A PROBLEM IN RESIDUAL ANALYSES.
AN M.S. THESIS PROBLEM.

BU-629-M* W. T. Federer October 1977

Abstract

Analyses of absolute values of residuals have been suggested for searching for discrepant blocks, treatments or other values. Owing to the correlation structure of residuals, questions arise concerning certain significance tests on residuals. Some suggested analyses of variance, degrees of freedom, computational formulas for sums of squares, and variance ratios are discussed. A simulation experiment is proposed for comparing the distribution of ratios of variances with percentage points of an F-distribution.

---

* Biometrics Unit Mimeo Series, Cornell University.

A PROBLEM IN RESIDUAL ANALYSES.

AN M.S. THESIS PROBLEM.

BU-629-M W. T. Federer October 1977

1. Introduction

The process of arriving at a "correct model" should cost the experimenter or analyst several degrees of freedom, but none of the standard textbooks consider this situation. They suggest finding a function of the responses which produces homoscedasticity, or additivity, but make no comment about loss in degrees of freedom. J. W. Tukey and coworkers have suggested that the consideration of a function of the data costs about 2/3rds of a degree of freedom for each function considered. In the following, "we have assumed" that the "correct" function of the responses has been obtained in previous experiments, and that we want to search for an outlying treatment or an outlying block. In this situation we are not searching for a "correct model", but we are assuming that we have such a model. Therefore, there need be no loss in degrees of freedom. The data to use for simulation would be standard normal deviates to which 5 has been added to each value producing a population with a mean of 5 and a variance of one.

H. C. Kirton, Department of Agriculture, New South Wales, Sydney, Australia, has suggested that after one has a "correct model", he should proceed as follows in searching for a discrepant treatment or block:

(i) compute estimated residuals,

(ii) use absolute values of the residuals, and

(iii) perform a standard ANOVA (the same one as used for the responses) and/or multiple comparisons procedures on the absolute values of the estimated residuals.

If the model is "correct", then the null hypothesis should be true for all categories in the ANOVA except for the residuals of the absolute values of residuals. That is, the expected value for each F-test is one. If the null hypothesis (hypotheses) is(are) not true, then this procedure can be used to pinpoint discrepant treatments, blocks, etc., in the experiment.

C. P. Quesenberry, Department of Statistics, North Carolina State University, has indicated to the author that an analysis of residuals has several problems related to the correlation structure among the estimated residuals. Owing to this fact, one should not immediately follow the above procedure of Kirton without first investigating the problem to some extent. Therefore, it is proposed that a computer simulation study be performed for a number of orthogonal experiment designs for varying numbers of treatments and replications of treatments. These numbers could vary between 2 and 10, say, for both treatments and replications. Some suggestions are made for correcting for degrees of freedom, for divisors in sums of squares, and for computing F-statistics. The "goodness" of these suggestions, as well as a comparison of the F-statistics with tabulated percentage points of the F-distribution, should be made.

2. Completely Randomized Design

Given the usual linear model, the solution for the ij^{th} residual is

$$Y_{ij} - \bar{y}_{i.} = e_{ij} \qquad (2.1)$$

where Y_{ij} is the response for the j^{th} observation ($j=1,2,\cdots,r$) on the i^{th} treatment and $\bar{y}_{i.}$ is the arithmetic mean of all responses from treatment i ($i=1,2,\cdots,v$). Let us denote the absolute value of the residual e_{ij} by a_{ij}.

Then, an ANOVA on the absolute value of the residuals takes the form:

Source of variation	d.f.	Sum of squares	Mean Square
Total	f	$\sum_{i=1}^{v}\sum_{j=1}^{r} a_{ij}^2$	T
Correction for mean	f_c	$\left(\sum_i\sum_j a_{ij} = a_{..}\right)^2/f$	C
Among v treatments	f_t	$\dfrac{\sum_{i=1}^{v}\left(\sum_{j=1}^{r} a_{ij} = a_{i.}\right)^2}{rf_c} - \dfrac{a_{..}^2}{f}$	A
Within treatments	f_e	$\sum_{i=1}^{v}\sum_{j=1}^{r} a_{ij}^2 - \sum_i a_{i.}^2/rf_c$	W

Now we know that $\Sigma_{i=1}^{v}\Sigma_{j=1}^{r} e_{ij}^2 = \Sigma_{i=1}^{v}\Sigma_{j=1}^{r} a_{ij}^2$ is distributed as $\sigma^2\chi^2$ with $v(r-1) = f$ degrees of freedom under the usual normality and homoscedasticity assumptions. Therefore, the divisor for the correction for the mean $\bar{a}_{..}$ has a divisor of f instead of vr as in the ANOVA on responses Y_{ij}. Since the total sum of squares is $\sigma^2\chi^2$, one wonders if each of the parts might not also be proportional to a χ^2 with x degrees of freedom. To obtain the column of "degrees of freedom" in the above ANOVA table, it is suggested that one proceed as follows (in this way the parts add to the total):

"d.f." in ANOVA Table on Y_{ij}		"d.f." in ANOVA Table on a_{ij}
Total	rv	$f = v(r-1)$
Correction for mean	1	$f_c = v(r-1)/rv = 1 - \frac{1}{r}$
Treatments	v-1	$f_t = (v-1)f_c$
Remainder	v(r-1)	$f_e = f_c v(r-1)$

Likewise, it is suggested that the divisor for the correction term for the a_{ij}'s be f instead of rv as in the ANOVA on Y_{ij} and that the divisor for the treatment totals (or means) squared be rf_c instead of r as in the ANOVA on Y_{ij}. The F-statistics suggested are:

$$F(f_c, f_e) = C/W \qquad (2.2)$$

$$F(f_t, f_e) = A/W \quad . \qquad (2.3)$$

Of course, an analytic result is much more valuable than a result obtained by simulation. However, owing to the lack of analytic results on the above F-statistics for analyses of residuals, it is suggested that a simulation experiment be performed for a range of values of v and r from 2 to 10. The distribution of the values from (2.2) and (2.3) should then be compared at various percentage points for the corresponding tabulated F-values.

3. Randomized Complete Block Design

Here, again, we consider the standard textbook linear model with the usual normality and homoscedasticity assumptions. The estimated residual e_{ij} is:

$$Y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot} = e_{ij} \qquad (3.1)$$

where Y_{ij} is the response of treatment i in the j^{th} block, i=1,2,$\cdots$,v, j=1,2,$\cdots$,b, $\bar{y}_{i\cdot}$ is the i^{th} treatment mean, $\bar{y}_{\cdot j}$ is the j^{th} block mean, and $\bar{y}_{\cdot\cdot}$ is the mean of the bv observations. As before, let $|e_{ij}| = a_{ij}$. Then, an ANOVA on the a_{ij}'s takes the form:

Source of variation	d.f.	Sum of squares	Mean square
Total	$f=(b-1)(v-1)$	$\sum_{i=1}^{v}\sum_{j=1}^{b} a_{ij}^2$	T
Correction for mean	f_c	$\left(\sum_{i=1}^{v}\sum_{j=1}^{b} a_{ij} = a_{\cdot\cdot}\right)^2/f$	C
Among blocks	f_b	$\dfrac{\sum_{j=1}^{b}\left(\sum_{i=1}^{v} a_{ij} = a_{\cdot j}\right)^2}{vf_c} - \dfrac{a_{\cdot\cdot}^2}{f}$	B
Among treatments	f_v	$\dfrac{\sum_{i=1}^{v}\left(\sum_{j=1}^{b} a_{ij} = a_{i\cdot}\right)^2}{rf_c} - \dfrac{a_{\cdot\cdot}^2}{f}$	V
Remainder	f_r	$\sum_{i=1}^{v}\sum_{j=1}^{b} a_{ij}^2 - \dfrac{\sum_{i=1}^{v} a_{i\cdot}^2}{rf_c} - \dfrac{\sum_{j=1}^{b} a_{\cdot j}^2}{vf_c} + \dfrac{a_{\cdot\cdot}^2}{f}$	R

where $f_c = (v-1)(b-1)/vb$, $f_b = f_c(b-1)$, $f_v = f_c(v-1)$, and $f_r = f_c(b-1)(v-1)$ $= (v-1)^2(r-1)^2/vr$. Then, $f_c + f_b + f_v + f_r = f$.

Now, Tf is distributed as $\sigma^2\chi^2$ with (b-1)(v-1) degrees of freedom. Therefore, a reasonable conjecture would be that the parts of $\Sigma_{i=1}^{v}\Sigma_{j=1}^{b} a_{ij}^2 = fT$ are

themselves proportional to a chi-square. (Note that the mean squares here are obtained by dividing the sum of squares by the degrees of freedom.) As before, we may compute ratios of mean squares and compare them with tabulated F-values at the various percentage points for the degrees of freedom listed. The F-ratios computed could be:

$$F(f_c, f_r) = C/R \tag{3.2}$$

$$F(f_b, f_r) = B/R \tag{3.3}$$

$$F(f_v, f_r) = V/R \ . \tag{3.4}$$

The simulation experiment could involve values of b and v from 3 to 10.

4. Simple Change-over Design and Latin Square Design

The standard textbook linear model with its associated normality and homoscedasticity assumptions is used here also. In this case the residuals are estimated by

$$Y_{hij} - \bar{y}_{h..} - \bar{y}_{.i.} - \bar{y}_{..j} + 2\bar{y}_{...} = e_{hij} \ , \tag{4.1}$$

where the i^{th} treatment in the j^{th} column and h^{th} row has response Y_{hij}, $\bar{y}_{h..}$, $\bar{y}_{.i.}$, and $\bar{y}_{..j}$ are the arithmetic means for the h^{th} row, i^{th} treatment, and j^{th} column respectively, $\bar{y}_{...}$ is the mean of bk = vr observations, h=1,2,$\cdots$,k, i=1,2,$\cdots$,v(=k), j=1,2,$\cdots$,b, and r = bk/v is the number of replications on treatment i . In the latin square design v = b = k = r. Let $|e_{hij}| = a_{hij}$. An ANOVA on the residuals would be of the form:

Source of variation	d.f.	Sum of squares	Mean square
Total	$f=(b-1)(k-2)$	$\sum_{hij} \sigma^2_{hij}$	T
Correction for mean	f_c	$\left(\sum_{hij} a_{hij} = a_{...}\right)^2/f$	C
Columns or blocks	f_b	$\frac{\sum_j \left(\sum_{hi} a_{hij}=a_{..j}\right)^2}{f_c k} - \frac{a^2_{...}}{f}$	B
Rows	f_r	$\frac{\sum_h \left(\sum_{ij} a_{hij}=a_{h..}\right)^2}{f_c b} - \frac{a^2_{...}}{f}$	R
Treatments	f_v	$\frac{\sum_i \left(\sum_{hi} a_{hij}=a_{.i.}\right)^2}{rf_c} - \frac{a^2_{...}}{f}$	V
Remainder	f_e	$\sum_{hij} a^2_{hij} - \frac{\sum_h a^2_{h..}}{bf_c} - \frac{\sum_i a^2_{.i.}}{rf_c} - \frac{\sum_j a^2_{..j}}{kf_c} + \frac{2a^2_{...}}{f}$	E

where $f_c = (b-1)(k-2)/bk$, $f_b = f_c(b-1)$, $f_r = f_c(k-1)$, $f_v = f_c(v-1)$, and $f_e = f_c(b-1)(k-2)$.

fT is distributed as $\sigma^2\chi^2$ with $(b-1)(k-2)$ degrees of freedom. The sums of squares which make up fT could be expected to be proportional to chi-squares. Then, the following ratios of variances would follow F distributions:

$$F(f_c, f_e) = C/E \tag{4.2}$$

$$F(f_b, f_e) = B/E \tag{4.3}$$

$$F(f_r, f_e) = R/E \tag{4.4}$$

$$F(f_v, f_e) = V/E \tag{4.5}$$

From a simulation experiment, these ratios could be compared with tabulated values of the F-distribution at the various percentage points. The simulation values could range from $v = k = 2, \cdots, 10$ and $b = 4$ to 10.

5. Multiple Comparisons and Other Designs

Instead of, or in addition to, computing F-statistics, one could use multiple comparisons procedures of various types and compare to experimental results with theoretical distributions. Likewise, other orthogonal or non-orthogonal experiment designs could be included in a simulation study. However, it is suggested that results from a completely randomized design, say, be studied thoroughly before extending a simulation procedure to other situations.