Agri Articles

(e-Magazine for Agricultural Articles)

Volume: 05, Issue: 01 (JAN-FEB, 2025)

Available online at http://www.agriarticles.com

©Agri Articles, ISSN: 2582-9882

Role of Outliers in Experimental Design

(*M. Nirmala Devi[1], M. Radha[2], V.M.Indumathi[3], S. Senthilnathan[4] and P. Jeyalakshmi[5])

[1]Department of Physical Science & Information Technology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

[2]Department of Agricultural Economics, Anbil Dharmalingam Agricultural College and RI, Trichy, Tamil Nadu, India

[3]Department of Agricultural and Rural Management, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

[4]Department of Agronomy, Tamil Nadu Agricultural University, Coimbatore (TN)

[5]Department of Agricultural Extension and Communication, VOC Agricultural College and Research Institute, Killikulam, Tamil Nadu, India

*Corresponding Author's email: pnirmalah@tnau.ac.in

In designed experiments, the assumption of independence of observations may be violated due to various types of dependence among the observations within a block, such as autoregressive, moving average type, and equi-correlation. Additionally, deviations from the basic assumptions can occur in the presence of disturbances like outlier(s). In this article an attempt has been made to brief the role of outliers and types of outliers in experimental designs.

Key words: Outliers-Statistical Assumptions-Robustness

Introduction

Statistical models describe specific aspects of observed phenomena. In the National Agricultural Research System (NARS), scientists conduct experiments that generate data, which is then analyzed using statistical tools to inform agricultural recommendations. Experimental design and data analysis improve research quality. However, models are only approximately correct and depend on assumptions like independent and identically distributed (*i.i.d.*) samples, which are rarely met in reality. Outliers, caused by issues such as disease infestations or data recording errors, can invalidate statistical results. Therefore, detecting and managing outliers efficiently is crucial for accurate conclusions.

Diagnostics are important in regression analysis as they provide ways to study residuals and assess the impact of observations on the regression equation. These diagnostics use the deletion of observations to evaluate the quality of the fitted model. Cook (1977, 1979) developed a statistic for assessing the influence of an observation on parameter estimates. In designed experiments, the focus is typically on estimating linear functions of parameters, such as treatment contrasts. Bhar and Gupta (2001, 2003) created statistics to assess influence for this purpose. Additionally, Cook's (1986) method for assessing local influence through model perturbation has been applied to designed experiments by Bhar (2013). The work of Bhar and Gupta (2001) has been applied to various design experiments (refer to Sarker et al. (2005), Nandi (2007), and Bhar and Ojha (2013)).

Outliers on Data - Causes and its Impact

Since humans began utilizing information from collected data to better understand their world, concerns have emerged regarding unrepresentative or outlying observations within data sets. An outlier is defined as an observation (or subset of observations) that significantly differs from the rest of the dataset. The presence of outliers is quite common across various fields involving data collection, often stemming from heavy-tailed distributions or simply resulting from erroneous data points. Almost every real dataset contains outliers, with some key causes including:

(i) **Malicious activity** - This can involve exaggerating yields from specific treatments, reporting significant decreases in pathogen populations in response to certain pesticides, cyber intrusions, destruction of crops due to grazing or other attacks, or inflating cultivation costs to secure farm loans or benefits from government schemes.

(ii) **Instrumentation error** - This includes defects in machine components or general wear and tear.

(iii) **Environmental variations** - Factors such as climate change, shifts in consumer buying patterns, genetic mutations, and global warming can lead to outliers.

(iv) **Human error** - This encompasses mistakes in reading measurements, reporting inaccuracies, and data entry errors.

Some Terminology Related to Outliers

No observation can be completely assured as a reliable reflection of the phenomena being studied. The likely credibility of an observation is indicated by its connection to other observations obtained under similar circumstances. Observations that an investigator perceives as distinct from the majority of the data have been referred to by various terms, including "outliers," "extreme observations," "discordant observations," "rogue values," "contaminants," "surprising values," "mavericks," or "dirty data." An outlier is characterized by its significant deviation from the other members of the sample in which it belongs. As defined by Daniel (1960), an outlier is "an observation whose value does not fit the pattern established by the remaining data." A broader definition from Beckman and Cook (1983) includes:

- **Discordant observation**: Observations that appear surprising or inconsistent to the investigator.
- **Contaminant**: Observations that do not stem from the target distribution.
- **Outlier**: A term that collectively refers to either a contaminant or a discordant observation.
- **Influential cases**: An outlier doesn't necessarily need to be influential in the sense that the results of an analysis may remain largely unchanged when an outlier is excluded. It's useful to view an influential observation as a specific type of outlier.

Next, we will discuss how outliers are treated in linear statistical models that represent cause-and-effect relationships for subsequent statistical analysis.

Outliers in Linear Models and Design of Experiments

Two well-known concepts for studying outliers were proposed by Dixon (1950). In linear models, outliers can be addressed using either the mean-shift model or the variance-inflation model. The variance-inflation model for a single outlier assumes that the basic model is valid, except that the variance of one unknown response is greater than the others. However, the mean-shift outlier model is more extensively explored and discussed in statistical literature. According to this model, studentized residuals are frequently utilized to identify outliers in linear model analyses. The observation with the largest absolute studentized residual typically receives special attention and is considered the most likely contaminant. This approach is justified by the premise that the basic normal theory model holds true, except that the mean of at least one unknown response is altered. Specifically, if the ith observation is deemed an outlier, its mean shifts from μi to μi+c, where c is a non-zero value, effectively dedicating a single parameter to that observation. This model has been predominant in the literature related to linear models, particularly within linear regression frameworks.

As mentioned in the introduction, outliers can also arise in data from designed experiments, fitting within the general linear model setup. The concern about outliers in experimental designs is not new; efforts to create statistically objective methods for addressing outliers began in the mid-19th century, leading to a wealth of literature on the topic. An excellent review can be found in Beckman and Cook (1983). Furthermore, several books—such as those by Atkinson (1985), Barnett and Lewis (1984), Belsley, Kuh, and Welsch (1980), Myers (1990), and Rousseeuw and Leroy (1987)—aggregate much of the relevant literature on outliers.

A significant aspect of most statistics developed in this field is whether outliers influence the estimation of parameters or the residual sum of squares. These statistics were initially created for linear models with a full column rank design matrix. Subsequently, researchers began applying these statistics to other linear models. Data generated from designed experiments is also susceptible to outlier occurrences. Although experimental design generally aligns with linear model principles, identifying and testing for outliers in this context presents unique challenges. First, the design matrix in experimental designs typically does not have full column rank, making standard test statistics inapplicable. Second, the focus in designed experiments often lies in a subset of parameters. For instance, in block designs, the primary interest is in estimating treatment contrasts, while other parameters, such as block effects and overall means, are treated as nuisance parameters. Consequently, researchers may want to understand how an outlying observation affects the estimation of treatment contrasts. Unfortunately, the literature on this subject, particularly in block designs, is limited.

Identification of Outliers in Designed Experiments

The issue of outliers in linear regression models has been widely studied. Research focuses on two main areas: (i) identifying outliers for further analysis, and (ii) adapting models or methods to account for outliers. Detection involves outlier identification, while robust estimation methods aim to lessen outlier impact on parameter inference. In designed experiments, statistics like Cook-statistic, AP-statistic, and Q_k_-statistic are used to detect outliers in block design setups (Bhar and Gupta, 2001). For more information, refer to Bhar (1997) and Bhar and Gupta (2001).

Multiple Outliers in Design of Experiments

Identifying a single outlier or influential point in linear regression is straightforward. However, when multiple outliers or influential points are present, masking makes detection difficult. Masking occurs when one outlier hides another. Although much research addresses masking in linear regression, there is limited work on this issue in experimental design, except for Bhar et al. (2013), who developed a technique to detect multiple outliers with uncorrelated errors.

Robustness of Experimental Designs

Before analyzing experimental data, we must apply diagnostic procedures to detect any outliers. Various methods can identify outliers for further examination. However, an outlier might not always be influential and managing them can be challenging. Simply ignoring or deleting them is not advisable. Instead, in experimental designs, adopting a robust design insensitive to outliers is recommended. These designs ensure that outliers do not impact parameter estimation. Robust designs resist disturbances and can handle violations of assumptions, such as missing observations, which have been extensively studied by researchers like Bhar and Gupta (2001) and Sarker et al. (2005).

References

1. Atkinson, A.C. (1985). *Plots, transformations and regression*. Oxford University Press (*pp* 292): Oxford
2. Barnett, V. and Lewis, T. (1984). *Outliers in statistical data*. 3rd ed., John Wiley: New York.
3. Beckman, R.J. and Cook, R.D. (1983). Outlier….s (with discussion). *Technometrics*, **25**, 119-163.
4. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*. John Wiley: New York.

5. Bhar, L. (1997). *Outliers in Experimental Designs*. Ph.D. thesis, IARI, New Delhi.
6. Bhar, L. and Gupta, V.K. (2001). A useful statistic for studying outliers in experimental designs. *Sankhya* B, **63**, 338-350.
7. Bhar, L. and Gupta, V.K. (2003). Study of outliers under variance-inflation model in experimental designs. *J. Ind. Soc. Agril. Stat.***, 56**(**2**), 142-154.
8. Bhar, L., Gupta, V.K., Parsad, R. (2013). Detection of Outliers in designed experiments in the presence of masking. *Statistics and Applications*, **11**(**1**&**2**), 147-160.
9. Bhar, L. and Ojha, S. (2013). Outliers in multi-response experiments**.** *Communications in Statistics-Theory and Methods***, 43**(**13**), 2782-2798.
10. Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics,* **19**, 15-18.
11. Cook, R.D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.*, **74**, 169-174.
12. Cook, R.D. (1986). Assesment of local infuence. *J. Roy. Statist. Soc.* B**, 48(2)**, 133-169.
13. Daniel, C. (1960). Locating outliers in factorial experiments. *Technometrics*, **20**, 385-395.
14. Dixon, W.J. (1950). Analysis of extreme values. *Ann. Math. Statist*., **21**, 488-506.
15. Myers, R.H. (1990). *Classical and Modern Regression Analysis with Applications,* 2nd Ed., PWS-Kent: Boston.
16. Nandi, P.K. (2007). *Design and Analysis of Multi-response Experiments.* Ph.D. thesis, IARI, New Delhi.
17. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection*. John Wiley: New York.
18. Sarker, S., Parsad, R. and Gupta, V.K. (2005). Outliers in block designs for diallel crosses. *Metron*, **63 (2)**, 177-191.