RESIDUALS IN MINIMAL RESOLUTION IV DESIGNS

Pen-Hwang Liau[1]

Abstract

In unreplicated factorial or fractional factorial experiments, the presence of one or more outliers can seriously affect the analysis of variance. Using the normal plot of t residuals to identify outliers in factorial or fractional factorial is an easy method to find these dubious points. In some cases, the t residuals form the identical pairs. One can not tell from the plot which is doubtful. This phenomenon occurs for all minimal designs of resolution IV, which fits the model containing all main effects and some two-factor interactions, whether it is orthogonal or not. In these kinds of models, when we drop one point or two points (not foldover pair) from the fraction, the phenomenon of identical pairs of t residuals may still occur. In this paper, the theoretical background of the phenomenon and its sequences will be investigated in detail.

1. Introduction

When a linear model is fitted to the observational or experimental data set, one can hardly be certain whether the model is appropriate or not before checking some basic assumptions. Residuals act an important role in the linear model. One usually uses residuals for detecting normal assumption, homogeneous variance, influential observations, outliers, *etc.* For the linear model, the random errors ε_i are generally assumed to be independent normal variables with mean 0 and constant variance σ^2. Unfortunately, the residuals e_i are not independent and do not have common variance. In performing the data analysis, the residuals and the scaled versions of the residuals have been used extensively to study validity of the linear model and its assumptions.

The heterogeneous variances of the residuals and the lack of independence among three types of residuals, the residuals (e_i), standardized residuals (r_i), and studentized residuals (t_i), complicate interpretation of their behavior. Three types of residuals are denoted as follows:

$$e_i, \quad r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \ .$$

We will explain the notations of these three types of residuals later. In spite of the problems associated with their usage, the three types of residuals have been proven usefully for detecting model's inadequacies and outliers. One has to concern that the presence of one or more outliers in factorial or fractional factorial designs can seriously affect the analysis of variance. This differs from the usual multiple regression analysis which has more degrees of freedom for the residual sum of squares, and one outlier may affect the residuals marginally only. Several authors apply F and/or t-tests to identify the spurious points, such as Quenouille (1953), Daniel (1960), Goldsmith and Boddy (1973), John and Prescott (1975), and Weisberg (1985). In fact, these tests based on the statistics that were used by these authors are equivalent to each other. Many other experimenters are used to applying the normal plot of residuals to detect various departures from the assumptions. It is also reasonable to use the normal plot of residuals to identify unusual large residuals. Box *et al.* (1978), for example, suggested this approach to find the unusual observations. Suppose that the normal plot of residuals is approximately a straight line, then it indicates that there are no outliers. Otherwise, possible outliers will be presented in the upper right, or the lower left off the line. Liau (1999) addressed that the normal plot of externally studentized residuals (or t residuals) is more sensitive than the normal plot of residuals for identifying the spurious points. If unfortunately the t residuals form the identical equal pairs in some situations, then one can not distinguish which point is an outlier from the model.

This article is organized as follows. After giving a brief description of the minimal resolution IV designs, we develop a theorem and give an example to explain the idea of the identical pair of residuals. In Section 3, we explore the properties of omitting or missing an observation or two observations from the model. Conclusion is given in Section 4.

2. MINIMAL RESOLUTION IV DESIGN

A two-level fractional factorial design is of resolution IV if the main effects are clear of two-factor interactions and some two-factor interactions are aliased with each other. Any 2^{n-p} fractional factorial design with resolution IV has to contain at least $2n$ points, and resolution IV design that contains exactly $2n$ points are called minimal designs. All minimal designs are foldover designs. This property has been showed by Webb (1968) and Margolin (1969). Another property which is worthy to mention here is that the minimal resolution IV design can convert any resolution III design for n factors into a resolution IV design for $n+1$ factors.

It is well known that the resolution IV design may be obtained from resolution III design by the process of fold over. That is, we can fold over a 2^{n-p} resolution III design, simply adding the original fraction to a second fraction with all the signs changed. And the plus signs in the identity column in the first fraction could be switched in the second fraction, then a $(n+1)^{th}$ factor could be associated with this column. For example, we can reverse all the signs of the 2^{3-1} design with defining relation $I = ABC$, and then the new design is defined by $I = -ABC$ with 4 points. We then add the fourth factor, say D, to the eight-point combined design with high level in the first fraction and low level in the second fraction. It is easy to verify that the resulting design is a 2^{4-1} minimal resolution IV design. This fraction is also a foldover design. It consists of four pairs of points which have foldover images of each other:

$$(1) \text{ and } abcd; \quad ab \text{ and } cd; \quad ac \text{ and } bd; \quad ad \text{ and } bc.$$

There are some other useful nonorthogonal minimal resolution IV designs. The 2^3 design with six points, 2^5 design with ten points, and 2^6 design with twelve points, for example, are also interesting designs in practice. If the analysis of variance procedure should include a set of main effects and some two-factor interactions, using residuals (or t residuals) for the model checking will meet some difficulties. As we mentioned in the introduction, the residuals act an important role for checking the assumptions of the model. In this situation, the t residuals occur in identical pairs, and one can not tell which of the two observations is to blame from the t residuals. We will give an example and the theoretical reasons to show this phenomenon in detail. In practice, we can easily get around the difficulty if we reduce the number of main effects.

We refer back to this difficulty that the model contains the main effects and some two-factor interactions. The following theorem provides the phenomenon

of identical pair of residuals (or t residuals).

THEOREM 1. *Suppose that the model includes all main effects and some two-factor interactions in the minimal resolution IV design for n factors. If the design matrix is nonsingular, then the t residuals of the foldover pair are identically equal.*

The proof of the theorem is given in Appendix.

EXAMPLE 1. We use an example of semiconductor experiment that is reported by Kackar and Shoemaker (1986) to show the properties of t residuals. This is a minimal resolution IV design for eight factors in sixteen runs. It takes a 2^4 with A, B, C, E and adds another four new factors by setting

$$D = ABC, \quad F = ABE, \quad G = ACE, \quad H = BCE.$$

The observed response is the mean epitaxial thickness of wafers fabricated. If the data set is fitted by all main effects and two-factor interactions, for example AB and AD, then the t residuals of the foldover pair are identical. The original data set and the t residuals are shown in Table 2.1 and Table 2.2, respectively.

TABLE 2.1 *Data for* 2^8 *minimal design*

(1)	adfg	bdfh	abgh	cdgh	acfh	bcfg	abcd
14.821	13.972	14.165	14.878	14.037	14.843	14.757	13.907
abcdefg	bceh	aceg	cdef	abef	bdeg	adeh	efgh
13.914	14.921	14.415	13.880	14.932	13.860	14.032	14.888

TABLE 2.2 t *residuals for* 2^8 *minimal design*

(1)	adfg	bdfh	abgh	cdgh	acfh	bcfg	abcd
−1.045	1.354	−1.511	−0.277	2.647	−0.115	0.200	−0.402
abcdefgh	bceh	aceg	cdef	abef	bdeg	adeh	efgh
−1.045	1.354	−1.511	−0.277	2.647	−0.115	0.200	−0.402

Hence the normal plot of the t residuals will consist of eight double points and each foldover pair is identically equal. That is, they have the property that the t residuals of two observations, which form a foldover pair, are identical.

In general, the normal plot can show that one, or both, of the points in a foldover pair is an outlier, but which one? Suppose that after the normal

plot of t residuals is performed, one of the points is clearly "off the line". The dubious point is the point and its complement. One, or both, of them may be an outlier but we can not separate them, and tell which is the offender. This kind of phenomenon occurs in minimal design no matter it is orthogonal or non-orthogonal.

3. OMITTING AN OBSERVATION OR TWO OBSERVATIONS FROM THE MODEL

When using the fractional factorial designs, sometimes an observation may be missing. This may happen because of carelessness, error, or for any reasons beyond the experimenter's control. Or sometimes we have to drop a point from the data set and then fit the model for the remaining points. When the fitted model is used to predict the new value for this dropping point, there are some particular properties happening in the minimal design. Suppose the minimal design is applied by the experimenter, if the model contains all main effects as well as some two-factor interactions, and an observation is omitted from the model for some reasons, then there will be some interesting things that are worthy to explore. We will show what happens when we omit an observation from the model and also predict this value from the remaining $2n - 1$ points.

COROLLARY 1. *Suppose that we drop one point from the model that contains all main effects and some two-factor interactions. The t residual of the other point in the foldover pair equals to zero, and the t residuals of the remaining foldover pair are still identically equal.*

From the procedure of the proof of Theorem 1 in Appendix, it is just enough to show that this phenomenon happens in residuals for the linear model. It means that the phenomenon happens in residuals. It also happens in t residuals. The proof of Corollary 1 is given in Appendix.

Suppose that a foldover pair is omitted from the model, the design matrix will become a singular matrix. It is just as John (1995) mentioned in his paper that the form of the design matrix remains the same, but corresponding rows have been dropped from $\mathbf{V}$ and $-\mathbf{V}$ (see Appendix). The rank of $\mathbf{V}^T\mathbf{V}$ is only $n - 1$, and then $\mathbf{X}^T\mathbf{X}$ is singular. In this situation, the degree of resolution of the design is no longer four.

COROLLARY 2. *Suppose that two points (not foldover pair) are omitted from the model that contains all main effects and some two-factor interactions. Then*

the t residuals of the other point in the corresponding foldover pair equal to zero,
and the residuals of the remaining foldover pair are still identically equal.

The proof of the above corollary is also given in Appendix.

4. CONCLUSION

Designed experiment with 2^n factorials are being used widely in industry. When we do the data analysis, any one or several of the features of the model, such as normality of the error term, constant variance, or outliers, may not be appropriate for the model assumptions. In this case, hence, residual analysis is a highly useful mean of examining the aptness of a model. The normal plot of t residuals is more sensitive for detecting outliers than the normal plot of residuals. Unfortunately, when one is using minimal design in the experiment, if the model contains all main effects and some two-factor interactions, the t residuals of the foldover pairs form the identical equal pairs. Thus when an outlier is presented, one can not identify the point. Furthermore, when we drop one observation or two observations (not foldover pair) from the model, the phenomenon still happens in the t residuals for the foldover pair. In fact we can easily get around the difficulty if we reduce the number of unimportant main effects from the model and leave more degrees of freedom for the residuals. The further study is required to find outliers on the problem when all main effects and some two-factor interactions are all significant in the model.

APPENDIX

PROOF OF THEOREM 1. Denote the levels of factors by "+1" for the high level and "−1" for the low level and fit a regression model with all main effects and some two-factor interactions. The design matrix with a column ones on the left can be written as

$$X = \begin{bmatrix} 1 & V & U \\ 1 & -V & U \end{bmatrix},$$

where V is a $n \times n$ square matrix for all main effects, and U is a $n \times k$ nonsingular matrix for some two-factor interactions. The matrix form of $X^T X$ is

$$X^T X = \begin{bmatrix} 2n & 0^T & 21^T U \\ 0 & 2V^T V & 0 \\ 2U^T 1 & 0^T & 2U^T U \end{bmatrix}.$$

Hence

$$2(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1/c & \mathbf{0}^T & -\mathbf{1}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}/c \\ \mathbf{0} & \mathbf{V}^{-1}(\mathbf{V}^T)^{-1} & \mathbf{0} \\ -(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}/c & \mathbf{0}^T & \mathbf{B}^{-1} \end{bmatrix},$$

where $c = n - \mathbf{1}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}$ and

$$\begin{aligned}
\mathbf{B}^{-1} &= \left(\mathbf{U}^T\mathbf{U} - \frac{1}{n}\mathbf{U}^T\mathbf{1}\mathbf{1}^T\mathbf{U}\right)^{-1} \\
&= \left\{\mathbf{I}_k - \frac{1}{n}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}\mathbf{1}^T\mathbf{U}\right\}^{-1}(\mathbf{U}^T\mathbf{U})^{-1} \\
&= \mathbf{I}_k + \frac{1}{n}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}\left\{1 - \mathbf{1}^T\mathbf{U}\left(\frac{1}{n}\right)(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}\right\}^{-1}\mathbf{1}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1} \\
&= \mathbf{I}_k + \frac{1}{c}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}\mathbf{1}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1} \\
&= (\mathbf{U}^T\mathbf{U})^{-1} + \frac{1}{c}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{1}\mathbf{1}^T\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}.
\end{aligned}$$

Let $\mathbf{H}_U = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. After the tedious algebraic operation, we can obtain

$$2c\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X} = 2c\mathbf{H} = \begin{bmatrix} \mathbf{A} & \mathbf{D} \\ \mathbf{D} & \mathbf{A} \end{bmatrix},$$

where $\mathbf{A} = (\mathbf{I}_n - \mathbf{H}_U)(\mathbf{J}_n - c\mathbf{I}_n - \mathbf{J}_n\mathbf{H}_U) + 2c\mathbf{I}_n$ and $\mathbf{D} = (\mathbf{I}_n - \mathbf{H}_U)(\mathbf{J}_n - c\mathbf{I}_n - \mathbf{J}_n\mathbf{H}_U)$. From this we get

$$2c(\mathbf{I}_{2n} - \mathbf{H}) = \begin{bmatrix} (\mathbf{I}_n - \mathbf{H}_U)(c\mathbf{I}_n + \mathbf{J}_n\mathbf{H}_U - \mathbf{J}_n) & (\mathbf{I}_n - \mathbf{H}_U)(c\mathbf{I}_n + \mathbf{J}_n\mathbf{H}_U - \mathbf{J}_n) \\ (\mathbf{I}_n - \mathbf{H}_U)(c\mathbf{I}_n + \mathbf{J}_n\mathbf{H}_U - \mathbf{J}_n) & (\mathbf{I}_n - \mathbf{H}_U)(c\mathbf{I}_n + \mathbf{J}_n\mathbf{H}_U - \mathbf{J}_n) \end{bmatrix},$$

where $\mathbf{J}_n$ is a $n \times n$ square matrix of ones, $\mathbf{I}_n$ is a $n \times n$ identity matrix. From above matrix form, it is clear to know that the ith and $(n+i)^{th}$ rows of $2c(\mathbf{I}_{2n} - \mathbf{H})$ are identical. That is, for the foldover pair $e_i = e_{n+i}$. Let h_{ii} be the ith diagonal element of $\mathbf{H}$. It is also not hard to obtain that $h_{ii} = h_{n+i,n+i}$ from the matrix $2c\mathbf{H}$. We may recall the formula from the general book of regression analysis; see, for example, Weisberg (1985).

The t residual is formed by

$$t_i = r_i \left(\frac{n - p' - 1}{n - p' - r_i^2}\right)^{\frac{1}{2}},$$

where $r_i = e_i/\{\hat{\sigma}(1 - h_{ii})^{1/2}\}$. For the foldover pair, $e_i = e_{n+i}$ and $h_{ii} = h_{n+i,n+i}$, it implies $t_i = t_{n+i}$. The theorem depends on the design that should be a minimal design and that the model should at least include all main effects. The theorem can be applied to both orthogonal and non-orthogonal designs.

Furthermore, if the model only contains all main effects, in this case, $c = n$ and the form of the matrix is

$$2n(\mathbf{I}_{2n} - \mathbf{H}) = \begin{bmatrix} n\mathbf{I}_n - \mathbf{J}_n & n\mathbf{I}_n - \mathbf{J}_n \\ n\mathbf{I}_n - \mathbf{J}_n & n\mathbf{I}_n - \mathbf{J}_n \end{bmatrix}.$$

This is the same as John's (1995) result.   □

PROOF OF COROLLARY 1. The model in terms of observations may be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Here, $\mathbf{Y}$ is a $2n \times 1$ vector of the observations, $\mathbf{X}$ is a $2n \times p'$ design matrix, $\boldsymbol{\beta}$ is a $p' \times 1$ vector of parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a $2n \times 1$ vector of random errors. Let $\mathbf{x}_i^T$ be the ith row of $\mathbf{X}$, and h_{ii} be the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Write $\hat{\boldsymbol{\beta}}_{(i)}$ for a $p' \times 1$ vector of regression coefficients estimated with the ith observation omitted, and e_i for the ith residual of the residual vector $\mathbf{e}$. The formula from the book of regression analysis, for example Weisberg (1985), can be obtained as

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_{ii}}.$$

It means that if the ith point is omitted and the model is refitted by the remaining $2n - 1$ points, then the relationship between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ can be expressed by the above formula. Furthermore, the new predicted value for the case i can be estimated from the remaining $2n - 1$ points and it equals to $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(i)}$. Let $\mathbf{r}$ be the vector of residuals for the new fit. Then

$$\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$$
$$= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$$
$$= \mathbf{e} + \mathbf{X}\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_{ii}}.$$

Hence the jth residual for the new fit is formed by

$$r_j = e_j + \frac{h_{ij}e_i}{1 - h_{ii}}.$$

Without loss of generality, we let $1 \leq i, j \leq n$. It is clear that $h_{ij} = h_{i,n+j}$ and $h_{ij} = h_{ji}$ by the above matrix form for $j \neq i$. Then we obtain

$$r_j = e_j + \frac{h_{ij}e_i}{1 - h_{ii}} = e_{n+j} + \frac{h_{i,n+j}e_i}{1 - h_{ii}} = r_{n+j}.$$

This means that the residuals of the foldover pair, except the ith pair, again occur in pairs. Hence, the t residuals also have the same phenomenon. For the omitted point

$$r_i = e_i + \frac{h_{ii}e_i}{1 - h_{ii}} = \frac{e_i}{1 - h_{ii}}.$$

Let $h_{ii} = (k_i + 2c)/2c$, then $h_{i,n+i} = k_i/2c$, where k_i is the ith diagonal element of $(\mathbf{I}_n - \mathbf{H}_U)(\mathbf{J}_n - c\mathbf{I}_n - \mathbf{J}_n\mathbf{H}_U)$. In this case, the residual of the omitted point in the foldover pair is

$$r_{n+i} = e_{n+i} + \frac{h_{i,n+i}e_i}{1 - h_{ii}} = e_{n+i} - e_i = 0.$$

It indicates that the other point of the pair fits the new regression model perfectly because its residual is zero. The residual (or t residuals) of the remaining points again occur in pairs. In this case it is still not easy to identify the exact outlier. □

PROOF OF COROLLARY 2. For convenience, we omit the first point from the fraction. Denote the new hat matrix by $\mathbf{H}^*$ and the matrix for the residuals by $\mathbf{I} - \mathbf{H}^*$ for the new model with $2n-1$ points. From Corollary 1, these two matrices shall be the form

$$\mathbf{I} - \mathbf{H}^* = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \mathbf{M} \\ \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{M} & \mathbf{0} & \mathbf{M} \end{bmatrix}$$

where $\mathbf{M}$ is a $(n-1) \times (n-1)$ matrix, and

$$\mathbf{H}^* = \begin{bmatrix} \mathbf{I} - \mathbf{M} & \mathbf{0} & -\mathbf{M} \\ \mathbf{0}^T & 1 & \mathbf{0}^T \\ -\mathbf{M} & \mathbf{0} & \mathbf{I} - \mathbf{M} \end{bmatrix}.$$

From the hat matrix $\mathbf{H}^*$, we obtain

$$h_{tu}^* = h_{t,n+u}^* = h_{n+t,u}^* \text{ for } t \neq u, \ 1 \leq t, u \leq n - 1 \text{ and } h_{tt}^* = h_{t,n+t}^*.$$

Suppose that the jth point is dropped from the fraction. If the first point has been omitted from the fraction, and the model is refitted by the remaining $2n - 2$

points. As we discussed in the proof of Corollary 1, the residual, say δ, for the new fit is formed by

$$\delta_k = r_k + \frac{h_{jk}^* r_j}{1 - h_{jj}^*}.$$

Since $r_k = r_{n+k}$ and $h_{jk}^* = h_{j,n+k}^*$, we have $\delta_k = \delta_{n+k}$ for the point k that is not a foldover pair with point j. Otherwise, the residual for the other point in the foldover pair with the point j is

$$
\begin{aligned}
\delta_{n+j} &= r_{n+j} + \frac{h_{j,n+j}^* r_j}{1 - h_{jj}^*} \\
&= r_{n+j} + \frac{(h_{jj}^* - 1) r_j}{1 - h_{jj}^*} \\
&= r_{n+j} - r_j \\
&= 0.
\end{aligned}
$$

With the same idea, it is easy to obtain that the residual for the other point in the foldover pair with the first point is still 0.   □

<h2>REFERENCES</h2>

BOX, G. E. P., HUNTER, W. G. AND HUNTER, J. S. (1978). *Statistics for Experimenters*, John Wiley, New York.

DANIEL, C. (1960). "Locating outliers in factorial experiments", *Technometrics*, **2**, 149–156.

GOLDSMITH, P. L. AND BODDY, R. (1973). "Critical analysis of factorial experiments and orthogonal fractions", *Applied Statistics*, **22**, 141–160.

JOHN, J. A. AND PRESCOTT, P. (1975). "Critical values of a test to detect outliers in factorial experiments", *Applied Statistics*, **24**, 56–59.

JOHN, P. W. M. (1995). "Pattern in residuals for 2^n designs", *Journal of Applied Statistics*, **22**, 253–259.

KACKAR, R. N. AND SHOEMAKER, A. C. (1986). "Robust design : A cost-effective method for improving manufacturing process", *AT & T Technical Journal*, **65**, 39–50.

LIAU, P-H. (1999). "Finding outliers in unreplicated factorial designs", *Journal of the Chinese Statistics Association*, **37**, 371–383.

MARGOLIN, B. H. (1969). "Resolution IV fractional factorial designs", *Journal of the Royal Statistical Society*, **B31**, 514–523.

QUENOUILLE, M. H. (1953). *The Design and Analysis of Experiment*, Griffin, London.

WEBB, S. R. (1968). "Non-orthogonal designs of even resolution", *Technometrics*, **10**, 291–299.

WEISBERG, S. (1985). *Applied Linear Regression*, 2nd ed., John Wiley, New York.