ANALYZING UNIFORM RESIDUALS FOR MISSPECIFICATION OF A LINEAR REGRESSION MODEL: I. TESTING THE TOTAL MODEL AND IDENTIFYING OUTLIERS

C. P. Quesenberry*

1. INTRODUCTION

Consider the general linear regression model

$$\underset{\sim}{y}_n = X_n \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_n , \qquad (1.1)$$

where $\underset{\sim}{y}_n = (y_1, \cdots, y_n)'$ is a column vector of observed values,
X_n is an $(n \times p)$ matrix of known values of rank p,
$\underset{\sim}{\beta} = (\beta_1, \cdots, \beta_p)'$ is a column vector of parameters, and
$\underset{\sim}{\varepsilon}_n = (\underset{\sim}{\varepsilon}_1, \cdots, \varepsilon_n)'$ is a column vector of i.i.d. $N(0, \sigma^2)$ rv's.
This model is, of course, of great importance in applied statistics, and because of its importance there has developed in the last two decades a large body of literature that considers methods for studying the validity of the model assumptions for a given $\underset{\sim}{y}_n$ and X_n.

Much of this work involves analysis of the residuals from the least squares fitted model. Early papers that considered least squares residuals were Anscombe (1961) and Anscombe and Tukey (1963). Many other papers considering analyses of least

*C. P. Quesenberry is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27650

squares residuals have appeared. Material on least squares residuals analysis now appears in a number of books including Seber (1977), Daniel and Wood (1971), Theil (1971) and Barnett and Lewis (1978).

Although the analysis of least squares residuals can undoubtedly be helpful in studying specification errors in the above regression model in many problems, there remain some difficulties in the interpretation of these analyses due to the fact that these residuals are dependent and, in general, not identically distributed. A problem of some concern with least squares residuals is that they tend to be "supernormal," as has been mentioned by a number of writers including Gnanadesikan (1977), Weisberg (1980a, 1980b), and Quesenberry and Quesenberry (1982). See Quesenberry and Quesenberry (1982) for a definition of "supernormal."

To avoid distributional difficulties with least squares residuals, Theil (1965, 1968) defined a set of $n - p$ residuals which are called best linear unbiased scalar, or, BLUS, residuals. There is now a substantial literature on BLUS residuals that has appeared since Theil's first papers. See Theil (1971) for a summary account of the BLUS literature until that date, and references to contributions to this and related literature. The BLUS residual vector has $n - p$ components which are linear functions of $\underset{\sim}{y}_n$, and are defined so that these residuals are independent normal random variables with zero means and variances that are known multiples of σ^2 (unknown). This exact distribution theory makes the B1US residuals convenient for use in testing and otherwise analyzing the data for model misspecification. There are two points which we particularly note about BLUS residuals. First, the elements of $\underset{\sim}{y}_n$ (and rows of X_n) must be written down in some particular order to compute these residuals and, in general, if this order is changed then different values for the residuals (and subsequent test statistics, *etc*.) will be obtained. This does *not*, of course, in any way invalidate the distribution theory for the residuals or the analyses based on them. It does mean, however, that if the same residual analysis was performed for two different orderings of the components of $\underset{\sim}{y}_n$, that different decisions might be made, *i.e*., analyses based on BLUS residuals are not invariant under permutations of the observations, in general. This lack of invariance to permutations of the observations will also be observed below for recursive residuals, including those to be proposed in this paper. Finally, it is noted that while the least squares residual e_i corresponds exactly to the regression observation $(y_i, x_{i1}, \cdots, x_{ip})$, the BLUS residual $\hat{e}_i$ is not so clearly associated with a particular observation. This fact may reduce the usefulness of these residuals for detecting outliers in the data.

Hedayat and Robson (1970) have posed stepwise or recursive residuals for testing the homoscedasticity of variances hypothesis. These are the LS residuals obtained by successively fitting the regression model to the first $p+2$ observations, then the first $p+3$ observations, $\cdots$, the first n observations. These $n-p$ residuals are independent, and identify in a one-to-one association with the last $n-p$ observations. They also depend upon the order of observations, and by ordering the observations by a variable, such as time, this property can be used to construct a test to detect a variance changing with time. Brown, Durbin and Evans (1975) also have considered recursive residuals and used them to construct tests to decide if a particular aspect of a regression model is changing with time. The recursive residuals considered in these last two papers have independent $N(0, \sigma^2)$ distributions. They are in concept related to the uniform residuals which we shall consider in this paper. The uniform residuals which we consider are also recursive in nature and are independent uniform random variables on the unit interval (0, 1), *i.e.*, are $U(0, 1)$ rv's.

Recursive residuals are especially attractive for testing for specification errors related to a particular ordering of the observations, as, for example, for testing against the alternative hypothesis that the variance σ^2 is increasing, (decreasing) with time; or to identify a point in time when a catastrophic event has disrupted a linear regression relationship in other ways. However, since the uniform residuals we consider here have exact known distribution theory under the regression model, they can also be used to make tests for misspecification that is not necessarily related to a particular ordering of the observations. In this paper we shall consider using uniform residuals to make an onmibus test and analysis of the entire model (1.1), and we propose some techniques for detecting outliers among multiple regression observations.

In section 2 the uniform residuals are defined and certain properties are given. In section 3 some omnibus tests and graphical methods for the entire model will be considered. In section 4 some methods for detecting outliers among the observations are proposed and illustrated with numerical examples.

2. UNIFORM RESIDUALS

Throughout this section it will be assumed that the assumptions of model (1.1) hold. Following O'Reilly and Quesenberry (1973), O-Q, Example 4.3, let $\underset{\sim}{x}_j' = (x_{j1}, \cdots, x_{jp})$ denote the jth row of X_n, X_j denote the first j rows of X_n, and assume that X_{p+2} is of rank p. If $\underset{\sim}{t}_j = X_j'\underset{\sim}{y}_j$ and $S_j^2 = \underset{\sim}{y}_j'[I - (X_j(X_j'X_j)^{-1}X_j']\underset{\sim}{y}_j$, then the statistic $\underset{\sim}{T}_j = (\underset{\sim}{t}_j', S_j^2)'$ is complete and sufficient for the multivariate normal distribution of model (1.1). Also, put

$$A_j = \frac{(j - p - 1)^{\frac{1}{2}}[y_j - \underset{\sim}{x}_j'(X_j'X_j)^{-1}\underset{\sim}{t}_j]}{\{[1 - \underset{\sim}{x}_j'(X_j'X_j)^{-1}\underset{\sim}{x}_j]\, s_j^2 - [y_j - \underset{\sim}{x}_j'(X_j'X_j)^{-1}\underset{\sim}{t}_j]^2\}^{\frac{1}{2}}} \quad . \qquad (2.1)$$

and

$$u_{j-p-1} = G_{j-p-1}(A_j)\ ; j = p + 2, \cdots, n\ ; \qquad (2.2)$$

for $G_\nu(\cdot)$ a Student-t df with ν degrees of freedom. Then from O-Q, Example 4.3 and Corallory 2.1, it follows that the vector $\underset{\sim}{u} = (u_1, \cdots, u_{n-p-1})'$ is a set of i.i.d. uniform random variables on the unit interval (0, 1). Also, the quantity A_j is a Student-t rv with $(j - p - 1)$ degrees of freedom, and the A_j's are independent.

We shall call $\underset{\sim}{u}$ the vector of <u>uniform</u> <u>residuals</u>. It is natural to call these quantities residuals because of the following considerations. First, we point out that the only method of estimation involved in any way in the distribution theory arguments leading to the above results is minimum variance unbiased (MVU) estimation, and it was only incidentally involved because certain conditional distribution functions are, in fact, MVU estimating distribution functions. In particular, neither least squares nor maximum likelihood estimation played any role in the derivation. Nevertheless, the quantity in square brackets in the numerator of A_j in (2.1) is the usual j^{th} least squares residual from the least squares fitting of the regression function $X_j\underset{\sim}{\beta}$ using $y_1, \cdots, y_j$. Moreover, the denominator of this quantity is the square root of an independent chi-squared rv with (j - p - 1) degrees of freedom, and therefore the A_j's themselves can reasonably be called "Studentized" residuals. Other very similar "Studentized" residuals were considered by Stefansky (1971, 1972).

We consider next two important distributional properties of uniform residuals, viz., that they are independent of the complete sufficient statistic for the model parameters, and that they possess important invariance properties.

Theorem 2.1 (Independence) Under the assumptions of model (1.1), the vector $\underset{\sim}{u}$ of uniform residuals is independent of the complete sufficient statistic $\underset{\sim}{T}_j = (\underset{\sim}{t}_j', S_j^2)'$; $j = p+2, \cdots, n$.

Proof. This result is immediate from the result of Basu (1955), and the fact that the distribution of $\underset{\sim}{u}_j$ is a constant function of the parameters $(\underset{\sim}{\beta}', \sigma^2)$.

Also, from (2.1) and (2.2) it can be seen that there is (almost surely) a one-to-one correspondence between the points $(u_1, \cdots, u_{n-p-1})$ and $(y_{p+2}, \cdots, y_n)$ in the conditional space for fixed $\underset{\sim}{T}_n' = (\underset{\sim}{t}_n', S_n^2)$. Thus by an argument similar to that given in Quesenberry and Starbuck (1976), Q-S, we see that under the model assumptions (1.1) the two vectors $\underset{\sim}{T}_n$ and $\underset{\sim}{u}$ essentially partition the information in $\underset{\sim}{y}_n$ into the information in $\underset{\sim}{T}_n$ available for making inferences about the parameters $(\underset{\sim}{\beta}', \sigma^2)$, and the independent information in the uniform residuals vector $\underset{\sim}{u}$ available for making inferences about the functional form of the distribution of the observations $\underset{\sim}{y}_n$. It can be shown, for example, that for any test of specification in the model (1.1), not involving parameter values, a test of equal power can be based on the $\underset{\sim}{u}$ vector alone. Thus a model analysis based only on $\underset{\sim}{u}$ ignores no relevant information. With these results in mind we shall consider analyzing $\underset{\sim}{u}$ to detect specification errors in the next two sections. The independence of $\underset{\sim}{T}$ and $\underset{\sim}{u}$ can be exploited as previously suggested in Q-S in a different context to assess overall error rates from misspecification and parametric inference errors under the model (1.1).

To complete this section we now show that the statistic $\underset{\sim}{u}$ possesses certain natural invariance properties. If a is a positive constant and $\underset{\sim}{c}_p = (c_1, \cdots, c_p)'$, a vector of constants, then put $\underset{\sim}{y}_n^* = a\underset{\sim}{y}_n + X_n\underset{\sim}{C}_p$.

<u>Theorem</u> <u>2.2</u> (Invariance) If $\underset{\sim}{y}_n$ satisfies the model (1.1), then (i) $\underset{\sim}{y}_n^*$ also satisfies the model assumptions (1.1), (ii) the uniform residuals vector computed from $\underset{\sim}{y}_n$ and $\underline{y}_n^*$ are the same.

<u>Sketch of proof</u>: Part (i) is immediate, and (ii) follows easily by substituting $\underset{\sim}{y}_n^*$ for $\underset{\sim}{y}_n$ in (2.1).

3. ANALYZING THE ENTIRE MODEL

The uniform residuals of (2.2) are i.i.d. uniform random variables when all of the assumptions of (1.1) are valid. If any of the assumptions of (1.1) are not correct, then the u_j's will not in general be i.i.d. uniform random variables on the unit interval (0, 1). In particular, if the ε_j's are not independent, if they are not identically distributed, or if their distribution is not of the normal form, then the u_j's will generally not be i.i.d. U(0, 1) rv's. Now, since analyses based on the normal linear regression model are valid only if all of the model assumptions hold, we would like to analyze the data $\underset{\sim}{y}_n$ so as to expose any misspecification in the model. In this section we shall consider some general omnibus tests for the model (2.1), and supplement these tests with graphical analyses.

Two Omnibus Tests for Uniformity

When the u values of (2.2) have been obtained, we shall wish to decide whether these are observed values of i.i.d. uniform random variables. That is, we wish to test the surrogate simple uniformity goodness-of-fit hypothesis H_0: $u_1, \cdots, u_N$ are i.i.d. U(0, 1) rv's, vs the alternative H_a: Negation of H_0.

A test for this problem is called a test of (simple) uniformity. There is a large number of goodness-of-fit tests which can be used to make tests for uniformity. Reasonably extensive power studies of tests for uniformity have been made by Quesenberry and Miller (1977), Q-M, and by Miller and Quesenberry (1979), M-Q. These papers and, also, Miller and Quesenberry (1974), review much of the literature in this area of goodness-of-fit testing. Based upon the results in these papers, we recommend two particular goodness-of-fit tests for uniformity. Our choice of tests for uniformity are the Neyman smooth test and Watson's U^2 test. (See M-Q, pp. 287, 8.)

Neyman (1937) posed a statistic designed to have high power for testing uniformity against certain classes of alternative distributions on the unit interval, (see M-Q, and Kendall and Stuart (1961); p. 444). The test is defined as follows. The Legendre polynomials π_r are given by; for r = 0, 1, 2, 3, 4; and $0 \le y \le 1$;

$$\pi_0(y) = 1,\ \pi_1(y) = \sqrt{12}(y-\tfrac{1}{2}),$$
$$\pi_2(y) = \sqrt{5}[6(y-\tfrac{1}{2})^2-\tfrac{1}{2}],\ \pi_3(y) = \sqrt{7}[20(y-\tfrac{1}{2})^3 - 3(y-\tfrac{1}{2})], \quad (3.1)$$
$$\pi_4(y) = 210(y-\tfrac{1}{2})^4 - 45(y-\tfrac{1}{2})^2 + 9/8.$$

Then put

$$t_r = \sum_{j=1}^{N} \pi_r(u_j) \text{ for } r = 1, 2, 3, 4,$$

and the Neyman smooth test rejects for large values of the statistic

$$p_4^2 = (1/N)\ (t_1^2 + t_2^2 + t_3^2 + t_4^2). \quad (3.2)$$

Neyman showed that when $u_1, \cdots, u_N$ are i.i.d. uniform rv's that p_4^2 has a limiting $\chi^2(4)$ distribution. Computations of upper .1, .05, and .01 percentage points in M-Q indicate that this approximation is reasonably good even for N as small as ten, or even smaller in some cases. This approximation is particularly convenient because it can be used to determine the observed significance level or p-value of the test for uniformity. Thus in order to obtain an overall assessment of the validity of the multiple linear regression model of (1.1), we compute the u_j's from (2.3), the value of p_4^2 from (3.2) and then evaluate

$$\text{NS4\_PV} \equiv \text{p-value} = P\{\chi^2(4) > p_4^2\} . \tag{3.3}$$

For $0 < \alpha < 1$, if NS4_PV $\leq \alpha$, then of course we reject the model (2.1) at the α level. In practice, we compute NS4_PV and view it as a general coefficient of validity of the regression assumption.

In addition to the Neyman smooth statistic p_4^2, we shall also frequently compute a second omnibus test for uniformity. Watson (1961) posed the statistic

$$U^2 = \frac{1}{12N} + \sum_{j=1}^{N} \{(2j-1)/2N - U_{(j)}\}^2 - N(\bar{u} - .5)^2 , \tag{3.4}$$

where $u_{(1)} \leq \cdots \leq u_{(N)}$ are the ordered u values and $\bar{u}$ is their mean. See, also, Q-M, section 2.5. This test is made by rejecting for large values of the following modified version of U^2,

$$U^2_{MOD} = [U^2 - 1/(10 \cdot N) + 1/(10 \cdot N^2)][1 + 0.8/N] \tag{3.5}$$

Stephens (1970) proposed using U^2_{MOD} rather than U^2 because its upper percentage points are approximately constant in N. The approximate critical values of U^2_{MOD}, due to Stephens, are given in Table 3.1.

TABLE 3.1

Approximate Critical Values of U^2_{MOD}

α	.10	.05	.01
Critical Value	.152	.187	.267

The accuracy of these values is adequate at the .10 and .05 levels for $N \geq 4$ and at the .01 level for $N \geq 9$.

<u>Graphing Techniques</u>

In addition to the general model tests described above, and the special purpose tests that will be given in the next section, it is sometimes helpful in spotting specification anomalies to plot the uniform residuals in various ways. Under the null hypothesis that the model of (1.1) is correct, the j^{th} order statistic $u_{(j)}$ is a beta random variable $B(j, N-j+1)$, and has expectation $E(u_{(j)}) = j/(N+1)$ for $j = 1, \cdots, N$. Thus if the points $(U_{(j)}, j/[N+1])$ are plotted, they should tend to follow the line $g(u) = u$, $0 \leq u \leq 1$. Plots of this type were made in Quesenberry, Whitaker and Dickens (1976). Quesenberry and Hales (1980) gave graphs which are helpful in interpreting these plots.

The uniform residuals can also be used in graphical procedures similar to those often used for least squares and related residuals. The residuals are plotted against one of the explanatory variables or an extraneous variable (such as time, perhaps) to look for trends. The uniform residuals will always fall in the interval (0, 1), and the patterns shown in these residuals are interpreted similarly to patterns in least squares and other residuals. When the model is valid the u_j's should tend to form a uniform band between the lines $u = 0$ and $u = 1$. The number of possible patterns that correspond to particular types of specification errors is large. The types of misspecification that lead to a particular pattern of points can sometimes be readily deduced by recalling the nature of the transformations in (2.1) and (2.2), viz., that A_j is a Studentized recursive residual and u_j is obtained by making a probability integral transformation of A, using the appropriate Student-t distribution function. Some of these plots are illustrated in the following example and in the examples of the next section.

Example 3.1 We consider the appropriateness of the multiple regression model (1.1) for the data given in Ostle (1954), p. 220. The y variable is the value of farm land per acre on January 1, 1920 in each of 25 Iowa counties, x_1 is the average corn yield per acre for the ten preceding years, x_2 is the percentage of farm land in small grain, etc. When these data are analyzed in the order given the last eighteen observations and their uniform residuals are given in Table 3.2.

Table 3.2 near here

TABLE 3.2

ANALYSIS OF OSTLE DATA

Y	X0	X1	X2	X3	X4	X5	U_VALUE
104	1	31	9	119	20	20	0.686550
141	1	36	13	106	53	27	0.716749
208	1	34	17	137	59	40	0.200092
115	1	30	18	136	40	19	0.349688
271	1	40	23	135	95	31	0.266479
163	1	37	14	98	41	25	0.991838
193	1	41	13	122	80	28	0.640635
203	1	38	24	173	52	31	0.034247
279	1	38	31	182	71	35	0.534078
179	1	34	16	124	43	26	0.933121
244	1	45	19	138	60	34	0.602859
165	1	34	20	148	52	30	0.105762
257	1	40	30	164	49	38	0.261714
252	1	41	22	96	39	35	0.961027
280	1	42	21	132	54	41	0.819332
167	1	35	16	96	41	23	0.886295
168	1	33	18	118	38	24	0.826376
115	1	36	18	113	41	21	0.029471

STATISTICS:

P_4^2	NS4_PV	U_{MOD}^2	$P_R(13)$	$P_L(25)$
2.389	.665	.063	.137	.416

The value NS4_PV = .665 and comparison of U^2_{MOD} with the values in Table 3.1 lead us to accept the regression model (1.1) for this data. Also, the p values of $P_R(13) \doteq .14$ and $P_L(25) = .42$ for the largest and smallest order statistics indicate that there are no outliers in this data, as will be explained in the next section.

We have plotted the ordered uniform residuals vs their expectations in Figure 3.1. This graph shows no anomalous behavior and supports the conclusion to accept the model stated above. As a further check on the model we have plotted the uniform residuals vs each of the explanatory variables. The graph of u vs x_1 is shown in Figure 3.2. This graph shows no anomalous trends in the residuals. The graphs for the other four explanatory variables are not shown here but they display similar regular behavior of the residuals.

Figure 3.1 near here

Figure 3.2 near here
----

4. DETECTING OUTLIERS

The problem of detecting one or more outliers among the observations of a regression data set is one of considerable practical importance. See, for example, the statement by Amscombe and Tukey (1963), p. 146. Outliers are data points

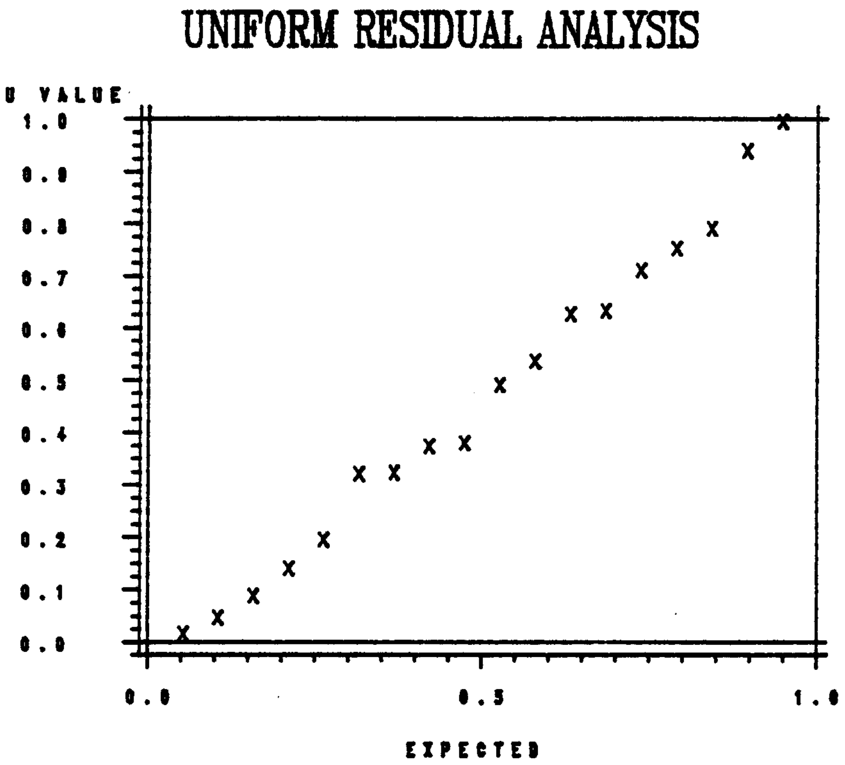

Figure 3.1: Ordered Residuals vs Expectations for Ostle Data

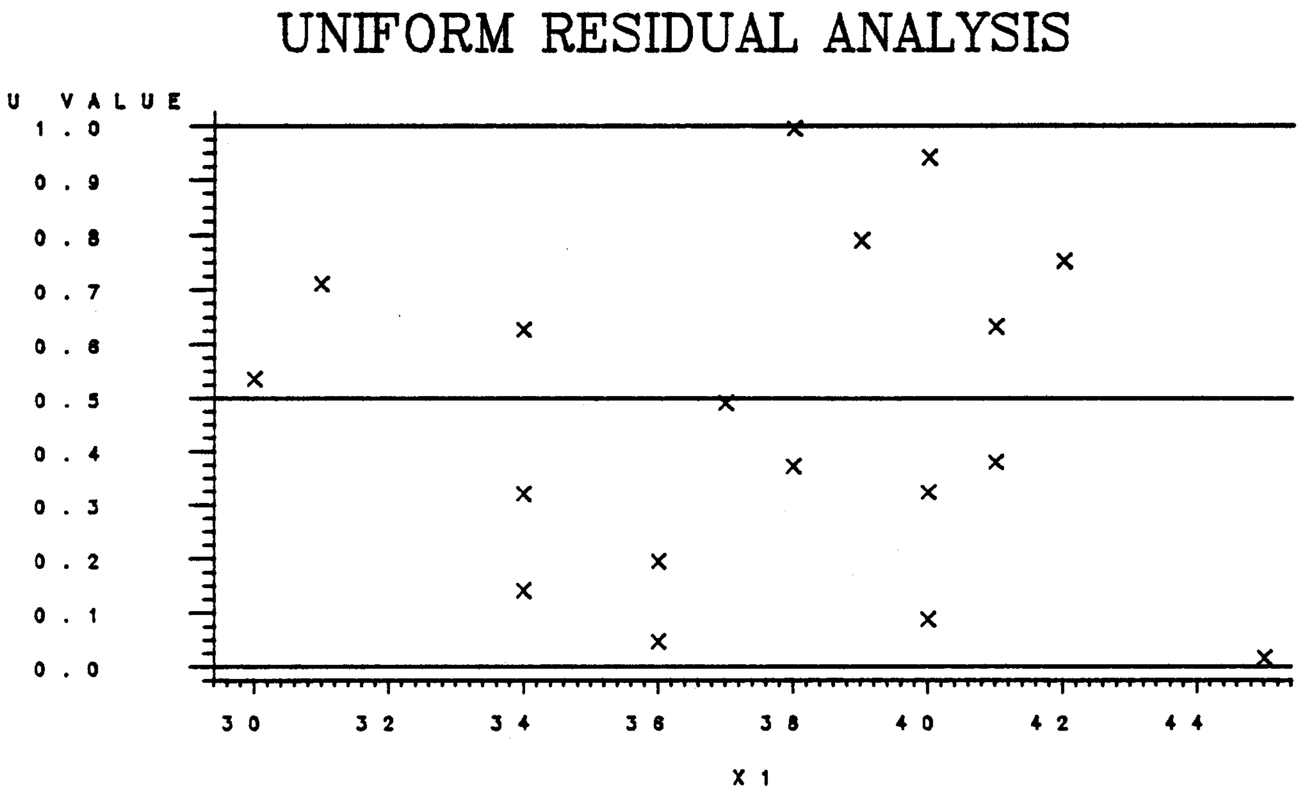

Figure 3.2: Residuals vs x_1 for Ostle Data

$(y, x_1, \cdots, x_p)$ that do not follow the data pattern indicated by the regression model (1.1). The presence of outliers may imply a failure of the regression model in some one or more of its attributes. Outliers can be due to an error distribution with tails thicker than the normal, to slippage in the means $X_n \underset{\sim}{\beta}$ that might arise from many causes, such as accidents with the data, or the model changing (slipping) with time, etc. We shall here be concerned only with the task of spotting outliers, but note in passing that it will sometimes be the case with real data that outliers themselves are especially important.

The exact distribution theory associated with $u_1, \cdots, u_n$; $N = n - p - 1$; is particularly useful for detecting outliers. If we could anticipate the number of outliers with A_j of (2.1) positive and the number with A_j negative, then we could use the i.i.d. $U(0, 1)$ distribution of the uniform residuals to design a test for this pattern of outliers. In practice, however, we usually do not have such information available. Thus in the following we pose an "omnibus" outlier detection scheme which we have tried on a number of examples and which appears to work well. We shall first describe our method and then report the results of applying it to some regression data sets.

With the observations given in some particular order, we compute the $N = n - p - 1$ uniform residuals as indicated in Table 4.1.

TABLE 4.1

Observations and Uniform Residuals

Obs. No.	Observation	Residual
1	$y_1, \quad x_{1,1}, \quad \ldots, x_{1,p}$	
⋮	⋮ ⋮ ⋮	⋮
p	$y_p, \quad x_{p,1}, \quad \ldots, x_{p,p}$	
p + 1	$y_{p+1}, x_{p+1,1}, \ldots, x_{p+1,p}$	
p + 2	$y_{p+2}, x_{p+2,1}, \ldots, x_{p+2,p}$	u_1
⋮	⋮ ⋮ ⋮	⋮
n	$y_n, \quad x_{n,1}, \quad \ldots, x_{n,p}$	u_N

From the N residuals in Table 4.1 we select the smallest and largest order statistics, say $u_{(1)}$ and $u_{(N)}$, respectively. We shall declare the observation associated with $u_{(1)}$ a negative or left outlier if $u_{(1)}$ is too small and the observation associated with $u_{(N)}$ a positive or right outlier if $U_{(N)}$ is too large. Let P_L denote the p value or observed significance level for testing that $u_{(1)}$ is too small, and similarly define P_R as the p value for testing that $u_{(N)}$ is too large. Then it is readily shown that under (1.1) and the distribution theory for uniform residuals that

$$P_L = 1 - (1 - u_{(1)})^N, \; P_R = 1 - u_{(N)}^N \, . \qquad (4.1)$$

For any $i = 1, \cdots, N$; put $P_L(i) = 1 - (1 - u_i)^N$ and $P_R(i) = 1 - U_i^N$. Then we shall declare observation i a left outlier at level α if $P_L(i) < \alpha$ and a right outlier at level α if $P_R(i) < \alpha$.

If the above procedure deletes at least one observation as an outlier we shall then proceed as follows. Delete the outliers from the observations and cyclically permute the remaining obserbations to place the first $p + 1$ observations in the last $p + 1$ positions of the data set. Then compute a new set of residuals and repeat the procedure above on this set of residuals. Continue making passes through the data until no observations are eliminated. It is well to bear in mind that due to the recursive nature of the transformations in (2.1), that the power for detecting an outlying observation that is near the end of the data set is somewhat better than for earlier observations. With this in mind we usually note "suspicious" observations as ones appearing early in the data set with $P_L(i)$ or $P_R(i)$ less than, say, 2α. If an observation is suspicious, we will verify its status by cyclically permuting this observation to the last position, recomputing its uniform residual and p values P_L and P_R.

We next illustrate these procedures with numerical examples. The values P_R and P_L given in Table 3.2 for the Ostle data are the p values for $u_{(18)}$ and $u_{(1)}$, respectively. Since $P_L \doteq .42$ and $P_R = .14$, we do not suspect outliers in this data.

Example 4.1 In this example we consider the regression data given in Brownlee (1960), sec. 13.12; and which was recently studied in detail in Daniel and Wood (1971), Chap. 5. The data and uniform residuals in the order given in Brownlee are given in Table 4.2. Using a nominal $\alpha = .05$ level, we see that observation 21 is a left outlier. The value of $P_R(15) = .966$ is, of course, not significant. Next we delete observation 21, cyclically permute the observations, and obtain on the second pass the results summarized in Table 4.3

Table 4.2 near here

TABLE 4.3

Brownlee Data, 2nd, 3rd, and 4th Passes

2nd Pass

OBS: 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 1 2 3 4 5

$P_R(1) = .0201$, $P_R(4) = .0395$

3rd Pass

OBS: 11 12 13 14 15 16 17 18 19 20 2 3 5 6 7 8 9 10

$P_R(3) = .0928$

4th Pass

OBS: 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 2 3

$P_R(3) = .0074$

TABLE 4.2

BROWNLEE DATA AND UNIFORM RESIDUALS, FIRST PASS

OBS.	Y	X0	X1	X2	X3	U_VALUE
1	42	1	80	27	89	—
2	37	1	80	27	88	—
3	37	1	75	25	90	—
4	28	1	62	24	87	—
5	18	1	62	22	87	—
6	18	1	62	23	87	0.078305
7	19	1	62	24	93	0.063449
8	20	1	62	24	93	0.456826
9	15	1	58	23	87	0.282464
10	14	1	58	18	80	0.613737
11	14	1	58	18	89	~~0.675124~~
12	13	1	58	17	88	0.537633
13	11	1	58	18	82	0.226155
~~14~~	~~12~~	~~1~~	~~58~~	~~19~~	~~93~~	0.362358
15	8	1	50	18	89	0.809920
16	7	1	50	18	86	0.640881
17	8	1	50	19	72	0.514957
~~18~~	~~8~~	~~1~~	~~50~~	~~19~~	~~79~~	0.577734
19	9	1	50	20	80	~~0.601348~~
20	15	1	56	20	82	0.702272
21	15	1	70	20	91	0.002119

STATISTICS:

NS4-PV	$P_L(21)$	$P_R(15)$
.126	.033	.966

Thus we eliminate observations 1 and 4, and cyclically permute the data to obtain on the third pass the results in Table 4.3. Since observation 3 gives a right p value of .0928, we permute it to the last position to obtain the results in Table 4.3 for the 4th pass. Since $P_R(3) = .0074$ we declare observation 3 an outlier, also. No other observation gives a value of either P_L or P_R that is near significance. Thus we have declared observations 1, 3, 4, and 21 to be outliers.

Next, since our outlier detection scheme depends upon the original order of the observations we ran the entire procedure with the order of observations reversed. The results are summarized in Table 4.4

Table 4.4 near here

Note that the algorithm rejects exactly the same set of outliers as for the data in its original order. Moreover, in their detailed study Daniel and Wood also identified this same set of outliers.

5. SUMMARY AND DISCUSSION

We have proposed analyzing uniform residuals to detect misspecification in a normal linear regression model. It follows from the conditional probability integral transformations theory introduced in O'Reilly and Quesenberry (1973) that these quantities are i.i.d. U(0, 1) rv's. In section 2

TABLE 4.4

Analysis of Brownlee Data in Reverse Order

Pass	Order of Observations	
	Base	Residuals Computed
1	21 20 19 18 17	16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
	$P_R(4) = .0043$, $P_L(2) = .9998$, Reject obs. 4 as a right outlier.	
2	16 15 14 13 12	11 10 9 8 7 6 5 3 2 1 21 20 19 18 17
	$P_R(3) = .0175$, $P_L(21) = .0069$, Reject obs. 3 and 21.	
3	11 10 9 8 7	6 5 2 1 20 19 18 17 16 15 14 13 12
	$P_R(1) = .0321$, $P_L(13) = .4122$, Reject obs. 1	
4	6 5 2 20 19	18 17 16 15 14 13 12 11 10 9 8 7
	$P_R(12) = .8174$, $P_L(18) = .5227$ No more outliers.	

SUMMARY: Observations 1, 3, 4, and 21 rejected.

it is shown that these residuals are independent of the sufficient complete statistics for the model parameters, and that they possess important natural invariance properties.

The Neyman smooth test and Watson's U^2 test are suggested as omnibus tests for the overall model. The Neyman smooth test is particularly convenient because it allows an easy evaluation of its p-value, which is essentially a model specification coefficient. The uniform residuals can be plotted against either the explanatory variables or other variables to attempt to discover anomolous trends in the data. Also, a plot of the ordered residuals against their expectations serves as a further check on the overall model.

The uniform residuals are particularly convenient to use in searching for outliers. Their exact distribution theory can be used to evaluate the observed significance levels of extreme order statistics. We have suggested a particular algorithm for detecting outlying observations. The procedures suggested for analyzing models are illustrated on two data sets.

REFERENCES

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

ANSCOMBE, F. J. and TUKEY, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141-160.

BARNETT, V. and LEWIS, T. (1978), "*Outliers in Statistical Data*," New York: John Wiley and Sons, Inc.

BASU, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhya*, 15, 377-380.

BROWN, R. L., DURBIN, J., and EVANS, J. M. (1975), "Techniques for Testing the Constancy of Regression Relationships Over Time," *Journal of the Royal Statistical Society*, B, 37, 149-192.

DANIEL, C. and WOOD, F. S. (1971), "*Fitting Equations to Data*," New York: Wiley-Interscience.

GNANADESIKAN, R. (1977), "*Methods for Statistical Data Analysis of Multivariate Observations*," New York: John Wiley and Sons, Inc.

HEDAYAT, A. and ROBSON, D. S. (1970), "Independent Stepwise Residuals for Testing Homoscedasticity," *Journal of the American Statistical Association*, 65, 1573-1581.

MILLER, F. L., JR. and QUESENBERRY, C. P. (1979), "Power Studies of Tests for Uniformity," *Communications in Statistics - Simulation and Computations*, B8(3), 271-290.

NEYMAN, J. (1937), "'Smooth' Test for Goodness of Fit," Skandinavisk Aktuarietidskrift, 20, 149-199.

O'REILLY, F. J. and QUESENBERRY, C. P. (1973), "The Conditional Probability Integral Transformation and Applications to Obtain Composite Chi-square Goodness-of-fit Tests, "Annals of Statistics, 1, 74-83.

OSTLE, B. (1954), "Statistics in Research," Ames, Iowa: The Iowa State University Press.

QUESENBERRY, C. P. and HALES, C. (1980), "Concentration Bands for Uniformity Plots," Journal of Statistical Computation and Simulation, 11, 41-53.

QUESENBERRY, C. P. and MILLER, F. L., JR. (1977), "Power Studies of Some Tests for Uniformity," Journal of Statistical Computation and Simulation, 5, 169-191.

QUESENBERRY, C. P. and QUESENBERRY, C., JR. (1982), "On the Distribution of Residuals from Fitted Parametric Models," Journal of Statistical Computation and Simulation," (To appear.)

QUESENBERRY, C. P. and STARBUCK, R. R. (1976), "On Optimal Tests for Separate Hypotheses and Conditional Probability Integral Transformations," Communications in Statistics, A5(6), 507-524.

QUESENBERRY, C. P., WHITAKER, T. B. and DICKENS, J. W. (1976), "On Testing Normality Using Several Samples: an Analysis of Peanut Aflatoxin Data," Biometrics, 32, 753-759.

SEBER, G. A. F. (1977), "Linear Regression Analysis," New York: John Wiley and Sons, Inc.

STEFANSKY, W. (1971), "Rejecting Outliers by Maximum Normed Residual," Annals of Mathematical Statistics, 42, 35-45.

STEFANSKY, W. (1972), "Rejecting Outliers in Factorial Designs," Technometrics, 14, 469-479.

STEPHENS, M. A. (1970), "Use of the Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics without Extensive Tables, "Journal of the Royal Statistical Society, B, 32, 115-122.

THEIL, H. (1965), "The Analysis of Disturbances in Regression Analysis," Journal of the American Statistical Association, 60, 1067-1079.

THEIL, H. (1968), "A Simplification of the BLUS Procedure for Analyzing Regression Disturbances, "Journal of the American Statistical Association, 63, 242-251.

THEIL, H. (1971), "Principles of Econometrics," New York: John Wiley and Sons, Inc.

WEISBERG, S. (1980a), "Comment (on a paper by White and MacDonald)," Journal of the American Statistical Association, 75, 28-31.

WEISBERG, S. (1980b), "Applied Regression Analysis," New York: John Wiley and Sons, Inc.