estimable linear parametric function must be a part of elements of $\boldsymbol{P}^{*}\boldsymbol{\theta}$.Hence $\boldsymbol{X}^{*}\boldsymbol{\theta}$which is estimable can be written as

$$\boldsymbol{X}^{*}\boldsymbol{\theta} = \boldsymbol{M}^{*}\boldsymbol{P}^{*}\boldsymbol{\theta} \#(5)$$

for some $n \times w$ matrix of $\boldsymbol{M}^{*}$of rank *w*, where the matrix $\boldsymbol{P}$is of dimension $w \times w$ with Rank = *w* such that

$$\boldsymbol{P}^{*}\boldsymbol{P}^{*\prime} = \boldsymbol{I}_{w-1} \#(6)$$

Now the transformed matrix can be written as

$$\boldsymbol{y}^{*} = \boldsymbol{X}^{*}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^{*} \#(7)$$

$$= \boldsymbol{M}^{*}\boldsymbol{\beta}^{*} + \boldsymbol{\varepsilon}^{*} \#(8)$$

where $\boldsymbol{\beta} = \boldsymbol{P}^{*}\boldsymbol{\theta}$**.** The model now becomes a full rank model with

$$\widehat{\boldsymbol{\beta}}^{*} = (\boldsymbol{M}^{*\prime}\boldsymbol{M}^{*})^{-1}\boldsymbol{M}^{*\prime}\boldsymbol{y}^{*} \#(9)$$

3.2 AP Statistic for Multi-Response Experiments with Correlated Errors

The linear model for multi – response with correlated errors can be written as

$$\boldsymbol{Y}^{0} = \boldsymbol{X}^{0}\boldsymbol{\theta}^{0} + \boldsymbol{\varepsilon}^{0} \#(10)$$

where $\boldsymbol{Y}^0 = [\boldsymbol{y}_1^* \quad \boldsymbol{y}_2^* \quad \dots \quad \boldsymbol{y}_p^*]'$is a $np \times 1$ vector of observations, $\boldsymbol{X}^0 = \bigoplus \boldsymbol{X}^* = \boldsymbol{I}_p \otimes \boldsymbol{X}^*$ is $np \times wp$ design matrix, $\boldsymbol{\theta}^0 = [\boldsymbol{\theta}_1^{*\prime} \quad \boldsymbol{\theta}_2^{*\prime} \quad \dots \quad \boldsymbol{\theta}_p^{*\prime}]'$ is a $wp \times 1$ the matrix of parameters and $\boldsymbol{\varepsilon}^0 = [\boldsymbol{\varepsilon}_1^{*\prime} \quad \boldsymbol{\varepsilon}_2^{*\prime} \quad \dots \quad \boldsymbol{\varepsilon}_p^{*\prime}]'$is $np \times 1$ matrix of error terms. The errors are distributed with mean $\boldsymbol{E}(\boldsymbol{\varepsilon}^0) = 0$ and dispersion matrix $D(\boldsymbol{\varepsilon}^0) = \boldsymbol{\Sigma}_{pp} \otimes \boldsymbol{\Omega} = \boldsymbol{\Psi}$, where the observations belonging to same replications are assumed to be autocorrelated. It is also assumed that the autocorrelation of multiple responses belonging to same replication is also same.

$\boldsymbol{\Sigma}_{pp}$is the variance covariance matrix of p responses and $\boldsymbol{\Omega} = diag(\boldsymbol{\Omega}_1 \quad \boldsymbol{\Omega}_2 \quad \dots \quad \boldsymbol{\Omega}_b)$where $\boldsymbol{\Omega}_j$is $k_j \times k_j$ matrix for correlation structure of j^{th} block, where k_j is the number of observations in the j^{th} block.

The matrix $\boldsymbol{M}$of order$pn \times pw$ with rank $(p \times w)$ is obtained in similar way as of $\boldsymbol{M}^*$ and the matrix is similar to that of $\boldsymbol{U}$defined in model (2) containing $\boldsymbol{u}_i$ vector of order $pn \times 1$ with all zeros and i^{th} observation being 1.

Now we define the following matrices

$$\boldsymbol{X}_0 = [\boldsymbol{M} \quad \boldsymbol{Y}^0] \; and \; \boldsymbol{Z}_0 = [\boldsymbol{M} \quad \boldsymbol{U} \quad \boldsymbol{Y}^0] \; \#(11)$$

where $\boldsymbol{M} = \boldsymbol{X}^0\boldsymbol{P}'(\boldsymbol{P}\boldsymbol{P}')^{-1}$

The AP statistic for multi response experiments with correlated errors can be written as

$$\boldsymbol{AP}_t = \frac{|\boldsymbol{Z}_0'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}_0|}{|\boldsymbol{X}_0'\boldsymbol{\Psi}^{-1}\boldsymbol{X}_0|} \; \#(12)$$

where,

$$|\boldsymbol{Z}_0'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}_0| = \begin{vmatrix} \boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M} & \boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} & \boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \\ \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{M} & \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} & \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{M} & \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} & \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \end{vmatrix}$$

$$= |\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M}| \left| \begin{pmatrix} \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} & \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} & \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{y} \end{pmatrix} - \begin{pmatrix} \boldsymbol{U}'\boldsymbol{\Psi}^{-1}\boldsymbol{M} \\ \boldsymbol{y}'\boldsymbol{\Psi}^{-1}\boldsymbol{M} \end{pmatrix} (\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M})^{-1} (\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{U} \quad \boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}) \right|$$

$$= |\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M}| \left| \begin{pmatrix} \boldsymbol{U}'\boldsymbol{H}\boldsymbol{U} & \boldsymbol{U}'\boldsymbol{H}\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{H}\boldsymbol{U} & \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} \end{pmatrix} \right|$$

$$= |\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M}| \left| \begin{pmatrix} \boldsymbol{U}'\boldsymbol{H}\boldsymbol{U} & \boldsymbol{U}'\boldsymbol{r}_0 \\ \boldsymbol{r}_0'\boldsymbol{U} & \boldsymbol{RSS} \end{pmatrix} \right|$$

where $\boldsymbol{H} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Psi}^{-1}$, $\boldsymbol{H}\boldsymbol{y} = \boldsymbol{r}_0$is the vector of residuals and $\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = \boldsymbol{RSS}$.

On similar process we can express $|\boldsymbol{X}_0'\boldsymbol{\Psi}^{-1}\boldsymbol{X}_0| = |\boldsymbol{M}'\boldsymbol{\Psi}^{-1}\boldsymbol{M}|(\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y})$

Thus

$$AP_t = |U'HU|\left(1 - \frac{r_2^{*\prime}(U'HU)^{-1}r_2^*}{RSS}\right) \#(13)$$

$r_2^*$is the residual corresponding to outlying observations.

Before applying to the real-world data, the developed methodology is tested over simulated data.

4. EXAMPLE:

Experimental data of autoregressive order 1 was simulated with auto-correlation values 0.39, 0.45 and -0.59 for 6 treatments and number of replications being 3. Assuming that the experiment was conducted in randomized block design and two responses were collected from the experiment conducted. The data obtained is as follows

Table 1: Simulated Multi – response experimental data with correlated errors

Treat-ments	Replication	Response 1	Response 2	Treat-ments	Replication	Response 1	Response 2
1	1	1.42	-0.88	4	2	6.28	13.56
2	1	25.37	33.97	5	2	9.06	20.02
3	1	5.34	1.38	6	2	18.04	13.25
4	1	26.24	10.84	1	3	-6.40	-2.18
5	1	16.02	-0.68	2	3	-8.42	-0.62
6	1	24.88	14.58	3	3	9.25	11.40
1	2	-0.41	3.01	4	3	22.85	28.27
2	2	-6.47	-3.93	5	3	10.14	17.52
3	2	-5.70	10.02	6	3	7.12	8.07

Before going for testing with the developed statistic, it is advisable to ensure that the underlying assumptions are valid. In the current example, the data was simulated based on the specified assumptions, thus confirming their satisfaction. Subsequently the data was tested over the developed statistic for identification of influential observations. For easy analysis code for the developed statistic was written in R – software and used for analysing the given data. The results thus obtained were presented in the Fig.1.

It was noted that 20th observation exhibited the lowest AP statistic value of 0.03 among all the observations, indicating its potential as a highly influential observation within the dataset. However, it is evident that the observations exert similar levels of influence overall, as noticed by the relatively insignificant change observed upon the removal of the 20th observation.

For further illustration we replace the 1st observation of second response with a very high value (*i.e.*, 100) and obtain the same statistics. The results thus obtained were presented in Fig. 2.

Upon the examination of AP statistic values, it becomes apparent that the 19th observation, which has been altered, exhibits the lowest AP statistic value. The observation stands out as highly influential within the dataset. Consequently, it is evident that AP statistic serves as a valuable tool for identifying outlying observations in multi – response experiments with correlated errors.

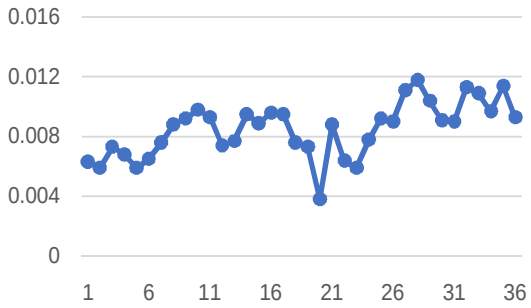

Fig. 1: AP statistic values of Table 1

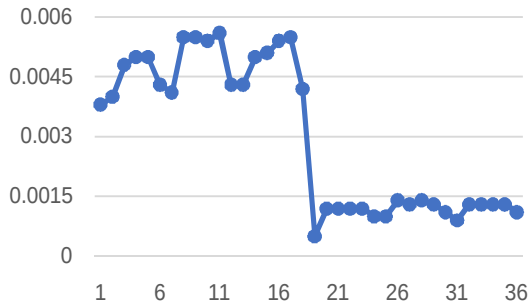

Fig 2: AP statistic values of Table 1 with modified 1st observation of second response

Upon comparison of Fig. 1 and Fig. 2, it becomes evident how the presence of outliers significantly impacts the estimation process. The outlier present in response 2 exerts a profound effect on the estimates, with a similar impact on response 1 as inferred from the decrease in their AP statistic value.

5. CONCLUSION:

Outliers play a significant role in data analysis, as evidenced by their impact on various statistical measures. Detecting and interpreting outliers can provide valuable insights into the underlying patterns. Understanding outliers and dealing with them is also important as they allow researchers to know whether the anomalies are genuine and base their decision on it. The test statistic was developed for multi – response experiments where the design matrix is not of full rank by suitably modifying AP statistic. The AP statistic also offers valuable insights into the relative impact of datapoints on overall analysis. This highlights its utility as a robust tool for discerning outliers and ensuring the integrity of statistical inferences in complex experimental designs. The developed statistic is tested on simulated data and showed significant results.

6. DECLARATIONS:

a. **Ethics Approval:**There is no ethicsapproval committee.

b. **Data Availability:** The data used for this study was simulated in R software and mentioned in **Table 1**.

c. **Consent to Publish declaration**: not applicable

7. REFERENCES:

Andrews D F and Pregibon D (1978) Finding the outliers that matter. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 40(1):85-93

Bates D, MaechlerM and Jagan M (2022) Matrix: Sparse and DenseMatrix Classes and Methods_. R package version 1.5-1. https://CRAN.R-project.org/package=Matrix.

Berenblut II and Webb G I (1974) Experimental design in the presence of autocorrelated errors. *Biometrika* 61(3):427-437.

Bhar L *M* (1997) Outliers in Experimental Designs [Unpublished Doctoral Dissertation, Indian Agriculture Research Institute]. *KrishiKosh*. http://krishikosh.egranth.ac.in/handle/1/2036687

Bhar L and Gupta V K (2001) A Useful Statistic for Studying Outliers in Experimental Designs. *Sankhyā: The Indian Journal of Statistics, Series B* 63(3):338–350. http://www.jstor.org/stable/25053185

Bhar L and Ojha S (2014) Outliers in multi-response experiments. *Communications in Statistics-Theory and Methods* 43(13):2782-2798.

Cook R D (1979) Influential observations in linear regression. *Journal of the American Statistical Association* 74(365):169-174.

Gerard D and Hoff P (2016) A higher-order LQ decomposition forseparable covariance models. *Linear Algebra and its Applications*505: 57-84. doi:10.1016/j.laa.2016.04.033

Lehmann E L, and Romano J P (2005) Generalizations of the familywise error rate. *The Annals of Statistics* 33(3):1138-1154.

Ojha S (2015) Outliers in Designed Experiments with Correlated Errors, [Unpublished Doctoral Dissertation, Indian Agriculture Research Institute]. *KrishiKosh*. http://krishikosh.egranth.ac.in/handle/1/5810010398.

Prasad R, Nandi PK, Bhar LM and Gupta V K (2008) Outliers in multi-response experiments.

R Core Team (2022) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. URL https://www.R-project.org/.

Venables W N and Ripley B D (2002) Modern Applied Statisticswith S. Fourth Edition. *Springer, New York*. ISBN 0-387-95457-0

Wickham H and Bryan J (2023) readxl: Read Excel Files_. R package version 1.4.3. https://CRAN.R-project.org/package=readxl.