THE ANALYSIS OF RESIDUALS

by

BRUMMETT JERALD McCLENDON

B. A., Texas Technological College, 1965

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics and
Statistical Laboratory

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

A. M. Feyerherm
Major Professor

CONTENTS

THE ANALYSIS OF RESIDUALS

1. Introduction

1.1 General Comments.

An analysis of the residuals--the observed value minus the fitted value--can answer many useful and sometimes necessary questions concerning data analysis. One of the most important of these questions concerns the assumptions inherent in a classical least-squares analysis. This will be the central theme of this paper.

The assumptions underlying a classical least-squares analysis (and the resulting F-test) are very reasonable for most situations. Also, the F-test has been found to be robust. However, in some situations one may doubt the validity of these assumptions. If large amounts of data are available or long-term projects are contemplated, it seems reasonable to first test these assumptions before doing the calculations and tests of hypothesis. This may lead to improvements in the consequent analysis, such as use of nonparametric techniques or transformations.

This paper will present both graphic methods and test statistics for testing assumptions. The distributions of test statistics will be found under the null hypothesis. Methods of modifying residuals and some comments on outliers will be presented. An example will be given to illustrate the procedures. The paper will conclude with some stated limitations and cautions.

1.2 Defining the Problem in Mathematical Notation.

Let $\underline{Y}$ be an (nxl) vector of given observations which are claimed to be independent determinations of the (nxl) vector of means $\underline{\mu}$ such that

$$\underline{\mu} = \underline{A}\ \underline{\theta} \quad , \tag{1.2a}$$

where $\underline{A}$ is a (nxp) matrix of given coefficients and $\underline{\mu}$ is a (pxl) vector of unknown parameters. Let $\hat{\underline{\theta}}$ denote the estimates of $\underline{\theta}$ obtained by least-squares. Then the fitted values become an (nxl) vector $\underline{Y}$ such that

$$\hat{\underline{Y}} = \underline{A}\ \hat{\underline{\theta}} \quad , \tag{1.2b}$$

and the residuals become an (nxl) vector $\underline{Z}$ such that

$$\underline{Z} = \underline{Y} - \hat{\underline{Y}} \quad . \tag{1.2c}$$

Let $\underline{Q} = (q_{ij})$ be an idempotent positive-semidefinite symmetric matrix which takes (y_i) into (z_i); i.e.,

$$z_i = \sum_j q_{ij} y_j \quad . \tag{1.2d}$$

If $\underline{A}$ has rank $(n-\nu)$, then $\underline{Q}$ has rank ν. To show this, note that the linear subspace C generated by the rows (columns) of $\underline{A}$ has dimension $(n-\nu)$. Also, $\underline{Z}$ is the projection of all n-component vectors orthogonal to C--call it $\bar{C}$. $\bar{C}$ has dimension ν which means $\underline{Q}$ has rank ν (see Rao, 1961). In an analysis of variance, ν becomes the error degrees of freedom.

Given $\underline{A}$, $\theta_1, \theta_2, \ldots, \theta_{n-\nu}$ is chosen so that the columns of $\underline{A}$ are linearly independent.

From (1.2a) and (1.2b),

$$\underline{Z} = \underline{Q}\,\underline{Y} \quad ,$$

$$\underline{Y} - \hat{\underline{Y}} = \underline{Q}\,\underline{Y} \quad :$$

It can be shown (Graybill, 1961) that $\hat{\underline{Y}} = \underline{A}\,\hat{\underline{\theta}} = \underline{A}\,(\underline{A'A})^{-1}\underline{A}'Y$. Let $V = (\underline{A'A})^{-1}$. Then,

$$\underline{Y} - \underline{A}\,\underline{V}\,\underline{A}'\underline{Y} = \underline{Q}\,\underline{Y} \quad ,$$

$$\underline{Q} = \underline{I} - \underline{A}\,\underline{V}\,\underline{A}' \quad .$$

The trace of $(q_{ij}) = \nu$.

In developing certain statistics it is convenient and sometimes necessary to put restrictions on $\underline{A}$ and $\underline{Q}$. They are the following:

(1.2e) (1) C contains the unit vector; that is, one of the parameter set $\underline{\theta}$ is the general mean, so that the corresponding column of $\underline{A}$ consists entirely of one's.

(1.2f) (2) The diagonal elements of $\underline{Q}$ are all equal.

A consequence of (1.2e) is that each row or column of $\underline{Q}$ sum to zero. A consequence of (1.2f) is that each diagonal element of $\underline{Q}$ is equal to ν/n, since the sum of the diagonal elements of a symmetric idempotent matrix is equal to its rank.

Eisenhart (1947) discusses the assumptions necessary for the analysis of variance, but for the purposes of this paper it will be stated that the estimation of $\underline{\theta}$ by least-squares procedures can be shown to be satisfactory if the following three assumptions are met:

(1.2g) (1) The effects in the model are additive.

(1.2h) (2) The error variance is independent of the mean (the components of $\underline{Y}$ are realizations of independent random variables all with equal variances).

(1.2i) (3) The components of $\underline{Y}$ are realizations of independent random variables all normally distributed.

Since the emphasis in this paper will be placed on the two-way classification, the following notation will be used in certain situations.

The $n = rc$ observations y_{ij}, $i = 1,2,\ldots,r$, and $j = 1,2,\ldots,c$, will have row means $\bar{y}_{i.} = \sum_i Y_{ij}/c$, column means, $\bar{y}_{.j} = \sum_{ij} Y_{ij}/r$, and a grand mean $\bar{y} = \sum_{ij} Y_{ij}/n$. The fitted values will be denoted by

$$\hat{y}_{ij} = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y} \quad ,$$

and the residuals will be denoted by

$$z_{ij} = y_{ij} - \hat{y}_{ij} \quad .$$

The analysis of variance table will be as follows:

Source	dF	SS
Rows	r-1	$SS\text{-}R = c\sum_i (\bar{y}_{i.} - \bar{y})^2$
Columns	c-1	$SS\text{-}C = r\sum_j (\bar{y}_{.j} - \bar{y})^2$
Residual	ν	$\nu S^2 = \sum_{ij} z_{ij}^2$

In order to get precise tests of the assumptions (1.2g) - (1.2i), one would need large samples and several sets of data. This, however, should not keep one from examining these assumptions on small samples or a single data set, but rather should indicate the weight put on the results of such an examination. It should be noted that some of the procedures discussed in this paper require modification when used for combined data sets. The main consideration in combining the data sets is the equality of variances. This usually can be corrected by using standardized residuals.

2. Graphic Techniques

Probably the most revealing and most important of all statistical techniques is the graph. This section presents some graphic techniques, using residuals, for testing assumptions (1.2g) - (1.2i).

2.1 First Scatter Plot.

A good way to begin an examination of residuals is to plot them against the fitted values. With increased use of computers, most authors suggest routine calculation of the residuals on each analysis. The suggested scatter plot is a simple next step. Outliers will show up as isolated points. As a guide for interpretation of this plot, note the following:

(1) The expected value of the residuals is zero.

(2) The linear regression of the residuals on the fitted values has zero slope.

A dependence of variability on level of response will reveal itself by causing greater dispersion in some sections of the scatter plot than in others. A non-linear graph will indicate that the effects in the model are not additive. The technique may be used on a combination of several data sets if the residuals are standardized.

2.2 FUNOP.

FUNOP (Full Normal Plot) is a graphic technique suggested by Tukey (1962) and Anscombe and Tukey (1963), which has greater sensitivity than the scatter plot or the cummulative plot on normal probability paper.

Let $\overset{\prime}{z}$ represent the median of the residuals and $a_{k/n}$ be a value chosen to be typical of the k-th value from the bottom in a sample of n from a unit normal distribution. Many choices are available for $a_{k/n}$. This paper will use Table XX of Fisher and Yates (1963). A plot of $(z_k - \overset{\prime}{z})/a_{k/n}$ against k (omitting the middle third of the k's) has the following properties:

(1) A single outlier will be revealed by a large value for $k = 1$ or $k = n$.

(2) If the plot turns up at both ends, one should suspect that he is either dealing with a long-tailed distribution or a number of outliers of each sign.

(3) Skewness will be revealed by the plot being higher at one end than at the other.

Therefore, a straight line with zero slope would indicate that assumptions (1.2g) - (1.2i) are met.

2.3 Second Scatter Plot.

This technique is presented by Anscombe and Tukey (1963) as a method of detecting removable nonadditivity. Consider the regression coefficient f of the residuals regressed on the fitted values. It can be shown (see Section 3.1) that $f = B/A$, where

$$B = \sum_i z_i \hat{y}_i^2 \quad ,$$

and A is the residual sum of squares obtained by doing the conventional analysis of variance on the square of the fitted values. If one plots the new residuals (Z_{ij}) against the old residuals (z_{ij}), removable nonadditivity will show up as a distinct straight line. A test of significance for nonadditivity using B and A is given in Section 3.1.

3. Test Statistics

Test statistics presented in this section are not claimed to have sweeping statistical properties. They are, however, claimed to be excellent guides, in absence of better techniques, for handling the important problem of testing the assumptions necessary for a classical least-squares analysis $\big((1.2g) - (1.2i)\big)$.

3.1 Additivity.

Tukey (1947) stated that failure of the assumption of additivity (1.2g) may be more serious than the others for tests of significance in an analysis of variance. In order to test additivity, consider the following statistic due to Anscombe (1961):

$$f = \frac{\sum_i z_i \hat{y}_i^2}{\sum_{ij} q_{ij} \hat{y}_i^2 \hat{y}_j^2} \quad , \tag{3.1a}$$

where (z_i) and (q_{ij}) are as defined in Section 1.2 and $\underline{\hat{Y}}^2$ is a matrix of the square of the fitted values (see 1.2b). It is claimed that f is a rough estimate of ϕ in some model $y = x + \phi(x' - \mu_o)^2$, where μ_o is some convenient central value. Note also that $f = B/A$, as discussed in Section 2.2.

The f statistic is motivated by the fact that

$$E(z_i) = E\left(\sum_j q_{ij} y_j\right) \quad ,$$

or

$$E(z_i) = E\left[\sum_j q_{ij}\left(x_j + \phi(x_j - \mu_o)^2\right)\right] \quad .$$

If the x_i are distributed as $N(\mu_i, \sigma^2)$ and (1.2d) holds, then

$$E(z_i) = \sum_j q_{ij}\mu_j + \phi\sum_j q_{ij}E(x_j - \mu_o)^2 \quad .$$

Next, note that restriction (1.2e) implies that

$$\sum_j q_{ij} \cdot (\text{constant}) = 0 \quad , \tag{3.1b}$$

since any row or column of (q_{ij}) sums to zero. Also, (1.2e) implies that

$$\sum_j q_{ij}\mu_j = 0 \quad ; \tag{3.1c}$$

since

$$\underline{Q}\ \underline{\mu} = (\underline{I} - \underline{A}(\underline{A'A})^{-1}\underline{A}')\underline{A}\ \underline{\theta} = \underline{0} \quad .$$

From (3.1c),

$$E(z_i) = \phi\sum_j q_{ij}E(x_j - \mu_o)^2 \quad .$$

By adding and subtracting $(\mu_j - \mu_o)^2$ and combining terms, it follows that

$$E(z_i) = \phi\sum_j q_{ij}E\left(x_j^2 - 2\mu_o(x_j - \mu_j) - \mu_j^2 + (\mu_j - \mu_o)^2\right) \quad . \tag{3.1d}$$

Since $E(x_j) = \mu_j$,

$$(3.1e) \qquad E\left(-2\mu_o(x_j - \mu_j)\right) = 0 \quad .$$

If x_j is $N(\mu_j,\sigma^2)$, then

$$(3.1f) \qquad E\left(x_j^{\,2} - \mu_j^{\,2}\right) = \sigma^2 \quad .$$

From (3.1c),

$$(3.1g) \qquad \phi\sum_j q_{ij}\sigma^2 = 0 \quad .$$

By substituting (3.1e), (3.1f), and (3.1g) into (3.1d), it follows that

$$E(z_i) = \phi\sum_j q_{ij}(\mu_j - \mu_o)^2 \quad .$$

Finally, applying (3.1c) again,

$$E(z_i) = \phi\sum_j q_{ij}\mu_j^{\,2} \quad .$$

If Q is such that $E(z_i)$ does not vanish, (z_i) has a linear regression on $(\sum_j q_{ij}\mu_j^{\,2})$. Remembering (1.2a) and (1.2b), the foregoing indicates a study of the statistic $\sum_{ij} z_i q_{ij}\hat{y}_j^{\,2} = \sum_i z_i\hat{y}_i^{\,2}$.

If (1.2f) holds, it can be shown (Anscombe, 1961) that

$$E\left(\sum_j z_i\hat{y}_i^{\,2}\right) \cong \phi\sum_{ij} q_{ij}\hat{y}_i^{\,2}\hat{y}_j^{\,2} \quad .$$

This suggests the estimate f of ϕ as given by equation (3.1a). Then if

$$G = \frac{(\sum_i z_i \hat{y}_i^2)^2}{\sum_{ij} q_{ij} \hat{y}_i^2 \hat{y}_j^2} ,$$

and W = (residual SS − G), G/W is approximately distributed as an F with one and (ν-1) degrees of freedom and is Tukey's one degree of freedom test for nonadditivity (Tukey, 1949).

For the two-way classification as defined in Section 1.2,

$$f = \frac{\frac{2}{rc}(\sum z_{ij})(c\bar{y}_{i.})(r\bar{y}_{.j})}{\frac{4}{rc}(SS-R)(SS-C)} ,$$

or

$$f = \frac{\sum z_{ij}\bar{y}_{i.}\bar{y}_{.j}}{2(SS-R)(SS-C)} .$$

If removable nonadditivity is found, often

$$P = 1 - 2f\bar{y}$$

will indicate the proper power transformation (Tukey, 1957). P equal to zero is to be interpreted as the logarithmetic transformation.

3.2 Testing Independence of the Mean and Variance.

In order to measure the dependence of variability upon the level of measurement, one may define a linear regression coefficient of the (z_i^2) on the $(\hat{y}_i)$. Anscombe (1961) suggests

$$\text{(3.2a)} \qquad h = \frac{\sum_i z_i^2(\hat{y}_i - \bar{Y})}{\sum_{ij}(q_{ij})^2(\hat{y}_i - \bar{Y})(\hat{y}_j - \bar{Y})S^2} = \frac{T}{HS^2} ,$$

where (z_i), $(\hat{y}_i)$ and (q_{ij}) are as defined in Section 1.2 and

$$\bar{Y} = \frac{\sum_i q_{ii} y_i}{\nu} \quad .$$

Assume that $(y_i - \mu_i)$ is distributed as $N(0,\sigma_i^2)$, where σ_i^2 is proportional to $\exp(\chi\mu_i)$. Then the h statistic (3.2a) is an estimate of χ .

Motivation for the h statistic comes from the fact that

$$E\left(\sum_i z_i^2 (\hat{y}_i - \bar{Y})\right) \simeq \sum_{ij} (q_{ij})^2 (\mu_i - \bar{\mu})(\mu_j - \bar{\mu})\chi\sigma^2 + 0(\chi^2) \quad ,$$

where

$$\bar{\mu} = \frac{\sum_i q_{ii} \mu_i}{\nu} \quad ,$$

and where $0(\chi^2)$ is the regression coefficient of z_i^2 on $(\hat{y}_i - \bar{Y})$, or $E\left(z_i^2 \middle| \hat{y}_i - \bar{Y}\right) = E(z_i^2) + 0(\chi^2)$. See Anscombe (1961) for derivation of the h statistic.

If one desires to test the significance of the deviation of χ from zero (equivalently, whether the variance is dependent on the level of measurement), a mean and standard error for the conditional distribution of h, given the fitted values, is needed.

Anscombe (1961) shows that

$$E\left(h \middle| \hat{y}_i\right) = 0 \quad ,$$

and

$$\mathrm{Var}\left(h \middle| \hat{y}_i\right) = \frac{2\nu}{(\nu+2)\ H} \quad .$$

For a one-way classification,

$$H = \frac{(k-1)SS(\text{means})}{k} ,$$

where k is the number of observations per treatment.

For a two-way classification,

$$H = \frac{(r-2)(c-1)}{rc} (SS\text{-}R) + \frac{(r-1)(c-2)}{rc} (SS\text{-}C) .$$

For an $r \times r$ latin square with k observations per cell

$$H = \frac{(r-2)(r-3)}{r^2} (SS\text{-}C + SS\text{-}R + SS\text{-}L) .$$

Anscombe and Tukey (1963) suggest that an overestimate of H can be obtained by totaling the sum of squares for all fitted values (treatments, blocks, etc.)

An estimate of the power transformation necessary for making the error variance constant ($\chi = 0$) is given by

$$P = 1 - (1/2)h \bar{Y} .$$

For the two-way classification, $\bar{Y} = \bar{y}$. See Tukey (1957) for a complete discussion of such transformations.

3.3 Normality.

If one desires to test hypotheses using the results from a least-squares analysis, the assumption of normality (1.2i) is necessary for many procedures. If sufficiently large samples are taken, most sampling situations produce approximately normal distributions. However, the consequences of grossly non-normal distributions are severe enough to merit a test for normality.

In assessing the shape of a distribution, the skewness coefficient γ_1 and the kurtosis coefficient γ_2 play an important part. They are defined as

$$\gamma_1 = \frac{1}{\sigma^3} E(y_i - \mu_i)^3, \qquad \gamma_2 = \frac{1}{\sigma^4} E\left[(y_i - \mu_i)^4\right] - 3 \quad .$$

The statistics g_1 and g_2, both functions of the residuals, will be proposed as estimates of γ_1 and γ_2. They are due to Anscombe (1961).

By definition,

$$E(\sum_i z_i{}^3) = E\left[\sum_i(\sum_j q_{ij}y_i)^3\right] \quad .$$

From (3.1c),

$$E(\sum_i z_i{}^3) = E\left\{\sum_i\left[\sum_j q_{ij}(y_j - \mu_j)\right]^3\right\} \quad .$$

If (1.2g) and (1.2h) hold and since the errors $(y_j - \mu_j)$ are independent with zero means, it follows that

$$E(\sum_i z_i{}^3) = \sum_{ij}(q_{ij})^3\gamma_1\sigma^3 \quad .$$

It is clear that

$$g_1 = \frac{\sum_i z_i{}^3}{\sum_{ij}(q_{ij})^3 s^3}$$

is a possible estimator for γ_1.

The sampling distribution of g_1 under the null hypothesis that $\gamma_1 = 0$ is determined by

$$E(g_1) = 0, \qquad \text{Var}\ (g_1) = \frac{E\left((\sum_i z_i^2)^2\right)}{\left(\sum_{ij} (q_{ij})^3\right)^2 E(s^6)} \quad .$$

But,

$$E(s^6) = \frac{(\nu+2)(\nu+4)\sigma^6}{\nu^2} \quad .$$

Using (1.2e) - (1.2i) and the fact that $\underline{Q}$ is p.s.d., Anscombe (1961) shows that

$$\text{Var}(g_1) = \frac{6n^2}{\nu(\nu+2)(\nu+4)\left(1 + (n-1)\bar{\rho}^{\,3}\right)} \quad ,$$

where $\bar{\rho}^{\,3}$ denotes the average cubed correlation between pairs (z_i, z_j).

For a two-way classification (r and c>2),

$$g_1 = \frac{n\sum z_{ij}^{\ 3}}{\nu(r-2)(c-2)s^3} \ , \qquad \text{Var}\ (g_1) = \frac{6n\nu}{(\nu+2)(\nu+4)(r-2)(c-2)} \quad .$$

Next, consider the estimation of γ_2.

$$E(\sum_i z_i^{\,4}) = E\left(\sum_i (\sum_j q_{ij} y_i)^4\right) \quad .$$

From (3.1c), one finds that

$$E(\sum_i z_i^{\,4}) = E\left\{\sum_i \left[\sum_j q_{ij} (y_j - \mu_j)\right]^4\right\} \quad .$$

If (1.2g) and (1.2h) hold and since the errors $(y_j - \mu_j)$ are independent with zero means, it follows that

$$E(\sum_i z_i{}^4) = \sum_{ij} (q_{ij}{}^4) E(y_j - \mu_j)^4 + T \quad ,$$

where

$$T = \sum_i \left[\sum_{j \neq j'} (q_{ij})^2 (q_{ij'})^2 E(y_j - \mu_j)^2 E(y_{j'} - \mu_{j'})^2 \right] \quad .$$

From (3.1c) and since $\underline{Q}$ is idempotent, it can be shown that (Anscombe, 1961)

$$E(\sum_i z_i{}^4) = \sum_{ij} (q_{ij})^4 \gamma_2 \sigma^4 + 3 \sum_i (q_{ii})^2 \sigma^4 \quad .$$

Finally, the estimate of γ_2 is

$$g_2 = \left(\frac{\sum_i z_i{}^4}{s^4} - \frac{3\nu \sum_i (q_{ii})^2}{\nu + 2} \right) D^{-1} \quad ,$$

where

$$D = \sum_{ij} (q_{ij})^4 - \frac{3 \left(\sum_i (q_{ii})^2 \right)^2}{\nu(\nu + 2)} \quad .$$

Under the null hypothesis that $\gamma_2 = 0$ and if (1.2g) and (1.2h) hold, Anscombe (1961) showed that

$$E(g_2) = 0 \quad ,$$

and

$$\mathrm{Var}\ (g_2) = \frac{24 n^2 \nu^2}{\left\{ \nu(\nu+2) \left(1 + (n-1) \bar{\rho}^{\,4} \right) - 3n \right\} (\nu+4)(\nu+6)} \quad .$$

For the two-way classification one finds that

$$\mathrm{Var}\ (g_2) = \frac{24n^2\nu^2}{\{(\nu+2)(r^2-3r+3)(c^2 - 3c+3)-3\nu^2\}(\nu+4)(\nu+6)} \ .$$

It is of interest to note that Srivastava (1959) has found that skewness has little effect on the power of the F-test in a one-way analysis of variance, but that kurtosis does.

3.4 Reduction to Two-Way Tables.

Anscombe and Tukey (1963) have found that, "data having a more complex pattern can often be studied effectively, so far as residuals go, in a two-way classification." This is the motivation for listing the results in the previous sections for a two-way classification.

Data in a 2^N factorial experiment is especially amenable to residual analysis based on a two-way structure. Consider, for example, a 2^6 factorial experiment in one replication. If one assumes that very few, if any, of the interactions are likely to be important, he can arrange the data in an 8×8 table. All combinations of three of the factors will constitute the rows while the columns will be all combinations of the other three factors. Other groupings of the factors, such as a 4×16 table, are also possible. For the residual analysis to be effective, the following two things are required:

(1) There must be at least one factor in each group which has a large "effect".

(2) Any extremely large interactions must be assigned to within-group, rather than between-group, status.

4. Modified Residuals

Most of the procedures presented in the first three sections of this paper are sensitive to extreme values (outliers). If extreme values are present, the rejection of a certain hypothesis may be due solely to their presence, rather than the hypothesis under consideration being false. Many papers have been written on this subject. Only one procedure, which relies heavily on the residuals, will be presented in this paper. Also, many types of complex systematic behavior may be present in residuals obtained from a conventional analysis which fits row, column, and grand means. This should be taken into consideration before residuals worthy of detailed study can be obtained. One procedure for dealing with a moderate amount of this complex systematic behavior is the vacuum cleaner. Since the vacuum cleaner is also sensitive to extreme values, a procedure called FUNOR-FUNOM usually precedes the vacuum cleaner.

4.1 Modifying Extreme Values.

Anscombe and Tukey (1963) propose that it is often advantageous to have available a definite rejection rule for outliers. This, of course, will sometimes lead to rejection of perfectly good observations. Since this will tend to increase the average error variance, one can regard the percentage increase in the error variance as the premium charged for protection against bad observations. Realizing that bad observations lead to a decrease in precision, one can see that this is a reasonable approach.

The procedure suggested by Anscombe and Tukey (1963) is to reject any observation which has a residual greater than DS. S is the square root of the error mean square and, if $\alpha = .05$,

$$D = 3.06(1 - \frac{1.85}{\nu})(\frac{\nu}{n})^{1/2} \quad ,$$

where ν and n are as defined in Section 1.2.

Many authors suggest that it is often more reasonable to modify an observation than to reject it completely {see Jeffreys (1948), de Finetti (1961), and Anscombe and Tukey (1963)}. Tukey (1962) presents a procedure called FUNOR-FUNOM which accomplishes the necessary modification.

A less complicated procedure, first suggested by C. P. Winsor (see Dixon, 1960), is to replace any suspected observation by the nearest nonsuspect observation. Therefore, any observation containing a residual which exceeds DS by only a small amount can be reduced by the difference in its residual and the largest nonsuspect residual. Any observation that has a residual exceeding DS by a large amount will be rejected completely unless there is definite reason for believing this observation is correct. If an observation is rejected or modified, the standard error S should be recomputed before testing another potential extreme value.

4.2 The Vacuum Cleaner.

The vacuum cleaner is a procedure presented by Tukey (1962) which yields residuals freer of systematic structures than the conventional mean fitting. Since some aberrations often simulate others, it is occasionally desirable to apply this procedure before the procedures in Section 3 are applied. It should be noted, however, that if this is done, the procedures in Section 3 will have to be modified.

Tukey (1962) first constructed a subprocedure which regressed the values in each row of a two-way table on the values in a separately given row, next regressed the values in each column on the values in a separately given column, thirdly, regressed the whole table on the two-way array consisting of all products of an entry in the separate row with an entry in the separately given column, and finally, subtracts this last regression from each of the other two. The result is a four-part breakdown:

> (original values) = (dual regression) + (deviations of row regression from dual regression) + (deviations of column regression from dual regression) + (residuals).

The basic vacuum cleaner becomes the application of this subprocedure twice. The first application removes row, column, and grand means. The second application removes row-by-row regression upon "column mean minus grand mean" and column-by-column regression on "row mean minus grand mean". The dual regression of the second application is the one degree of freedom for non-additivity.

5. An Example

The following example (Table 5.a) presented by Gill (1966) will serve to illustrate some of the techniques presented in the previous sections. Results of the conventional analysis are given in Table 5.b and the resulting residuals in Table 5.c.

Table 5.a

Oil Content of Flaxseed Inoculated at Different Stages of Growth

Treatment	Block					
	1	2	3	4	$y_{i.}$	$\bar{y}_{i.}$
Seedling	4.4	5.9	6.0	4.1	20.4	5.1
Early Bloom	3.3	1.9	4.9	7.1	17.2	4.3
Full Bloom	4.4	4.0	4.5	3.1	16.0	4.0
Full (1/100)	6.8	6.6	7.0	6.4	26.8	6.7
Ripening	6.3	4.9	5.9	7.1	24.2	6.05
Uninoculated	6.4	7.3	7.7	6.7	28.1	7.02
$y_{.j}$	31.6	30.6	36.0	34.5	132.7	
$\bar{y}_{.j}$	5.27	5.1	6.0	5.75		$\bar{y}=5.53$

Table 5.b

Analysis of Variance

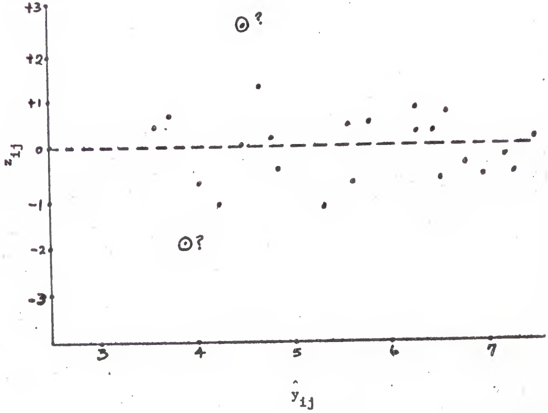
Source	dF	SS	MS	F-test
Blocks	3	3.14	1.05	
Treatments	5	31.65	6.33	4.83 $p(<.01)$
Residual	15	19.72	1.31	
Total	23	54.51		

Table 5.c

Residuals						
i, j	y_{ij}	$\hat{y}_{ij}$	z_{ij}	rank=k	$a_{k/n}$	$(z_{ij}-\overset{\prime}{z})/a_{k/n}$
1, 1	4.4	4.84	−0.44	16	−0.37	1.40
1, 2	5.9	4.67	+1.23	2	+1.50	0.77
1, 3	6.0	5.57	+0.43	7 tie	+0.54*	0.65
1, 4	4.1	5.32	−1.22	23	−1.50	0.87
2, 1	3.3	4.04	−0.74	21	−1.04	0.79
2, 2	1.9	3.87	−1.97	24	−1.95	1.05
2, 3	4.9	4.77	+0.13	12	+0.05	1.00
2, 4	7.1	4.52	+2.58	1	+1.95	1.28
3, 1	4.4	3.74	+0.66	5	+0.88	0.66
3, 2	4.0	3.57	+0.43	7 tie	+0.54*	0.65
3, 3	4.5	4.47	+0.03	13	−0.05	1.00
3, 4	3.1	4.22	−1.12	22	−1.24	0.97
4, 1	6.8	6.44	+0.36	9	+0.37	0.76
4, 2	6.6	6.27	+0.33	10	+0.26	0.96
4, 3	7.0	7.17	−0.17	14	−0.16	1.56
4, 4	6.4	6.92	−0.52	17	−0.48	1.25
5, 1	6.3	5.79	+0.51	6	+0.73	0.59
5, 2	4.9	5.62	−0.72	20	−0.88	0.91
5, 3	5.9	6.52	−0.62	19	−0.73	0.96
5, 4	7.1	6.27	+0.83	3	+1.24	0.60
6, 1	6.4	6.76	−0.36	15	−0.26	1.69
6, 2	7.3	6.59	+0.71	4	+1.04	0.60
6, 3	7.7	7.49	+0.21	11	+0.16	0.81
6, 4	6.7	7.24	−0.54	18	−0.60	1.03
	$\bar{y}$=5.53		$\overset{\prime}{z}$=(+0.13+0.03)/2 =+0.08		*ave. of 7 & 8 (+0.60+0.48)/2=0.54	

The first step is to plot the residuals against the fitted values (see Section 2.1). Figure 5.d gives this plot.

Figure 5.d



The following results are noted:

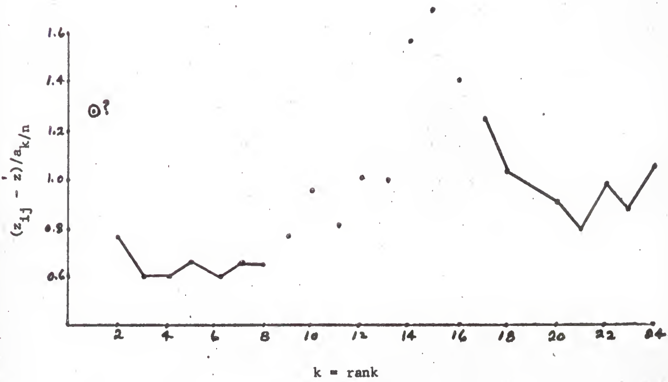
(1) Two "potential" outliers are Early Bloom, Blocks 2 and 4.

(2) Excluding these two values, the variance apparently is independent of the level of response.

(3) The assumption of additivity (1.2g) appears to be met since the plot is linear.

The second step is FUNOP (2.2). This result is given in Figure 5.e.

Figure 5.e



The following results are noted:

(1) One "potential" outlier is Early Bloom, Block 4.

(2) The upturned tails give some indication of kurtosis; the higher plot on the right indicates some skewness. This causes one to believe that the assumption of normality may not be met.

The next step is to apply the procedure suggested in Section 4.1. The following calculations are noted:

$$(1)\quad S = (1.31)^{1/2} = 1.145$$

$$(2)\quad D = 3.06\ (1 - \frac{1.85}{\nu})(\nu/n)^{1/2}$$

$$= 3.06\ (1 - \frac{1.85}{15})(15/24)^{1/2}$$

$$= 2.12$$

$$(3)\quad DS = 2.43$$

Since the residual for Early Bloom in Block 4 (2.58) exceeds DS by a small amount, it will be modified. The nearest nonsuspect residual is 1.23. The modified value is then 7.1 - (2.58-1.23) = 5.75. A second two-way analysis, using the modified value, yields Table 5.f.

Table 5.f

Analysis of Variance after Modification of First Outlier

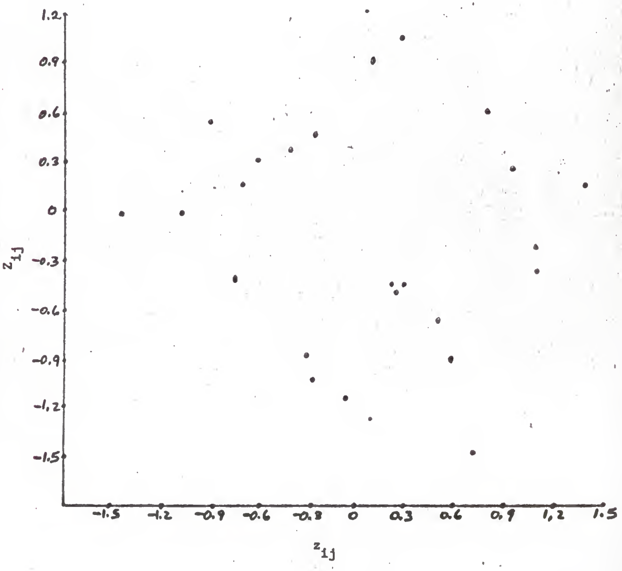
Source	dF	SS	MS	F-test
Blocks	3	2.77	0.92	
Treatments	5	35.35	7.07	27.19 p(<.01)
Residual	15	3.89	0.26	
Total	23	42.01		

This result is quite startling. By simply modifying one observation, the residual mean square has been reduced by 80% of its original value. When the sole purpose of an experiment is to test for treatment differences, the validity of using such a procedure on a small data set is questionable.

However, for the purpose of testing the assumptions necessary for doing an analysis of variance, this procedure seems to be adequate. By repeating the method of Section 4.1 the other suspected outlier (Early Bloom, Block 2) is also modified (from 1.9 to 2.64). No other observations are modified as the procedure is repeated.

The next step is the second scatter plot (2.3). Figure 5.g gives the desired plot. The obvious lack of a linear trend indicates that the assumption of additivity is probably satisfied.

Figure 5.g



Next, consider the test statistics presented in Section 3. The analysis of variance after modification of both outliers is given in Table 5.h.

Table 5.h

Analysis of Variance after Modification of Both Outliers

Source	dF	SS	MS	F-test	
Blocks	3	2.29	0.76	39.76	p(<.01)
Treatments	5	41.77	8.35		
Residual	15	3.19	0.21		
Total	23	47.25			

To test for removable non-additivity, consider the following:

$$f = \frac{B}{A} = \frac{\sum z_{ij}\bar{y}_{i.}\bar{y}_{.j}}{2(SS-R)(SS-C)} = \frac{-1.43}{12.02} \quad ,$$

where the z_{ij} are the residuals after modification of the suspect observations. Also, consider the following:

$$\frac{B^2}{A} = 0.17 \quad ,$$

and the remainder sum of squares is (3.19 - 0.17) = 3.02. Consequently, $F(1,14) = 0.17/3.02 = 0.06$, so it appears that the assumption of additivity is met.

To test for dependence of variability upon the level of response, consider

$$\begin{aligned} H &= \left((r-2)(c-1)(SS\text{-}R)/rc\right) + \left((r-1)(c-2)(SS\text{-}C)/rc\right) \\ &= \left((4)(3)(41.77)/24\right) + \left((5)(2)(2.29)/24\right) \\ &= 21.84 \end{aligned}$$

$$\begin{aligned} h &= \left(\sum z_{ij}{}^{2}(\hat{y}_{ij} - \bar{y})\right)/S^{2}H \\ &= -6.335/(3.19)(21.84) \\ &= -0.09 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(h) &= 2\nu/\left((\nu+2)H\right) \\ &= 2(15)/\left((17)(21.84)\right) \\ &= 0.08 \end{aligned}$$

To test H_o: $\chi = 0$,

$$t(15dF) = h/\left(\mathrm{Var}(h)\right)^{1/2} = -0.32$$

There is no good evidence for dependence of variability on the level of response.

To check on the assumption of normality, consider

$$\begin{aligned} g_1 &= n\sum z_{ij}{}^{3}/\left(\nu(r-2)(c-2)S^{3}\right) \\ &= (20)(-3.034)/\left((15)(4)(2)(3.19)^{3/2}\right) \\ &= -0.089 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(g_1) &= 6n\nu/\left((\nu+2)(\nu+4)(r-2)(c-2)\right) \\ &= (6)(20)(15)/\left((15+2)(15+4)(6-2)(4-2)\right) \\ &= 0.697 \end{aligned}$$

$$
\begin{aligned}
g_2 &= \frac{n^2\nu^2\left\{\left[\left(\frac{\nu+2}{\nu}\right)\left(\sum z_{ij}{}^4\right)/(\nu S^2)^2\right] - 3/n\right\}}{(\nu+2)(r^2-3r+3)(c^2-3c+3)-3\nu^2} \\
&= \frac{(20)^2(15)^2\left\{\left[\left(\frac{15+2}{15}\right)(13.67)/\left(15(3.19)\right)^2\right] - 3/20\right\}}{(15+2)(6^2-18+3)(4^2-12+3)-3(15)^2} \\
&= 7.07
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(g_2) &= 24n^2\nu^2/\left\{\left((\nu+2)(r^2-3r+3)(c^2-3c+3)-3\nu^2\right)(\nu+4)(\nu+6)\right\} \\
&= 24(20)^2(15)^2/(1824)(19)(21) \\
&= 2.97
\end{aligned}
$$

To test H_o: $\gamma_1 = 0$ (no skewness present),

$$t(15\mathrm{dF}) = g_1/\left(\mathrm{Var}(g_1)\right)^{1/2} = -0.107 \ .$$

There is no evidence of skewness.

To test H_o: $\gamma_2 = 0$ (no kurtosis present), use

$$t(15\mathrm{dF}) = g_2/\left(\mathrm{Var}(g_2)\right)^{1/2} = 4.11 \ .$$

There may be kurtosis present. This would need confirmation with other data sets. Since kurtosis is a flattening out of a distribution, its presence would make true treatment differences more difficult to detect.

6. Conclusion

The techniques in this paper provide useful insight into the validity of the assumptions in a classical least-squares analysis. They are fairly easy to apply, but have no sweeping statistical properties. Not all possible information about these assumptions can be obtained from the residuals and the results obtained must be qualified in the sense that it usually requires several data sets or large amounts of data to determine such things as the shape of the distribution with any degree of confidence. One should also remember that most of these techniques are sensitive to outliers and that it is very easy for one violation to act like another. One should allow these techniques to guide his judgment but not dictate courses of action. If, however, gross violations of these assumptions are detected, one should consider either using nonparametric techniques (see Lohrding, 1966) or making a transformation of the data.

ACKNOWLEDGEMENTS

I am very much indebted to Dr. A. M. Feyerherm for his guidance in preparing this report. In addition, I am also indebted to Arthur D. Dayton for providing a valuable reference. I am very grateful to my wife, Karen, for her encouragement and help in editing and typing my notes.

REFERENCES

Anscombe, F. J. (1961). Examination of Residuals. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press), *vol. 1*, 1-36.

Anscombe, F. J. and Tukey, J. W. (1963). The Examination and Analysis of Residuals. *Technometrics 5*:141-160.

de Finetti, Bruno. (1961). The Bayesian Approach to the Rejection of Outliers. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press), *vol. 1*, 1-36.

Dixon, W. J. (1960). Simplified Estimation from Censored Normal Samples. *The Ann. of Math. Stat. 31*:385-391.

Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrics 3*:1-21.

Fisher, R. A. and Yates, Frank. (1963). *Statistical Tables for Biological, Agriculture and Medical Research*. Oliver and Boyd, London.

Gill, J. L. (1966). The Use of Residuals in the Study of Nonconformity of Data to a Two-way Classification Model. Unpublished mimeograph, Michigan State University.

Graybill, Franklin A. (1961). *An Introduction to Linear Statistical Models*, McGraw-Hill Book Company, Inc., New York.

Jeffreys, H. (1948). *Theory of Probability*. University Press, Oxford.

Lohrding, R. K. (1966). Nonparametric Analogues of Analysis of Variance, Master's report, Kansas State University.

Rao, C. Radhakrishna. (1965). *Linear Statistical Inference and Its Applications*. John Wiley and Sons, Inc., New York.

Srivastava, A. B. L. (1959). Effect of Non-normality on the Power of the Analysis of Variance Test. *Biometrica 46*:114-122.

Tukey, J. W. (1949). One Degree of Freedom for Nonadditivity. *Biometrics 5*: 232-242.

Tukey, J. W. (1957). On the Comparative Anatomy of Transformations. *The Ann. of Math. Stat. 28*:602-632.

Tukey, J. W. (1962). The Future of Data Analysis. *The Ann. of Math. Stat. 33*:1-67.

THE ANALYSIS OF RESIDUALS

by

BRUMMETT JERALD McCLENDON

B. A., Texas Technological College, 1965

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics and
Statistical Laboratory

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

ABSTRACT

This paper provides methods of testing the validity of the assumptions necessary for using the classical least-squares procedures. These procedures are based on the residuals--the observed value minus the fitted value. Both graphic techniques and test statistics are presented for detecting removable nonadditivity, dependence of variance on the level of measurement, and nonnormality.

The first graphic technique is a simple plot of the residuals against the fitted values. It is the first step in gaining insight into the appropriateness of the least-squares techniques. The second graphic technique presented, called FUNOP, uses the residuals in a somewhat more complicated way to indicate the shape of the underlying distribution. A third graphic technique, called the second scatter plot, is presented for detecting removable nonadditivity.

The test statistics are presented for detecting removable nonadditivity, dependence of variance on the level of measurement, skewness, and kurtosis (f, h, g_1, and g_2, respectively). The mean and variance for each statistic is found under the null hypothesis that the corresponding population parameter which the statistic estimates is zero.

The above procedures are presented only as a guide to better data analysis, not as completely valid techniques with sweeping optimum properties. If applied along with good statistical judgment, they are excellent guides for handling important problems faced in data analysis.