A General Definition of Residuals

By D. R. Cox and E. J. Snell

Imperial College

[Read at a Research Methods Meeting of the Society, March 13th, 1968, Professor R. L. Plackett in the Chair]

Summary

Residuals are usually defined in connection with linear models. Here a more general definition is given and some asymptotic properties found. Some illustrative examples are discussed, including a regression problem involving exponentially distributed errors and some problems concerning Poisson and binomially distributed observations.

1. Introduction

Residuals are now widely used to assess the adequacy of linear models; see Anscombe (1961) for a systematic discussion of significance tests based on residuals, and for references to earlier work. A second and closely related application of residuals is in time-series analysis, for example in examining the fit of an autoregressive model.

In the context of normal-theory linear models, the $n \times 1$ vector of random variables $\mathbf{Y}$ is assumed to have the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X}$ is a known matrix, $\boldsymbol{\beta}$ a vector of unknown parameters and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of unobserved random variables of zero mean, independently normally distributed with constant variance. If $\hat{\boldsymbol{\beta}}$ is the vector of least-squares estimates of $\boldsymbol{\beta}$, the residuals $\mathbf{R}^*$ are defined by

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{R}^*. \tag{2}$$

Provided that the number of parameters is small compared with n, most of the properties of $\mathbf{R}^*$ are nearly those of $\boldsymbol{\epsilon}$, i.e. $\mathbf{R}^*$ should have approximately the properties of a random sample from a normal distribution. In fact, $\mathbf{R}^*$ being linear in $\mathbf{Y}$, the random variable $\mathbf{R}^*$ has, under (1), a singular normal distribution and hence the properties of significance tests can be studied in some detail (Anscombe, 1961).

The main types of departure from the model (1) likely to be of importance are:

(i) the presence of outliers;

(ii) the relevance of a further factor, omitted from (1), detected by plotting the residuals against the levels of that factor;

(iii) non-linear regression on a factor already included in (1), detected by plotting the residuals against the levels of that factor and obtaining a curved relationship;

(iv) correlation between different ϵ_i's, for example between ϵ_i's adjacent in time, detected from scatter diagrams of suitable pairs of R_i^*'s, or possibly from a periodogram analysis of residuals;

(v) non-constancy of variance, detected by plotting residuals or squared residuals against factors thought to affect the variance, or against fitted values;

(vi) non-normality of the distribution of the ϵ_i's, detected by plotting the ordered residuals against the expected order statistics from a standard normal distribution (Pearson and Hartley, 1966, Table 28).

Corresponding to the graphical analyses suggested in points (i)–(vi), statistics can be constructed for formal tests of significance. The idea of inspecting residuals is very old, but the systematic calculation of residuals, particularly from extensive data, has become practicable only recently; their thorough graphical analysis as a routine is feasible only with a suitable computer graphical output device.

The examination of the adequacy of a model by such analyses may be contrasted with a more formal approach in which there is fitted:

either (a) a more general model containing one or more additional parameters and reducing to (1) for particular values of the new parameters;

or (b) a different family of models, adequacy of fit being assessed, say, by the maximum log likelihood achieved.

One example of (a) is the family of models considered by Box and Cox (1964), in which model (1) is considered as applying to an unknown power of the original observations. The advantages of the more formal techniques are that they have sensitive significance tests associated with them and that they are directly constructive in the sense that, if the initial model does not fit, a specific better-fitting model is obtained immediately from the analysis. On the other hand, analysis of residuals, especially by graphical techniques, does not require committal in advance to a particular family of alternative models. It will indicate the nature of a departure from the initial model, but not explicitly how to extend or replace the model. With very extensive data, significance testing is relatively unimportant and the types of departure that can be detected are more numerous than can be captured in advance in a few simple parametric models. It is in such applications that the analysis of residuals is likely to be most fruitful.

2. A More General Definition

The main object of the present paper is to give a more general definition of residuals and to illustrate some of its properties and applications. Consider a model expressing an observed vector random variable $\mathbf{Y}$ in terms of a vector $\boldsymbol{\beta}$ of unknown parameters and a vector $\boldsymbol{\epsilon}$ of independent and identically distributed unobserved random variables. More particularly we assume that each observation Y_i depends on only one of the ϵ's, so that we can write

$$Y_i = g_i(\boldsymbol{\beta}, \epsilon_i) \quad (i = 1, \ldots, n). \tag{3}$$

This assumption excludes applications to time series and also to component of variance problems in which several random variables enter into each observation. Models involving discrete distributions, such as the binomial and Poisson, are not in the first place included, because, for example, Poisson-distributed observations with different means cannot be expressed in terms of transformations of identically distributed observations. Later, however, in Sections 7–9, we extend the methods to deal with Poisson and binomial distributions.

To define residuals for (3), let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimate of $\boldsymbol{\beta}$ from $\mathbf{Y}$. It would be possible to work with other asymptotically efficient estimates, or even with inefficient estimates, but the details of Section 4 would be different.

Now suppose that the equation

$$Y_i = g_i(\hat{\boldsymbol{\beta}}, R_i) \tag{4}$$

has a unique solution for R_i, namely

$$R_i = h_i(Y_i, \hat{\boldsymbol{\beta}}). \tag{5}$$

Note that

$$\epsilon_i = h_i(Y_i, \boldsymbol{\beta}). \tag{6}$$

We take (5) as defining the residual corresponding to Y_i and the model (3); later we shall introduce a minor modification of (5) and then call R_i the crude residual.

Example 1. If (3) is the normal-theory linear model (1) with known variance, the residuals (5) are the same as those, R_i^*, of Section 1, equation (2). If the variance is an additional unknown parameter, then

$$R_i = R_i^* \{\sum R_j^{*2}/n\}^{-\frac{1}{2}},$$

and so R_i is essentially equivalent to R_i^*.

Example 2. Feigl and Zelen (1965) discussed some leukemia data in which for the ith individual Y_i is the time to death in weeks and x_i is the log of the initial white blood cell count. Feigl and Zelen considered primarily linear regression of Y_i on x_i with exponential errors, but here we work with the model, mentioned briefly by Feigl and Zelen,

$$Y_i = \beta_1 \exp\{\beta_2(x_i - \bar{x})\}\,\epsilon_i, \tag{7}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independently exponentially distributed with unit mean and $\bar{x} = \sum x_i/n$. The advantage of (7) over a linear regression model is that for all $\beta_1 > 0$ and all β_2, x_i, the random variable on the left-hand side of (7) is non-negative.

For this model, if $\hat{\beta}_1, \hat{\beta}_2$ are maximum likelihood estimates of β_1, β_2, then

$$Y_i = \hat{\beta}_1 \exp\{\hat{\beta}_2(x_i - \bar{x})\}\,R_i,$$

i.e.

$$R_i = [\hat{\beta}_1 \exp\{\hat{\beta}_2(x_i - \bar{x})\}]^{-1}\,Y_i. \tag{8}$$

Example 3. For some purposes it is convenient for analysing a random sample $Y_1, \ldots, Y_n$ from a Weibull distribution to write the model in the form

$$Y_i = (\beta_1 \epsilon_i)^{\beta_2}, \tag{9}$$

where again $\epsilon_1, \ldots, \epsilon_n$ have an exponential distribution of unit mean.

Some further examples are given in Section 10.

Often the number of parameters is small compared with the number of observations and the configuration is such that all relevant combinations of parameters are estimated with small standard error of order $n^{-\frac{1}{2}}$. Then a residual R_i will differ from ϵ_i by an amount of order $n^{-\frac{1}{2}}$ in probability and most statistical properties of the R's will differ little from those of the ϵ's.

We examine the properties of the R's more carefully in Section 4. This will be done by expanding $R_i - \epsilon_i$ in a Taylor series in terms of $\hat{\beta}_s - \beta_s$. We need some of the properties of maximum likelihood estimates, in particular an expression for their bias, and these are developed briefly in Section 3.

3. Some Properties of Maximum Likelihood Estimation

Bartlett (1952), incidentally to his study of large-sample confidence intervals, gave a simple expression for the bias to order n^{-1} of the maximum likelihood estimate from a single random sample, there being one unknown parameter. Haldane (1953) and Haldane and Smith (1956) further discussed asymptotic expansions for the properties of a maximum likelihood estimate dealing with random samples and one or two unknown parameters; for further discussion and extensions see Shenton and Wallington (1962) and Shenton and Bowman (1963).

With a single parameter and observations that are independent, but not necessarily identically distributed, the log likelihood is

$$L(\beta) = \sum \log p_j(Y_j, \beta),$$

where $p_j(Y_j, \beta)$ is the p.d.f. of Y_j. For a regular problem the maximum likelihood equation $L'(\hat{\beta}) = 0$ is to first order

$$L'(\beta) + (\hat{\beta} - \beta) L''(\beta) = 0. \tag{10}$$

Write

$$U^{(j)} = \frac{\partial \log p_j(Y_j, \beta)}{\partial \beta}, \quad V^{(j)} = \frac{\partial^2 \log p_j(Y_j, \beta)}{\partial \beta^2} \tag{11}$$

and replace $-L''(\beta)$ by its expectation

$$I = \sum E(-V^{(j)}),$$

where I is the total information in the sample.

We thus have the standard first-order expressions

$$\hat{\beta} - \beta = \frac{U^{(\cdot)}}{I}, \quad \text{var}(\hat{\beta}) = \frac{1}{I}, \tag{12}$$

where $U^{(\cdot)} = \sum U^{(j)}$, the dot indicating a sum over the sample.

To obtain a more refined answer, we replace (10) by the second-order equation

$$L'(\beta) + (\hat{\beta} - \beta) L''(\beta) + \tfrac{1}{2}(\hat{\beta} - \beta)^2 L'''(\beta) = 0. \tag{13}$$

Take expectations in (13), thereby obtaining

$$E(\hat{\beta} - \beta) E\{L''(\beta)\} + \text{cov}\{\hat{\beta} - \beta, L''(\beta)\} + \tfrac{1}{2} E(\hat{\beta} - \beta)^2 E\{L'''(\beta)\} + \text{cov}\{\tfrac{1}{2}(\hat{\beta} - \beta)^2, L'''(\beta)\} = 0. \tag{14}$$

Now, approximately, by (12)

$$\text{cov}\{\hat{\beta} - \beta, L''(\beta)\} = \frac{1}{I} \text{cov}(U^{(\cdot)}, V^{(\cdot)}) = \frac{J}{I}, \tag{15}$$

where

$$J = \sum E(U^{(j)} V^{(j)}).$$

Also if

$$W^{(j)} = \frac{\partial^3 \log p_j(Y_j, \beta)}{\partial \beta^3}, \quad K = E(W^{(\cdot)}),$$

then

$$E\{L'''(\beta)\} = K.$$

Note that I, J, K refer to a total over the sample and are of order n.

Finally, a calculation similar to (15) shows that the final term in (14) is $O(n^{-1})$, whence (14) gives (Bartlett, 1952), for the terms of order 1

$$-IE(\hat{\beta}-\beta)+\frac{J}{I}+\frac{K}{2I}=0,$$

i.e.

$$b\equiv E(\hat{\beta}-\beta)=\frac{1}{2I^2}(K+2J), \tag{16}$$

which is of order n^{-1}.

When there are parameters $\beta_1, \ldots, \beta_p$, we define

$$U_r^{(j)}=\frac{\partial \log p_j(Y_j,\boldsymbol{\beta})}{\partial \beta_r}, \quad V_{rs}^{(j)}=\frac{\partial^2 \log p_j(Y_j,\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s},$$

$$W_{rst}^{(j)}=\frac{\partial^3 \log p_j(Y_j,\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s \partial \beta_t},$$

$$I_{rs}=E(-V_{rs}^{(.)}), \quad J_{r,st}=E\{\textstyle\sum U_r^{(j)} V_{st}^{(j)}\}, \quad K_{rst}=E(W_{rst}^{(.)}). \tag{17}$$

Expansion of the equation

$$[\partial L/\partial \beta_r]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}=0$$

replaces the first-order equation (12) by

$$\hat{\beta}_r-\beta_r=I^{rs}U_s^{(.)}, \quad \operatorname{cov}(\hat{\beta}_r,\hat{\beta}_s)=I^{rs}. \tag{18}$$

The superscripts denote matrix inversion and the summation convention is applied to multiple suffices referring to parameter components.

The second-order equation (13) becomes

$$\frac{\partial L}{\partial \beta_r}+(\hat{\beta}_s-\beta_s)\frac{\partial^2 L}{\partial \beta_r \partial \beta_s}+\tfrac{1}{2}(\hat{\beta}_t-\beta_t)(\hat{\beta}_u-\beta_u)\frac{\partial^3 L}{\partial \beta_r \partial \beta_t \partial \beta_u}=0.$$

On taking expectations, we have that

$$E(\hat{\beta}_s-\beta_s)I_{rs}=\tfrac{1}{2}I^{tu}(K_{rtu}+2J_{t,ru}), \tag{19}$$

a set of simultaneous linear equations for the biases, with solution

$$b_s\equiv E(\hat{\beta}_s-\beta_s)=\tfrac{1}{2}I^{rs}I^{tu}(K_{rtu}+2J_{t,ru}). \tag{20}$$

In the right-hand side of (20), which is of order n^{-1}, consistent estimates of parameters can be inserted.

4. Further Properties of Residuals

It would be useful to know the joint distribution of the R_i's defined by (5). The distribution of any suggested test statistic could then be found, and the properties of graphical procedures evaluated. It is, however, not feasible to determine this joint distribution in general and, as a first step, we consider the expectations and covariances of the R_i's.

For this, expand (5) in series, obtaining to order n^{-1}

$$R_i = \epsilon_i + (\hat{\beta}_r - \beta_r) H_r^{(i)} + \tfrac{1}{2}(\hat{\beta}_r - \beta_r)(\hat{\beta}_s - \beta_s) H_{rs}^{(i)}, \tag{21}$$

where

$$H_r^{(i)} = \frac{\partial h_i(Y_i, \boldsymbol{\beta})}{\partial \beta_r}, \quad H_{rs}^{(i)} = \frac{\partial^2 h_i(Y_i, \boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}. \tag{22}$$

Thus

$$E(R_i) = E(\epsilon_i) + E(\hat{\beta}_r - \beta_r) E(H_r^{(i)}) + \operatorname{cov}(\hat{\beta}_r - \beta_r, H_r^{(i)}) + \tfrac{1}{2} E\{(\hat{\beta}_r - \beta_r)(\hat{\beta}_s - \beta_s)\} E(H_{rs}^{(i)}), \tag{23}$$

the neglected terms being $o(n^{-1})$.

In (23), the second term is given by (20). The fourth term is given to sufficient accuracy by the usual large-sample result (18) and to evaluate the third term, we have, again by (18),

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_r - \beta_r, H_r^{(i)}) &= E(I^{rs} U_s^{(.)} H_r^{(i)}) \\ &= I^{rs} E(U_s^{(i)} H_r^{(i)}); \end{aligned} \tag{24}$$

on the right-hand side the summation convention does not apply to the superscript i.

Thus, to order n^{-1},

$$\begin{aligned} E(R_i) &= E(\epsilon_i) + b_r E(H_r^{(i)}) + I^{rs} E(H_r^{(i)} U_s^{(i)} + \tfrac{1}{2} H_{rs}^{(i)}) \\ &= E(\epsilon_i) + a_i, \end{aligned} \tag{25}$$

say. In the same way, squaring (21) and taking expectations, we have to the same order that

$$E(R_i^2) = E(\epsilon_i^2) + 2b_r E(\epsilon_i H_r^{(i)}) + 2I^{rs} E(\epsilon_i H_r^{(i)} U_s^{(i)} + \tfrac{1}{2} H_r^{(i)} H_s^{(i)} + \tfrac{1}{2} \epsilon_i H_{rs}^{(i)}) \tag{26}$$

and that for $i \neq j$

$$E(R_i R_j) = \{E(\epsilon_i)\}^2 + (a_i + a_j) E(\epsilon_i) + I^{rs} E(\epsilon_i H_r^{(j)} U_s^{(i)} + \epsilon_j H_r^{(i)} U_s^{(j)} + H_r^{(i)} H_s^{(j)}). \tag{27}$$

We can summarize (25), (26), and (27) as follows:

$$\left.\begin{aligned} E(R_i) &= E(\epsilon_i) + a_i, \\ \operatorname{var}(R_i) &= \operatorname{var}(\epsilon_i) + c_{ii}, \\ \operatorname{cov}(R_i, R_j) &= c_{ij}, \end{aligned}\right\} \tag{28}$$

where a_i, c_{ii}, c_{ij} can be found in terms of the right-hand sides of (25), (26), and (27) and are of order n^{-1}. Note that the summation convention applies to the parameter suffices only and not to c_{ii}.

A simple example of these formulae is given in Section 6.

5. Application of Results of Section 4

There are broadly three ways in which the above results can be used.

Firstly, if all the correction terms in (28) are numerically small this gives some assurance that treating the R_i's as having the same statistical properties as the ϵ_i's is reasonable. If, say, the correction terms are small for all residuals except one, we might look at that residual separately and then omit it from the rest of the analysis.

Secondly, for particular types of test statistic, the results can be used to approximate to its distribution. Thus if we consider

$$T = \sum R_i z_i,$$

where the z_i's are constants, then

$$E(T) = E(\epsilon_i)\sum z_i + \sum a_i z_i,$$

$$\text{var}(T) = \text{var}(\epsilon_i)\sum z_i^2 + \sum z_i z_j c_{ij}.$$

A statistic used for testing possible dependence on z_i of $\text{var}(\epsilon_i)$ is

$$T' = \sum R_i^2(z_i - \bar{z}), \tag{29}$$

where $\bar{z} = \sum z_i/n$. Here the results of Section 4 give only that

$$E(T') = \sum c_{ii}(z_i - \bar{z}) + 2E(\epsilon_i)\sum a_i(z_i - \bar{z}).$$

In principle, it is possible to extend the arguments to obtain $E(R_i^4)$ and $E(R_i^2 R_j^2)$ and hence to reach an approximation to $\text{var}(T')$.

Thirdly, we may use (28) to define a modified residual R_i' having more nearly the properties of ϵ_i. How best to do this depends somewhat on the particular case, but one fairly general procedure is to write

$$R_i' = (1 + k_i) R_i + l_i, \tag{30}$$

where k_i, l_i are small constants. If we require that, to order n^{-1},

$$E(R_i') = E(\epsilon_i), \quad \text{var}(R_i') = \text{var}(\epsilon_i), \tag{31}$$

two equations determining k_i, l_i follow from (28).

A serious limitation to this discussion is that it applies only indirectly to the examination of distributional form, by plotting ordered residuals against the expected order statistics for the distributional form proposed for the ϵ_i's. For this we would like, in particular, to calculate $E(R_{(i)})$, where $R_{(i)}$ is the ith largest residual; alternatively, we would like to introduce modified residuals R_i'' such that

$$E(R_{(i)}'') = E(\epsilon_{(i)});$$

of course, it is easy to formulate even more ambitious aims. It is plausible, but not certain, that the modification (30), designed to produce residuals with approximately the same marginal mean and variance, is an advantage also from the point of view of plots to examine distributional form.

6. An Example

We consider further the data of Feigl and Zelen (1965). The model (7) is

$$Y_j = \beta_1 \exp\{\beta_2(x_j - \bar{x})\}\epsilon_j, \quad j = 1, 2, \ldots, n.$$

We first find the bias in $\hat{\beta}_1, \hat{\beta}_2$. Since ϵ_j is exponentially distributed with unit mean, we have, writing $(x_j - \bar{x}) = d_j$,

$$p_j(Y_j, \boldsymbol{\beta}) = [\exp\{-Y_j \exp(-\beta_2 d_j)/\beta_1\}]/\{\beta_1 \exp(\beta_2 d_j)\},$$

$$\log p_j(Y_j, \boldsymbol{\beta}) = -\{Y_j \exp(-\beta_2 d_j)/\beta_1\} - \log\beta_1 - \beta_2 d_j, \tag{32}$$

which, on differentiating and taking expectations, leads to

$$I_{rs} = \begin{cases} n/\beta_1^2, & r = s = 1, \\ \sum d_j^2, & r = s = 2, \\ 0, & r \neq s. \end{cases}$$

Equations (20) therefore give

$$b_1 = \tfrac{1}{2} I^{11}\{I^{11}(K_{111} + 2J_{1,11}) + I^{22}(K_{122} + 2J_{2,12})\},$$

with a similar equation for b_2. From (17) and (32), we have, without the summation convention,

$$\begin{aligned} J_{l,1t} &= E(\textstyle\sum U_l^{(j)} V_{1t}^{(j)}) \\ &= \begin{cases} -2n/\beta_1^3, & t = 1, \\ -\sum d_j^2/\beta_1, & t = 2, \end{cases} \end{aligned}$$

and

$$\begin{aligned} K_{1lt} &= E(W_{1lt}^{(.)}) \\ &= \begin{cases} 4n/\beta_1^3, & t = 1, \\ \sum d_j^2/\beta_1, & t = 2. \end{cases} \end{aligned}$$

Thus

$$b_1 = -\tfrac{1}{2}\beta_1/n.$$

A similar calculation gives

$$b_2 = -\tfrac{1}{2}\textstyle\sum d_j^3/(\sum d_j^2)^2.$$

To find $E(R_i)$ and $E(R_i^2)$, given by (25), (26), we write

$$h_i(Y_i, \boldsymbol{\beta}) = Y_i \exp(-\beta_2 d_i)/\beta_1$$

from which we evaluate $H_r^{(i)}$, etc., and obtain

$$\begin{aligned} E(R_i) &= 1 + \tfrac{1}{2}n^{-1} + \tfrac{1}{2}(d_i \textstyle\sum d_j^3 - d_i^2 \sum d_j^2)/(\sum d_j^2)^2 \\ &= 1 + a_i, \end{aligned} \tag{33}$$

$$\begin{aligned} E(R_i^2) &= 2 + 2(d_i \textstyle\sum d_j^3 - 2d_i^2 \sum d_j^2)/(\sum d_j^2)^2 \\ &= 2 + c_{ii}^{\dagger}, \end{aligned} \tag{34}$$

where the connection with (28) is that $c_{ii}^{\dagger} = c_{ii} + 2a_i$.

Although Feigl and Zelen compared two groups of observations it is sufficient here to consider only one of the groups. The data are given in Table 1, together with the corresponding values $a_i, c_{ii}^{\dagger}$ calculated from (33), (34); values of the crude residuals R_i, defined by (5), are also given.

In order to calculate modified residuals R_i' we note, since ϵ_i has an exponential distribution, that we require a transformation which adjusts the mean and variance and yet restricts R_i' to be positive. Hence we take

$$R_i' = \{R_i/(1 - l_i)\}^{1+k_i}, \tag{35}$$

where both l_i and k_i are small. Assuming this transforms R_i to an exponential distribution with unit mean, it can be shown that

$$E(R_i) = (1-l_i)\,\Gamma\{1+1/(1+k_i)\}$$

and

$$E(R_i^2) = (1-l_i)^2\,\Gamma\{1+2/(1+k_i)\}.$$

TABLE 1

Leukemia data (Feigl and Zelen, 1965). Log white blood cell count, x_i. Survival time, weeks, Y_i. Crude and modified residuals, R_i and R_i'

x_i	Y_i	a_i	$c_{ii}^{\dagger}$	R_i	R_i'
3·36	65	−0·013	−0·340	0·56	0·49
2·88	156	−0·088	−0·946	0·79	0·70
3·63	100	0·013	−0·134	1·17	1·12
3·41	134	−0·007	−0·293	1·23	1·20
3·78	16	0·022	−0·063	0·22	0·19
4·02	108	0·029	−0·003	1·94	1·91
4·00	121	0·029	−0·005	2·13	2·11
4·23	4	0·028	−0·012	0·09	0·07
3·73	39	0·019	−0·082	0·51	0·46
3·85	143	0·025	−0·039	2·12	2·11
3·97	56	0·028	−0·009	0·96	0·90
4·51	26	0·015	−0·109	0·80	0·74
4·54	22	0·013	−0·131	0·71	0·65
5·00	1	−0·037	−0·528	0·05	0·03
5·00	1	−0·037	−0·528	0·05	0·03
4·72	5	−0·002	−0·249	0·19	0·15
5·00	65	−0·037	−0·528	3·47	4·17

If we equate these expressions to (33) and (34), take logarithms and expand, ignoring high-order terms of k_i and l_i, we find

$$a_i = -l_i-(1-\gamma)\,k_i, \quad c_{ii}^{\dagger} = -4l_i-2(3-2\gamma)\,k_i,$$

where $\gamma = \Gamma'(1)/\Gamma(1)$. Thus

$$k_i = \tfrac{1}{2}(4a_i-c_{ii}^{\dagger}), \quad l_i = 0{\cdot}21c_{ii}^{\dagger}-1{\cdot}85a_i. \tag{36}$$

Values of R_i' are given in Table 1; the most noticeable difference between the crude and modified residuals occurs in the final entry, at $x = 5{\cdot}00$. A plot of R_i' against x_i shows no evidence that the mean or dispersion of the residuals vary systematically with x. The modified residuals R_i' are shown plotted in Fig. 1 against the expected values of exponential order statistics for a sample of size $n = 17$; the assumption of an exponential distribution is clearly confirmed.

A supplement to the graphical analysis is the calculation of a test statistic designed to examine consistency with the exponential distribution; see Cox and Lewis (1966, pp. 161–163) for a brief review of such tests applied to simple random samples.

One such test, although not normally the best, is based in effect on comparing the variance with the square of the mean. Now

$$T^* \equiv \sum R_i^2 = 31{\cdot}07$$

and the result (34) leads, with $n = 17$, to

$$E(T^*) = 2(n-2) = 30.$$

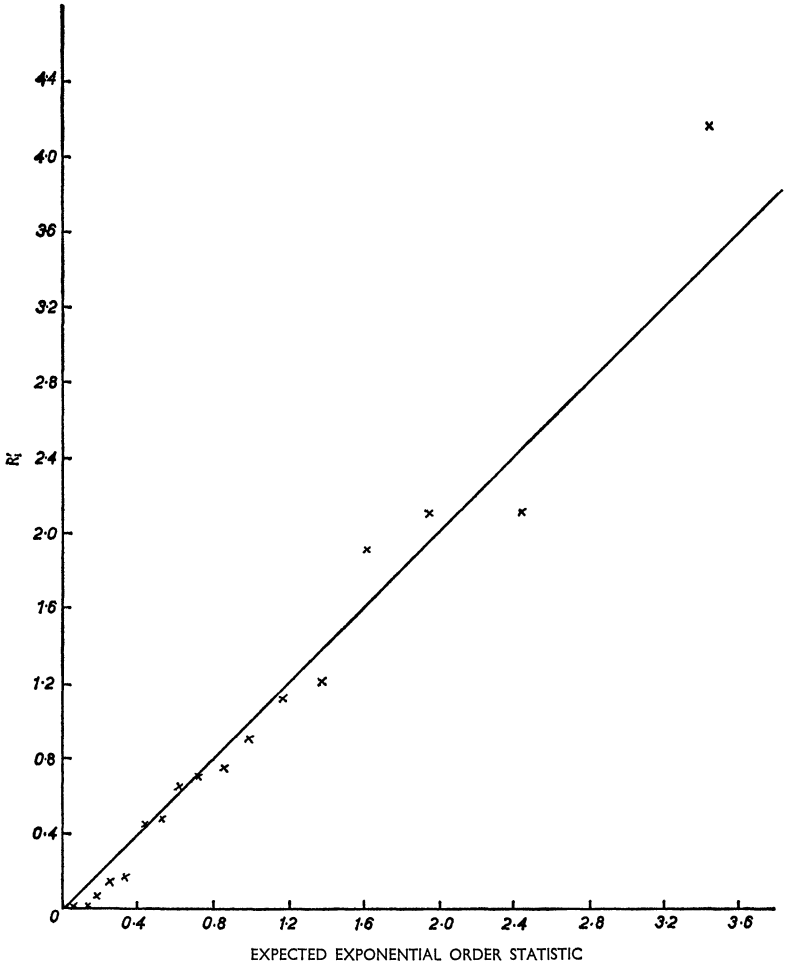

FIG. 1. Leukemia data. Modified residual, R_i', versus expected exponential order statistics. Straight line corresponds to unit exponential distribution.

Now if the residuals were obtained directly from a random sample of size n, i.e. $R_i = Y_i/\bar{Y}$, and the test statistic is

$$T^* = \sum (Y_i/\bar{Y})^2,$$

then it is easy to show that, to the order considered,

$$E(T^*) = 2(n-1).$$

This suggests that T^* should be regarded as derived from a random sample of size $n-1$, i.e. a "degree of freedom" subtracted for the extra parameter fitted. Again the close fit to the exponential distribution is confirmed.

The covariance between different residuals can be calculated from (25), (27). In fact

$$\operatorname{cov}(R_i, R_k) = -n^{-1} - d_i d_k / \sum d_j^2. \tag{37}$$

This is identical, to the order considered, with the corresponding formula for ordinary linear regression. Since the residuals here have approximately unit variance, (37) is also the correlation between residuals. In this example, the numerically greatest correlation is about 0·2. The presence of a substantial correlation between a particular pair of residuals would have been a warning of possible difficulties of interpretation, especially if both residuals had appeared to correspond to outliers.

In one way the fact that the R_i' and the R_i differ relatively little is an anticlimax. A more encouraging way of looking at the conclusions is, however, that they suggest that, even with the very small number of observations considered here, the distortion introduced by the fitting is small and that the unmodified residuals may in practice often be adequate.

7. POISSON DATA

In order to apply the methods of the preceding sections to Poisson-distributed observations, we must first consider how to define residuals so as to obtain nearly identically distributed variables. The difference from the earlier discussion is that, there, the model was defined directly in terms of independently distributed random variables. Here we proceed indirectly, defining R_i as

(a) $(Y_i-\mu_i)/\sqrt{\mu_i}$,

(b) $2(\sqrt{Y_i}-\sqrt{\mu_i})$,

or

(c) $\{\psi(Y_i)-\psi(\mu_i)\}/\{\psi'(\mu_i)\sqrt{\mu_i}\}$,

where $\mu_i = \mu_i(\hat{\boldsymbol{\beta}})$ is the expected frequency, on some model dependent upon parameters $\beta_1, \ldots, \beta_p$, of the Poisson observation, Y_i. Each of these, asymptotically, defines a standard normal deviate; (c) is a generalization of (a) and (b), and in it $\psi(x)$ is an arbitrary function.

The choice of an appropriate transformation depends upon the requirements; the need for a direct interpretation might lead to (a) or, alternatively, for a multiplicative model, to (c) with $\psi(x) = \log x$. But since our immediate object is to detect departures, rather than to explain them, it is desirable to find a transformation to give a set of residuals with a distribution as near as possible to some known form. Anscombe (1953) suggests that an appropriate transformation for normalizing Poisson observations is

$$\{Y_i^{\frac{2}{3}} - (\mu_i - \tfrac{1}{6})^{\frac{2}{3}}\}/(\tfrac{2}{3}\mu_i^{\frac{1}{6}}). \tag{38}$$

This can be derived by considering the Taylor expansion for moments of a power Y_i^h and equating the coefficient of skewness to zero; see also Moore (1957).

We therefore define $h_i(Y_i, \boldsymbol{\beta})$ by (38); we have also

$$p_j(Y_j, \boldsymbol{\beta}) = \exp(-\mu_j)\,\mu_j{}^{Y_j}/Y_j!$$

and hence we can apply the results of Section 4 to obtain $E(R_i)$ and $E(R_i^2)$. To do so, however, involves certain approximations and it is necessary to distinguish between terms that become small when n, the number of distinct Poisson observations, is large and those that become small when μ, the expectation of a typical observation, is large. The general results of Section 5 refer to large n.

The biases are given by

$$b_s = -\tfrac{1}{2} I^{rs} I^{tu} \sum \frac{1}{\mu_j} \frac{\partial \mu_j}{\partial \beta_r} \frac{\partial^2 \mu_j}{\partial \beta_t \partial \beta_u} \tag{39}$$

with

$$I_{rs} = \sum \frac{1}{\mu_j} \frac{\partial \mu_j}{\partial \beta_r} \frac{\partial \mu_j}{\partial \beta_s}$$

and hence b_s is of order μ^{-1}.

In obtaining $E(R_i)$ and $E(R_i^2)$, there is appreciable simplification if we consider leading terms in expansions in powers of $1/\mu$; this leads to terms of order $\mu^{-\frac{1}{2}}$ for $E(R_i)$ and of order 1 and μ^{-1} for $E(R_i^2)$. Thus if μ_i is sufficiently large it will be adequate to take as an approximation only the terms of order 1 and this gives

$$E(R_i) = 0, \quad E(R_i^2) = 1 - I^{rs} \frac{\partial \mu_i}{\partial \beta_r} \frac{\partial \mu_i}{\partial \beta_s} \frac{1}{\mu_i}. \tag{40}$$

The expression for $E(R_i)$ to order $\mu^{-\frac{1}{2}}$ is, in fact,

$$E(R_i) = -\frac{b_r}{\mu_i^{\frac{1}{2}}} \frac{\partial \mu_i}{\partial \beta_r} + I^{rs} \left(\frac{1}{6\mu_i^{\frac{3}{2}}} \frac{\partial \mu_i}{\partial \beta_r} \frac{\partial \mu_i}{\partial \beta_s} - \frac{1}{2\mu_i^{\frac{1}{2}}} \frac{\partial^2 \mu_i}{\partial \beta_r \partial \beta_s} \right). \tag{41}$$

The transformations (a) and (b) also lead to the approximation (40), to the order considered.

8. BINOMIAL DATA

Following the arguments of Section 7, we define a transformation $\phi(Y_i/m_i)$ of the observation Y_i from a binomial distribution with parameters $\theta_i(\boldsymbol{\beta})$ and m_i. By considering the Taylor expansion and equating the skewness to zero, we obtain a differential equation of which a solution is

$$\phi(u) = \int_0^u t^{-\frac{1}{3}} (1-t)^{-\frac{1}{3}} \, dt, \quad 0 \leqslant u \leqslant 1. \tag{42}$$

Blom (1954) suggests equation (42) as a normalizing transformation but does not apply it. In order to simplify its application we have computed Table 2. This gives values of $\phi(u)/\phi(1)$, i.e. the incomplete beta function $I_u(\frac{2}{3}, \frac{2}{3})$, which is symmetrical about $u = 0{\cdot}5$; multiplication by $B(\frac{2}{3}, \frac{2}{3}) = 2{\cdot}0533$ gives the value of (42). For example, $\phi(0{\cdot}2) = 2{\cdot}0533 \times 0{\cdot}257 = 0{\cdot}528$, $\phi(0{\cdot}8) = 2{\cdot}0533\,(1 - 0{\cdot}257) = 1{\cdot}526$.

Introducing the mean and variance of the transformed binomial variate, we define

$$h_i(Y_i, \boldsymbol{\beta}) = [\phi(Y_i/m_i) - \phi\{\theta_i - \tfrac{1}{6}(1-2\theta_i)/m_i\}] / \{\theta_i^{\frac{1}{2}}(1-\theta_i)^{\frac{1}{2}}/\sqrt{m_i}\}; \tag{43}$$

this reduces to (38) for small θ_i. Plots on probability paper suggest that the transformation is very effective, even for values as small as $m_i = 5$, $\theta_i = 0{\cdot}04$. Often the bias correction $-\frac{1}{6}(1-2\theta_i)/m_i$ can be omitted.

TABLE 2

Values of the incomplete beta function $I_u(\frac{2}{3}, \frac{2}{3})$

u	*0·000*	*0·001*	*0·002*	*0·003*	*0·004*	*0·005*	*0·006*	*0·007*	*0·008*	*0·009*
0·00	0	0·007	0·012	0·015	0·018	0·021	0·024	0·027	0·029	0·032
0·01	0·034	0·036	0·038	0·040	0·043	0·045	0·046	0·048	0·050	0·052
0·02	0·054	0·056	0·058	0·059	0·061	0·063	0·064	0·066	0·068	0·069
0·03	0·071	0·072	0·074	0·075	0·077	0·079	0·080	0·082	0·083	0·084
0·04	0·086	0·087	0·089	0·090	0·092	0·093	0·094	0·096	0·097	0·098
0·05	0·100	0·101	0·102	0·104	0·105	0·106	0·108	0·109	0·110	0·112
0·06	0·113	0·114	0·115	0·117	0·118	0·119	0·120	0·122	0·123	0·124
0·07	0·125	0·126	0·128	0·129	0·130	0·131	0·132	0·134	0·135	0·136
0·08	0·137	0·138	0·139	0·141	0·142	0·143	0·144	0·145	0·146	0·147
0·09	0·149	0·150	0·151	0·152	0·153	0·154	0·155	0·156	0·157	0·158
0·10	0·160	0·161	0·162	0·163	0·164	0·165	0·166	0·167	0·168	0·169
0·11	0·170	0·171	0·172	0·173	0·174	0·176	0·177	0·178	0·179	0·180
0·12	0·181	0·182	0·183	0·184	0·185	0·186	0·187	0·188	0·189	0·190
0·13	0·191	0·192	0·193	0·194	0·195	0·196	0·197	0·198	0·199	0·200
0·14	0·201	0·202	0·203	0·204	0·205	0·206	0·207	0·208	0·209	0·210
0·15	0·211	0·212	0·213	0·214	0·214	0·215	0·216	0·217	0·218	0·219
0·16	0·220	0·221	0·222	0·223	0·224	0·225	0·226	0·227	0·228	0·229
0·17	0·230	0·231	0·232	0·232	0·233	0·234	0·235	0·236	0·237	0·238
0·18	0·239	0·240	0·241	0·242	0·243	0·244	0·244	0·245	0·246	0·247
0·19	0·248	0·249	0·250	0·251	0·252	0·253	0·254	0·254	0·255	0·256
0·20	0·257	0·258	0·259	0·260	0·261	0·262	0·262	0·263	0·264	0·265
0·21	0·266	0·267	0·268	0·269	0·270	0·270	0·271	0·272	0·273	0·274
0·22	0·275	0·276	0·277	0·277	0·278	0·279	0·280	0·281	0·282	0·283
0·23	0·284	0·284	0·285	0·286	0·287	0·288	0·289	0·290	0·290	0·291
0·24	0·292	0·293	0·294	0·295	0·296	0·296	0·297	0·298	0·299	0·300
0·25	0·301	0·302	0·302	0·303	0·304	0·305	0·306	0·307	0·308	0·308
0·26	0·309	0·310	0·311	0·312	0·313	0·313	0·314	0·315	0·316	0·317
0·27	0·318	0·318	0·319	0·320	0·321	0·322	0·323	0·323	0·324	0·325
0·28	0·326	0·327	0·328	0·328	0·329	0·330	0·331	0·332	0·333	0·333
0·29	0·334	0·335	0·336	0·337	0·338	0·338	0·339	0·340	0·341	0·342
0·30	0·342	0·343	0·344	0·345	0·346	0·347	0·347	0·348	0·349	0·350
0·31	0·351	0·351	0·352	0·353	0·354	0·355	0·355	0·356	0·357	0·358
0·32	0·359	0·360	0·360	0·361	0·362	0·363	0·364	0·364	0·365	0·366
0·33	0·367	0·368	0·368	0·369	0·370	0·371	0·372	0·372	0·373	0·374
0·34	0·375	0·376	0·376	0·377	0·378	0·379	0·380	0·380	0·381	0·382
0·35	0·383	0·384	0·384	0·385	0·386	0·387	0·388	0·388	0·389	0·390
0·36	0·391	0·392	0·392	0·393	0·394	0·395	0·396	0·396	0·397	0·398
0·37	0·399	0·400	0·400	0·401	0·402	0·403	0·403	0·404	0·405	0·406
0·38	0·407	0·407	0·408	0·409	0·410	0·411	0·411	0·412	0·413	0·414
0·39	0·414	0·415	0·416	0·417	0·418	0·418	0·419	0·420	0·421	0·422
0·40	0·422	0·423	0·424	0·425	0·425	0·426	0·427	0·428	0·429	0·429
0·41	0·430	0·431	0·432	0·433	0·433	0·434	0·435	0·436	0·436	0·437
0·42	0·438	0·439	0·440	0·440	0·441	0·442	0·443	0·443	0·444	0·445
0·43	0·446	0·447	0·447	0·448	0·449	0·450	0·450	0·451	0·452	0·453
0·44	0·454	0·454	0·455	0·456	0·457	0·457	0·458	0·459	0·460	0·461
0·45	0·461	0·462	0·463	0·464	0·464	0·465	0·466	0·467	0·468	0·468
0·46	0·469	0·470	0·471	0·471	0·472	0·473	0·474	0·474	0·475	0·476
0·47	0·477	0·478	0·478	0·479	0·480	0·481	0·481	0·482	0·483	0·484
0·48	0·485	0·485	0·486	0·487	0·488	0·488	0·489	0·490	0·491	0·491
0·49	0·492	0·493	0·494	0·495	0·495	0·496	0·497	0·498	0·498	0·499

The p.d.f. of Y_j is

$$p_j(Y_j, \boldsymbol{\beta}) = \binom{m_j}{Y_j} \theta_j^{Y_j}(1-\theta_j)^{m_j - Y_j},$$

from which we obtain the biases

$$b_s = -\tfrac{1}{2} I^{rs} I^{tu} \sum \frac{m_j}{\theta_j(1-\theta_j)} \frac{\partial \theta_j}{\partial \beta_r} \frac{\partial^2 \theta_j}{\partial \beta_t \partial \beta_u}, \tag{44}$$

To obtain $E(R_i)$ and $E(R_i^2)$, we consider expansions in powers of m^{-1} and get, analogous to (40), the approximation

$$E(R_i) = 0, \quad E(R_i^2) = 1 - I^{rs} \frac{\partial \theta_i}{\partial \beta_r} \frac{\partial \theta_i}{\partial \beta_s} \frac{m_i}{\theta_i(1-\theta_i)}. \tag{45}$$

In the numerical example which follows, we checked that the higher order terms neglected in (45) are indeed negligible.

9. A Further Example

Dyke and Patterson (1952) present the analysis for a 2^4 factorial design of the proportions of respondents who achieve good scores on cancer knowledge; some details of the data are given in columns (1)–(3) of Table 3. They assume a logit transformation of the proportions, the expected value of the transformed variate being a linear function of parameters representing main effects and interactions. Values of the parameters are estimated by maximum likelihood. We consider their solution and apply the methods of Section 8 to examine residuals from the fitted model.

Following Dyke and Patterson (with slight changes in notation) we write the model as

$$\theta_j(\boldsymbol{\beta}) = \{1 + \exp(-2z_j)\}^{-1},$$

where $z_j = l_{jr}\beta_r$, summed over the parameters; $l_{jr} = \pm 1$. Then

$$\frac{\partial \theta_j}{\partial \beta_r} = 2\theta_j(1-\theta_j) l_{jr},$$

$$\frac{\partial^2 \theta_j}{\partial \beta_r \partial \beta_s} = 4\theta_j(1-\theta_j)(1-2\theta_j) l_{jr} l_{js}$$

and substitution into (44) gives

$$b_s = -4 I^{rs} I^{tu} \sum m_j \theta_j(1-\theta_j)(1-2\theta_j) l_{jr} l_{jt} l_{ju}.$$

From (45), we have

$$E(R_i) = 0$$

and

$$E(R_i^2) = 1 - 4 I^{rs} m_i \theta_i(1-\theta_i) l_{ir} l_{is}$$
$$= 1 + c_{ii}^{\dagger}.$$

Dyke and Patterson fit a model with five parameters, representing the overall mean and the four main effects. They quote the values of I^{rs} obtained in the course of their solution and we use these values to calculate b_s and $c_{ii}^{\dagger}$.

In order to calculate modified residuals, we use (30) writing $l_i = 0$ since $E(R_i) = 0$; solving for k_i, remembering k_i is small, we obtain

$$R_i' = (1 - \tfrac{1}{2}c_{ii}^{\dagger})\, R_i.$$

Values of $c_{ii}^{\dagger}$, R_i and R_i' are given in Table 3. The biases b_s were all extremely small; none exceeded $2\frac{1}{2}$ per cent of the standard error of the estimate of the parameter.

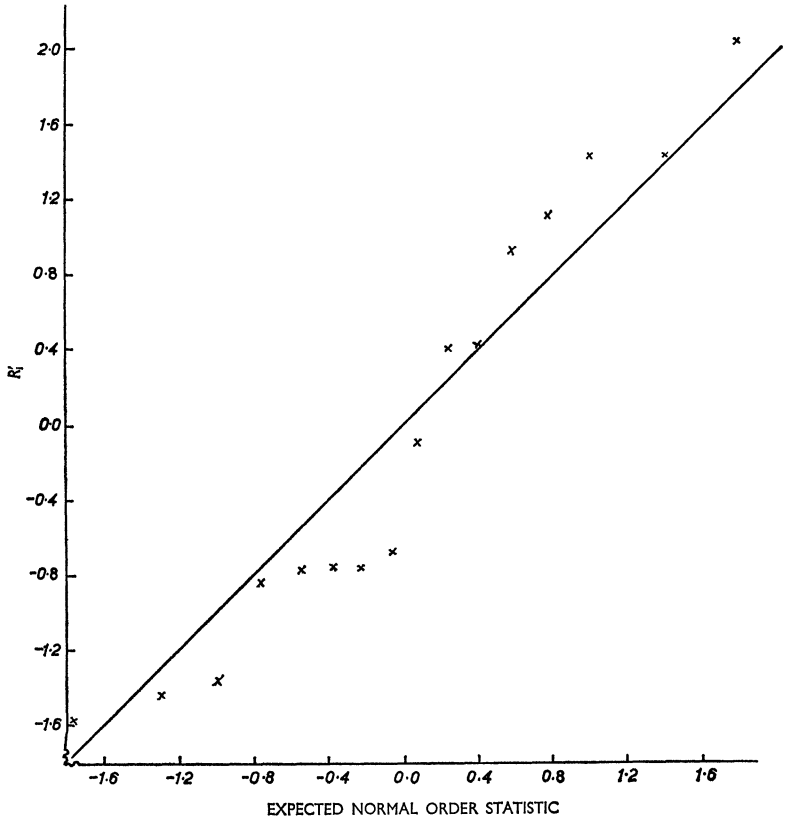

FIG. 2. 2^4 factorial design. Modified residual, R_i', versus expected normal order statistics. Straight line corresponds to unit normal distribution.

The modified residuals R_i' are plotted against the expected normal order statistics in Fig. 2 and show agreement with the assumption of a standard normal distribution.

Dyke and Patterson go on to fit a model with extra parameters to represent the interactions AD, BD and CD. Before proceeding to extend the model, we examine the residuals R_i' from the simple main effects model. If we regard the values of R_i' as observations in a 2^4 design, we can analyse them in the usual way and obtain the sums of squares given in Table 4. This is an unweighted analysis and, as such, can be used only for guidance; the existence of any apparent effect can be established only by fitting a model containing the appropriate parameters. Also, for the same reason, the sums of squares due to main effects in Table 4 are not zero. Nevertheless,

the magnitude of the AC effect suggests it to be worth including in the model along with AD, BD and CD; the three-factor interaction ACD also is large but has not been included in the further analysis. We therefore fitted a model with nine parameters;

TABLE 3

2^4 factorial design (Dyke and Patterson, 1952). r_i number of "good" responses out of m_i. Crude and modified residuals, R_i and R'_i

Treatment	m_i	r_i	$-c^{\dagger}_{ii}$	R_i	R'_i
abcd	31	23	0·248	0·36	0·41
abc	169	102	0·540	−0·46	−0·68
abd	12	8	0·135	1·34	1·44
ab	94	35	0·372	−0·07	−0·09
acd	45	27	0·379	−0·59	−0·75
ac	378	201	0·711	−0·45	−0·84
ad	13	7	0·133	1·04	1·12
a	231	75	0·527	0·65	0·94
bcd	4	1	0·050	−1·33	−1·37
bc	32	16	0·190	0·38	0·43
bd	7	4	0·076	1·38	1·44
b	63	13	0·225	−0·68	−0·78
cd	11	3	0·123	−1·47	−1·57
c	150	67	0·505	1·45	2·06
d	12	2	0·100	−0·73	−0·77
1	477	84	0·705	−0·78	−1·43

TABLE 4

2^4 factorial designs. Sums of squares of modified and crude residuals

Effect	*Sums of squares*	
	R'_i	R_i
A	0·78	0·80
B	0·26	0·20
C	1·09	1·14
D	0·01	0·00
AB	0·04	0·04
AC	2·53	0·98
AD	1·85	1·57
BC	0·32	0·20
BD	2·06	1·69
CD	4·88	3·89
ABC	2·51	1·26
ABD	0·07	0·10
ACD	3·77	1·97
BCD	0·03	0·05
ABCD	0·03	0·03

the estimated values and standard errors are given in Table 5. Comparison of the estimates with their standard errors confirms that the AC interaction is at least as significant as the AD and BD interactions.

TABLE 5

2^4 factorial design. Estimated parameters and their standard errors

Parameter	*Estimate*	*Standard error*
Mean	−0·13	0·06
A	0·25	0·06
B	0·12	0·05
C	0·17	0·05
D	0·11	0·06
AC	−0·05	0·03
AD	0·10	0·06
BD	0·06	0·05
CD	−0·11	0·05

For comparison, the corresponding analysis on the crude residuals, R_i, is also given in Table 4; it is interesting that the AC interaction does not stand out in this case. Note, however, that 9 parameters are being fitted to 16 observations, so that the applicability of the asymptotic formulae is in doubt.

10. FURTHER WORK

Some possible extensions of the work fall under the following three broad headings.

(i) There are applications, for example to time series and components of variance problems, in which more than one random variable contributes to each observation. Durbin and Watson (1950, 1951) have considered a significance test for serial correlation based on residuals from a fitted regression.

(ii) There could be more refined distributional studies of the quantities discussed above. In particular, the more detailed study of order statistics of residuals would be useful both in connection with detecting outliers and for tests of distributional form. There is obvious scope for simulation.

(iii) Further special applications can be considered. Two general types concern (a) the examination of distributional form from random samples, for example by using the probability integral transformation to produce residuals having a theoretical rectangular distribution, and (b) more complex situations that are essentially generalizations of regression.

As an example of (a), consider the Weibull distribution (Section 2), where a transformation can be made to the unit exponential distribution, or the circular normal distribution, where the probability integral transformation can be applied. As further examples of (b), consider two problems closely associated with normal-theory regression, namely the calculation of residuals after fitting a transformation (Box and Cox, 1964) or after fitting a non-linear model by iterative least squares.

ACKNOWLEDGEMENT

We are grateful to Mrs E. A. Chambers and Mr B. G. F. Springer for programming the calculations. Their work was supported by the Science Research Council.

REFERENCES

ANSCOMBE, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. *J. R. Statist. Soc.* B, **15**, 229–230.

—— (1961). Examination of residuals. *Proc. 4th Berkeley Symp.*, **1**, 1–36.

BARTLETT, M. S. (1952). Approximate confidence intervals. II. *Biometrika*, **40**, 306–317.

BLOM, G. (1954). Transformations of the binomial, negative binomial, Poisson and χ^2 distributions. *Biometrika*, **41**, 302–316.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc.* B, **26**, 211–252.

COX, D. R. and LEWIS, P. A. W. (1966). *The Statistical Analysis of Series of Events*. London: Methuen.

DURBIN, J. and WATSON, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, **37**, 409–428.

—— (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, **38**, 159–178.

DYKE, G. V. and PATTERSON, H. D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.

FEIGL, P. and ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant observation. *Biometrics*, **21**, 826–838.

HALDANE, J. B. S. (1953). The estimation of two parameters from a sample. *Sankhyā*, **12**, 313–320.

HALDANE, J. B. S. and SMITH, S. M. (1956). The sampling distribution of a maximum likelihood estimate. *Biometrika*, **43**, 96–103.

MOORE, P. G. (1957). Transformations to normality using fractional powers of the variate. *J. Am. Statist. Ass.*, **52**, 237–246.

PEARSON, E. S. and HARTLEY, H. O. (1966). *Biometrika Tables for Statisticians*, 3rd ed. Cambridge University Press.

SHENTON, L. R. and BOWMAN, K. (1963). Higher moments of a maximum-likelihood estimate. *J. R. Statist. Soc.* B, **25**, 305–317.

SHENTON, L. R. and WALLINGTON, P. A. (1962). The bias of moment estimators with an application to the negative binomial distribution. *Biometrika*, **49**, 193–204.

DISCUSSION ON THE PAPER BY PROFESSOR COX AND MRS SNELL

Mr M. J. R. HEALY (Medical Research Council): From one point of view, the statistical developments of the last half-century or so may be regarded as the detailed exploitation of the linear model. That these developments are not over is evidenced by the programme of next week's ordinary meeting of the Society† (to which Professor Cox is making another of his characteristically stimulating contributions); but there is an ever-growing interest nowadays in the use of non-linear models for the statistical study of a wide diversity of phenomena. One reason for this is the feeling that the physical situation generating the observations will often demand the use of non-linear formulae, and these are bound to be needed as soon as we deal with either thresholds or asymptotes.

Parallel with this change in the climate of statistical thought has been the growing interest in the study of residuals. As soon as a feeling of dissatisfaction with one's model sets in, it is natural to look at the discrepancies between its predictions and what is actually observed, and there is now a fair-sized literature dealing with methods for doing this. Both the tendencies I refer to have been fostered by advances in computing facilities which have brought to practical trial hitherto theoretical notions and techniques.

It is when both tendencies intersect that we reach the subject of tonight's paper. If our model involves, for example, a curve with a horizontal asymptote, it is almost inevitable that the standard notion of additive homoscedastic errors with a symmetric distribution

† The discussion at that meeting is published in Part 3 of Series A of the *journal*.

will be quite inadequate. A model which takes the underlying physical situation seriously enough to influence the expectations might also allow it to affect the error structure, and doing so is likely to lead to just the sort of problems dealt with here. The solutions to such problems may well be more important than they would be in the linear case. If a linear model is admittedly an over-simplification, we can tolerate even some systematic departures from it, weighing these against the advantages of simplicity and ease of calculation; whereas the extra cost of the non-linear model may not be recouped by better insight into the true situation if discrepancies between model and observations are allowed to remain uninvestigated.

I would like to make a few specific comments on the paper, in particular on the second example, and to ask the authors whether they have compared their incomplete beta-function transform with any alternatives. The maximum-likelihood calculations can often be viewed as weighted least squares following a transformation, and thus can readily be persuaded to produce a set of quasi-residuals as a by-product. Even simpler, perhaps, is a procedure I have myself found useful for informal investigation; it consists of tabulating or plotting (observed − expected)/(expected)$^{\frac{1}{2}}$ the sum of squares of these quantities being simply the goodness-of-fit χ^2. The surprising thing about the transformation tabulated in the paper is perhaps its near linearity between 10–90 per cent.

I confess that I share the authors' reservations over treating 16 − 9 as a large number. I would also like to ask whether they know any way of taking advantage of the fact that their "ordered plots" should in the null situation not only give straight lines but straight lines with known position and slope. Perhaps one might examine the residuals.

In summary, my reaction to this paper has been that which is so often stimulated by papers in which Professor Cox has had a hand—a mixture of somewhat envious admiration with an impatience to go away and apply the suggested methods to data a satisfactory treatment of which has so far eluded me. I would like to propose a very hearty vote of thanks to both authors.

Mr P. J. HARRISON (Imperial Chemical Industries Ltd): In tonight's paper Professor Cox and Mrs Snell deal with a topic which has always been of importance to practising statisticians and which with the advent of computers has grown in importance. The main questions which we ask of residuals have been detailed in Section 1 of the paper and these are broadly:

(i) Does the fitted model adequately describe the data?
(ii) Are the data adequate for determining the model? Here we may be concerned with the need for data editing such as blocking observations or eliminating outliers. We may also be concerned about whether more observations are required for model discrimination.
(iii) Is the assumption about the nature of the error distribution correct and if not is a re-analysis necessary?

Before statisticians had ready access to computers they had a reasonable degree of control over residuals. This was primarily due to the fact that observations largely resulted from statistically designed experiments and there was often prior information about the nature of the experimental error. When this was not the case then computing limitations usually restricted investigations to models with few parameters.

However, immediately computers became available for statistical analyses, multiple linear regression and non-linear estimation programs were written which were capable of analysing models involving a large number of parameters.

Initially most of these programs were applied somewhat indiscriminately, and generally to data which had not been statistically planned. Residuals were listed but their analysis was left to the statistician who probably contented himself with a quick scan for outliers and distribution form. More recently these programs have attempted to automate the examination of model adequacy.

In one I.C.I. program there is an option for considering the family of models considered by Box and Cox (1964). The paper defines this as a more formal approach. Procedures are also available for dealing with outliers, for testing distributional form and for plotting residuals with selected variables. It is also possible to test the hypothesis that the residuals are distributed in the multivariable space according to a given law. In this case interest might lie in detecting and defining regions of the space in which the residuals are significantly correlated. Clump or Cluster Analysis is a possibility here.

In general the residuals which are analysed are those the paper refers to as Crude Residuals and Professor Cox and Mrs Snell draw our attention to the fact that perhaps we should work in terms of Modified Residuals. However, the paper comes to no definite recommendation that we should use the Modified Residual except perhaps when the number of observations is small compared with the number of parameters to be estimated, and I would agree with the authors that this is encouraging.

The paper suggests that further work might be undertaken in defining residuals related to time series and component of variance problems. Certainly in such situations the problems are more difficult. Box and Jenkins discuss the identification of non-stationary time series models using combinations of differencing, regression, overparameterization and simulation techniques. In industry the most frequently encountered non-stationary model can be represented in terms of a first-order random walk generating process:

$$\left.\begin{aligned} d_t &= m_t + \epsilon_t, \\ m_t &= m_{t-1} + \gamma_t, \end{aligned}\right\} \tag{1}$$

where d_t is the observation at time t and ϵ and γ are each a set of identically independently distributed random variables with zero mean.

This model is a particular case of the polynomial random walk in which each derivative is subjected to random impulses. The linear least-squares predictor for (1) is

$$\hat{d}_{t+1} = \hat{d}_t + Ae_t,$$

where $e_t = d_t - \hat{d}_t$, and A depends on the magnitude of $V(\epsilon)/V(\gamma)$.

In a dynamic situation the residuals are closely scrutinized and often analysed using a series of Wald Sequential tests in the form of a Backward Cumulative Sum test. If the test shows a significant departure from the null hypothesis that the predictor is satisfactory then the residuals are used to assess (i) whether an "outlier" has occurred in the ϵ set in which case that particular outlier is ignored (ii) whether an outlier has occurred in the γ set in which case adjustments to the predictor are made or (iii) whether the model has changed.

In the study of chemical reactions and processes both variance components and stochastic variation are to be expected and present difficulties in the analysis of residuals. Professor Cox and Mrs Snell have taken one step towards the clarification of residuals and I have much pleasure in seconding the vote of thanks.

The vote of thanks was put to the meeting and carried unanimously.

Professor F. J. ANSCOMBE (Yale University): I heartily endorse Mr Healy's appreciation of this paper. It is a most timely study of a topic that will surely command widespread interest—a topic that no doubt many of us have been uneasily aware of and are now delighted to see unfold so clearly.

Back in the days when I frequented this hall, an honoured tradition permitted a contributor to the discussion sometimes to make remarks only indirectly related to the subject of the paper. I hope I shall be pardoned for speaking now about computing.

One fateful day last June I was taken by a colleague to see a demonstration of a new computer language. The language is based on a book by K. E. Iverson called *A Programming Language* (Iverson, 1962). The book was concerned with the precise and concise

expression of algorithms, using a notation and way of thinking closely resembling established mathematics. The present implementation of a modified version of the language as a computer coding language is known as APL. (The modest first letter has been retained, though some of us think that TPL would be more fitting. The language should not be confused with PL/1, developed for IBM's 360 series of computers.) APL has been implemented experimentally at IBM's Thomas J. Watson Research Center, Yorktown Heights, N.Y., for computation in conversational mode through typewriter terminals. At present it has not been generally released.

Requirements for statistical computing are many and various, because the persons who have occasion to do such computing have very diverse degrees of interest in statistics and in computing. No one system or method can be satisfactory for all. A professional statistician (like me) needs to be able to experiment freely in computing, and ought if possible to be able to do so without a vast effort in mastering a computer language and with little time spent in coding or otherwise specifying what he wants done. It seems to me that, unlike previous general-purpose languages such as FORTRAN or ALGOL, APL is sufficiently powerful and sufficiently easy to learn to meet this need. I have been preparing a paper to show how the language can be used for typical statistical calculations, to show enough of its character that a reader could have some basis for judging whether to take an active interest. Far less than that can be said now.

Statistical computing, like other computing, requires negotiation of arrays. Various languages and systems have been proposed permitting matrix operations to be called for easily. APL also is designed to handle arrays. The unique feature of APL is the generality of its definitions, leading to a high degree of consistency that not only makes the language easy to remember but also gives it a peculiar dignity and reasonableness. One example must suffice.

Any language or system designed to handle matrices must obviously encompass matrix addition. It must permit a command like:

$$\text{ADD} \quad \mathbf{A} \quad \mathbf{B},$$

where **A** and **B** are matrices of the same size. In APL matrix addition is denoted by

$$\mathbf{A}+\mathbf{B}.$$

What is peculiar is that this notation refers not to a special operation, matrix addition, but to a general method of combining two arrays that are not necessarily matrices by a function that is not necessarily addition. In fact **A** and **B** can be any two arrays of the same size—vectors, or matrices, or rectangular arrays of any number of dimensions. The function "+" can be replaced by any other standard function f having the same syntax, that is by a symbol f such that for any scalar arguments **A** and **B** the combination $\mathbf{A}\,f\,\mathbf{B}$ is scalar. Thus if **A** and **B** are two 17-dimensional arrays of the same size, $\mathbf{A}\times\mathbf{B}$ means the array of the same size formed by multiplying each element of **A** by the corresponding element of **B**. Similarly for **A** * **B** (where * means exponentiation) and **A** | **B** (where | means the residue function) and **A** = **B** (where = is the logical function taking the value 1 when the arguments are equal and 0 otherwise), and so on for a dozen more functions. We might say that for the cost of a capability in matrix addition we have obtained free many other capabilities, many of which turn out to be just as useful. The whole of APL has been constructed with this kind of generality.

Whatever may be the fate of this particular implementation of APL, something like it must surely eventually command widespread approval.

The relevance of computing to the paper by Professor Cox and Mrs Snell is that until a computer has been adequately tamed, residuals have only theoretical interest. Such a study was not done many years ago in the Fisherian era because computers were not available then. I am eager to try out these methods, which promise to be a valuable weapon in the continual struggle to fit theory to facts.

Professor J. DURBIN (London School of Economics): I am strongly in favour of the examination of the residuals after carrying out regression analysis, linear or non-linear, as a means of examining the adequacy of the model specification. But it must be recognized that this can be a highly treacherous business where an intuitive analysis can lead one seriously astray. In addition to the inspection of visual plots of various kinds one will often wish to carry out some form of test of significance. If the test statistic based on the fitted residuals is $\mathbf{a}(R_1, \ldots, R_n)$, which we may take to be suitably normalized, it might appear reasonable to suppose that its distribution converges for large n to that of the corresponding statistic based on the true residuals $\mathbf{a}(\epsilon_1, \ldots, \epsilon_n)$ on the ground that $R_i - \epsilon_i$ usually converges stochastically to zero for all n. If this were true one could construct large-sample tests of fit by treating statistics calculated from the fitted residuals as if they had been based on the true residuals. The method is usually invalid, however, except in particularly favourable circumstances.

Let $\boldsymbol{\beta}$ denote the parameters of interest and suppose that if $\boldsymbol{\beta}$ were known, $\mathbf{a}$ would estimate a vector $\boldsymbol{\alpha}$ of nuisance parameters. The model specification is then equivalent to a hypothesis of the form $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$. Let $\mathbf{b}$ be the maximum-likelihood estimate of $\boldsymbol{\beta}$ assuming $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ and let $\mathbf{a}$ be the maximum-likelihood estimator of $\boldsymbol{\alpha}$ assuming $\boldsymbol{\beta} = \mathbf{b}$. In the present context $\mathbf{a}$ would be a statistic based on the fitted residuals. Assuming the usual type of regularity conditions one can then show that $\sqrt{(n)}\,(\mathbf{a} - \boldsymbol{\alpha}_0)$ is asymptotically normal with zero mean and variance matrix $\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\mathbf{B}^{-1}\mathbf{C}'\mathbf{A}^{-1}$ where

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{C} \\ \hline \mathbf{C}' & \mathbf{B} \end{array}\right] = -\frac{1}{n} E \begin{bmatrix} \dfrac{\partial^2 \log L}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}'} & \dfrac{\partial^2 \log L}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\beta}'} \\ \dfrac{\partial^2 \log L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\alpha}'} & \dfrac{\partial^2 \log L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} \end{bmatrix}.$$

On the other hand, the intuitive procedure would take $\mathbf{a}$ to have the same asymptotic distribution as the maximum-likelihood estimator of $\boldsymbol{\alpha}$ assuming the true value of $\boldsymbol{\beta}$ to be known, that is $\sqrt{(n)}\,(\mathbf{a} - \boldsymbol{\alpha}_0)$ would be assumed to be $N(\mathbf{0}, \mathbf{A}^{-1})$. Since the two distributions can be very different the intuitive test can be seriously misleading. It is interesting to note that any error is always in the same direction, that is, towards the underestimation of significance. The intuitive test is asymptotically valid if $E\{(\partial^2 \log L)/(\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\beta}')\} = \mathbf{0}$. These results are given in a paper "Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables" which is to appear in *Econometrica*.

To get an idea of how the results apply to problems of the type considered by Professor Cox and Mrs Snell consider a simplified form of their model (7), namely

$$y_i = \exp(\beta x_i)\,\epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independently exponentially distributed and the null hypothesis is that they have mean unity. Suppose we require a test based on the residuals that will detect a shift in mean. The example is of course artificial but I believe it brings out the main point. The intuitive test would be to compute

$$a = \frac{1}{n}\sum_1^n R_i$$

and to take this as asymptotically $N(1, n^{-1})$. Applying the above theory we find that in fact $\sqrt{(n)}\,a$ has asymptotic variance

$$1 - n\bar{x}^2 \Big/ \sum_1^n x_i^2 = \sum_1^n (x_i - \bar{x})^2 \Big/ \sum_1^n x_i^2.$$

The intuitive procedure therefore underestimates the variance of the test statistic by an amount that can be very large.

Dr V. BARNETT (University of Birmingham): I should like to echo the sentiments of earlier speakers and say how much I enjoyed this paper. There are just two points I should like to make. The first and major one arises from the form of the model (3) and defined residuals (5). Implicit in the definition (5) is that any parameters which relate specifically to the distribution of "errors" or unobserved variables ϵ are incorporated in the set of model parameters, β. Later discussion and examples in the paper make this quite apparent. For instance, the authors suggest that an unknown variance in the normal-theory linear model should be handled in this way (example 1, Section 2); and the ϵ-distributions for the two practical examples discussed are assumed to be in standardized form—this is an essential feature of the ordered residual plots of Figs. 1 and 2, requiring the linear relationships to be of unit slope and through the origin. The mathematical convenience of this approach is quite clear and well demonstrated in this paper. Also, in many cases the basic behaviour of the model is bound up with the parameters related to the error distribution (examples 2 and 3 in Section 2 are of this type), and we have no alternative but to simultaneously estimate all the parameters in fitting the model. But in certain cases this is not so; for a regression model, say, with independent identically distributed (not necessarily normal) errors it is reasonable (and indeed customary) to fit the parameters in the basic model separately, and without regard to the scale parameter in the error distribution. This is not to deny the essential need of studying residuals in order to validate the model; but there would seem to be no reason why the residuals should have been fully standardized for this purpose. More generally, some parameters in the error distribution may be of secondary practical interest, and not warrant the complication of the model fitting process by their inclusion.

The procedure of plotting the ordered residuals (or ordered modified residuals) remains a useful tool for obtaining a visual validation of the model in such cases. But it can provide the further service of yielding quick estimates of any error distribution parameters omitted from the fitted model. Also, although such omitted parameters may have secondary practical interest, we are not absolved from trying to get as good estimates of them as we are able, within the scope of the (limited) effort we are prepared to put into this. This raises the general question of estimation via probability plotting procedures. Such methods find wide practical application in many fields, for example in engineering, meteorology, hydrology, etc., usually for estimating scale and location parameters in distributions with density function of the form $(1/\sigma)f\{(x-\mu)/\sigma\}$. However, there would seem to have been little serious study of the statistical *rationale* of this approach to estimation. The method consists simply of plotting ordered sample values, $y_{(i)}$, at preassigned values, x_i, of a variable x; and estimating the scale and location parameters from the best straight-line relationship. Various sets of plotting positions, x_i, have been employed. These might be, as in this paper, the expected values of the standardized order statistics or, more commonly in practical application, the inverse probabilities of the cumulative sampling distribution, for example

$$F^{-1}\left(\frac{i}{n+1}\right) \quad \text{or} \quad F^{-1}\left(\frac{i-\frac{1}{2}}{n}\right).$$

It is in the choice of x_i that more detailed work is needed, since it is quite apparent that the statistical properties (bias and relative efficiency) of the resulting estimates of μ and σ may vary *wildly* with the plotting positions chosen.

Little detailed formal discussion of this problem appears in the literature; although there are papers by Benard and Bos-Levenbach (1953) who propose modified plotting positions for general use; and Chernoff and Lieberman (1954), concerned specifically with the normal distribution. Some more detailed results on suitable choice of plotting positions for estimation can be easily obtained. For instance, if μ happens to be the *mean* of the distribution, unbiasedness of the estimate of μ constrains the x_i to always produce the sample mean $\bar{y}$, which for certain distributions is most inefficient. For the plotting positions proposed in this paper no problems of bias will arise for, say, a scale parameter σ

(if we ignore inter-residual correlations); but we may well sacrifice a lot of information on σ in relation to some other choice of plotting positions. This is well illustrated, for example, by the case of uniformly distributed residuals with half range σ.

The choice of plotting positions relates also to the question of model validation. We are invited to consider the plotted ordered residuals in Figs. 1 and 2 as evidence of the adequacy of the respective models for these examples. It is largely a subjective matter how convincing we find the evidence in these two cases—personally Fig. 2 worries me a little! More objective criteria of the adequacy of the model can of course be constructed and applied to these plots, as the authors have mentioned. But again the properties, and relative convenience, of relevant test statistics for this purpose are going to depend (perhaps quite strongly) on the choice of plotting positions, x_i.

My second and concluding point is something of a personal "hobby horse" and I will be brief. In the Introduction the authors comment on the use of computers in studying residuals. There is a growing tendency, in some circles, to regard the computer as a substitute for common sense, thought and observation. I certainly would not presume to criticize the authors on these grounds; the whole spirit of their paper denies this attitude; nor would I dispute the essential value of the computer in large-scale studies of residuals. But it would be unfortunate if the lack of a "suitable computer with graphical output device" was allowed to discourage, or excuse, the study of residuals for model validation. A pencil and piece of graph paper may still work wonders in this respect!

Dr S. C. PEARCE (East Malling Research Station): I studied the text of this paper with the greatest interest and found it both stimulating and provocative. After a time I came to suspect that the authors know rather more about the quantities, R'_i, than they say, so perhaps I may provoke them in turn, hoping they will add to their remarks and so make an excellent paper even better.

These quantities are derived from the crude residuals, R_i, by a transformation that is intended to give them the same mean and variance as the random variables, ϵ_i. Accordingly they go only part of the way towards the desired quantities, $R''_{(i)}$, each of which shall be an estimate of one of the ϵ_i's. Where they first appear at equation (30), there are some cautionary words about particular cases and fairly general procedures, but when they are actually used at equation (35) the word is *Hence*; I suggest that it should be *Let*. The transformation proposed is completely reasonable, but it is not unique for the purpose intended.

In fact the transformation in this instance proves to be rather too powerful, as is shown by Fig. 1. In general, it will make small R_i into even smaller R'_i whereas it will increase large R_i. In Fig. 1 we see that all the smallest residuals lie below the line while the largest is awkwardly above it, and we should actually have done better not to have transformed at all. Turning to Fig. 2, the transformation has again been too powerful; with one exception all the negative residuals are too small and all the positive ones too large. Here let me agree that these two Figures provide most impressive support for the essential rightness of the authors' approach; my point is merely that there are in fact systematic deviations, which may be due to the arbitrary element in the transformation.

However, there are other explanations possible. Perhaps the systematic deviations result from a quantity having been badly estimated on account of some quirk in the data. In that case the fact of the transformation having been too powerful in two instances may be of no more importance than a coin having been tossed twice and come down heads on both occasions. Alternatively everything may be the result of using R'_i instead of $R''_{(i)}$ or ϵ_i. Admittedly the method of deriving the R' is reasonable, because each ϵ_i must be a function of the values R_i and must depend chiefly upon the R_i to which it corresponds, but a transformation such as the one used can hardly be exact. There is another point, how certain can we be *a priori* that the random residuals, ϵ_i, are in fact distributed exponentially (Fig. 1) or normally (Fig. 2)? (In the latter case, as a matter of fact, it is scarcely conceivable that they should be.) For my own part I can see no way of judging where the systematic deviations come from, but perhaps the authors can help.

I have rather trailed my coat because I would like to hear the extended comments of the authors on this point. I would, however, advance a suggestion. Scandalous as the suggestion may seem at a meeting of the Royal Statistical Society, there are occasions when real data are a nuisance and fudged-up figures are better. This is perhaps an occasion for simulation. If we knew for certain what the parameters and random residuals were, we could apply the methods of this valuable paper and observe how the values of R_i' actually behave. It could be that we do not need to seek much further.

Mr A. M. WALKER (University of Cambridge): I would like to ask a very simple question in connection with Example 2 in Section 2 of the paper, where the model is given by equation (7). Why did the authors not take logarithms of each side of the equation, so as to obtain a linear regression model of the usual form, with independently and identically distributed residuals $\log \epsilon_i = \eta_i$ (say)? If that were done, and the basic parameters taken to be $\log \beta_1$ and β_2, the analysis would seem to be somewhat simpler, because the parameters occur linearly in the definition of the residuals, which become $\log R_i$, $1 \leqslant i \leqslant n$, in the authors' notation. There would of course be no difference as regards maximum likelihood estimation (and least-squares estimators might well not be particularly efficient, but derivation of formulae analogous to (33) and (34) on p. 255 giving the approximate mean and variance of $\log R_i$ would certainly be more straightforward, and if behaviour of residuals not in accordance with the model was expected to be associated with the value of the independent variable x (log white blood cell count), the logarithmic transformation would be a fairly natural one. Also the estimated residuals are no longer restricted to be positive, so that one does not have to introduce a transformation such as that given by equation (35); incidentally, taking logarithms in this equation obviously just brings one back to equation (30) with R_i replaced by $\log R_i$. The distribution of η_i is, admittedly, less simple than that of ϵ_i, but is still fairly easy to handle. For example its moment generating function $E(\exp t\eta) = \Gamma(1+t)$, so that its rth cumulant is $\psi^{(r-1)}$ (1), where $\psi(x) = d/dx\{\log \Gamma(x)\}$ denotes the digamma function.

This question illustrates the point that the general definition of residuals given at the beginning of Section 2 is not unique; one can apply an arbitrary 1–1 transformation to the ϵ_i and still satisfy it, as the resulting random variables remain independently and identically distributed. In many problems there will be a natural transformation to use, but that will not always be so, and even when it is, is the "natural" transformation necessarily the most appropriate? I would be grateful if Professor Cox and Mrs Snell could comment on this.

The following written contributions were received after the meeting.

Dr R. M. LOYNES (University of Cambridge): I should like to say that I enjoyed the paper, and to observe very briefly that it is possible to obtain the distribution of the residuals R_i, and then to find a function $\rho_i(R_i)$ which has the same distribution as ϵ_i, thus producing a different modified residual $R_{(i)}''$; these statements are to hold if account is taken of terms of order n^{-1} but of no higher order.

A numerical comparison with the values of R_i' in Table 1 shows, as one would expect, no great differences: all except the last row showing differences no greater than 0·03, and a difference of 0·09 in the last row.

One can similarly consider joint distributions of the R_i, and more complicated adjustments of them in an attempt to reduce the dependence. It is not possible to be quite as thorough now, but on general grounds it seems unlikely that much can be done: for example in the simplest possible (normal, linear) case there are n independent ϵ_i, while the R_i satisfy a linear relationship.

A full account will be given elsewhere.

Dr C. L. MALLOWS (Bell Telephone Laboratories): I would like to congratulate the authors on their important paper, and to report some results that bear on the questions raised at the end of Section 5 and in Section 10 (ii). Suppose that $X_1, \ldots, X_n$ are random variables (for example, residuals) that are nearly, but not quite, independent and identically distributed. In a forthcoming paper I give a general method for obtaining the distributions of the order-statistics of $X_1, \ldots, X_n$, and of pairs, triads, and so on of these order-statistics. These distributions are obtained as expansions in series; approximate results can be obtained by truncation. A key result is that through first correction terms, the joint distribution of any pair of order-statistics is determined as soon as the pairwise joint distributions of $X_1, \ldots, X_n$ are known. Thus first-order corrections to the means, variances and covariances of the order-statistics can be determined from the means, variances and covariances of the original variables. For example, suppose $X_1, \ldots, X_n$ are multinormal with zero means, unit variances and small correlations. (If $X_1, \ldots, X_n$ are rescaled residuals, the correlations will be known.) One finds that if squares and higher powers of the correlations are neglected, then through second moments the order-statistics behave as though all the correlations were equal to their average. Thus to this approximation the configuration of the order-statistics has the same distributional properties as that derived from a set of independent normal variables, so that probability plotting remains an appropriate technique.

Dr M. B. PRIESTLEY (University of Manchester): I should like to join the other discussants in expressing my thanks to the authors for a very interesting and stimulating paper. The techniques which they propose will no doubt find many useful applications, but I would like to mention one or two points which I feel require further consideration. In the first place, equation (3) does not by itself seem to define a unique set of $\{\epsilon_i\}$. As stated, the $\{\epsilon_i\}$ are defined as a set of independent identically distributed random variables. However, as Mr Walker has pointed out, there are many transformations on any given set of $\{\epsilon_i\}$ which will result in a new set of random variables which are also independent and identically distributed. Since the authors' definition of the residuals $\{R_i\}$ (equation (5)) assumes a knowledge of the maximum-likelihood estimates $\hat{\boldsymbol{\beta}}$, and hence implicit knowledge of the distributional form of the $\{\epsilon_i\}$, it might be as well to include this information in the basic model (equation (3)), so that the $\{R_i\}$ would then be defined more explicitly with respect to a given family G of functions $\{g_i\}$ and a given family F of distributions for $\{\epsilon_i\}$.

The "interaction" between the unknown parameters of G ($\boldsymbol{\beta}$) and the unknown parameters of F has already been mentioned by Dr Barnett, and is related to the general problem of interpreting the observed $\{R_i\}$. The authors have investigated some of the sampling properties of the $\{R_i\}$, and have shown how these properties may be used to detect departures from the original model. However, these results were deducted on the *basis of the original model.* If, therefore, one decides (on the result of some test on the R_i) that this model was inadequate, there would seem no reason to suppose, *a priori*, that the $\{R_i\}$ will still follow closely the pattern of the (modified) $\{\epsilon_i\}$. In the case of a polynomial regression model it is probably true that, in general, an examination of the $\{R_i\}$ (or $\{R_i'\}$) would correctly indicate when further terms were needed. However, in the context of time-series analysis (that is, when the $\{\epsilon_i\}$ are allowed to be correlated in "time") it is doubtful whether the same conclusion would hold, and a periodogram analysis of the residuals (as suggested by the authors in Section 1) might be misleading. As an example, consider a linear system with stationary input X_t and transfer function A, and suppose that the output, W_t, contains an additive "noise" term, ϵ_t. Then the system may be described by the equation:

$$W_t = AX_t + \epsilon_t.$$

Suppose now that one observes X_t and W_t over some interval of time and wishes to estimate the transfer function A. As an initial model one might suppose that ϵ_t was a white noise

process uncorrelated with X_t, and estimate A by the usual technique of cross-spectral analysis. However, if these assumptions were incorrect then the estimate of A could be seriously in error, and any subsequent model fitted to the residuals may be invalid. In such situations it would be nice if one could appeal to some type of "orthogonality" property—but this is hardly likely to be applicable in many situations.

Professor J. TUKEY (Princeton University): The broad conclusions to be supported or drawn from this work are relatively clear. These include:

(i) Since residuals are mainly used to look for what might be going on beyond what is already in the model, we can go for this with first-order answers. Small corrections, whether or not of higher order, are often negligible. (Studies of the distribution of residuals may prove exceptions and may not.) The important thing is to look at "residuals"; details of definition matter much less.

(ii) The authors have made available a useful and well-behaved normalizing transformation for the binomial.

(iii) The authors have provided approximate formulae for bias and variance of any residual-like function of the observations and the parameters, however selected. (How complex the situations are, under which we will, in fact, use these formulae is not yet clear.)

There remain a number of minor points which I fail to understand. These include: The omission (in (v) on p. 248) of distributional behaviour as a possible clue to heterogeneous variance. The statement, in the second paragraph of Section 2, that Poisson variates cannot be expressed in terms of identically distributed quantities. (The inverse of the probability-integral transformation always applies—presumably the interchanged statement was meant.) Why it is not more convenient to introduce the logarithm of time to death in Example 2, thus obtaining residuals that are both additive and unbounded? Why should we be more concerned with variances than with covariances for the R_i of Table 3? And finally how much the nine-parameter fit leading to Table 5 increases our belief in the reality of the AC effect beyond that coming from Table 4?

Mrs SNELL replied briefly at the meeting and the authors subsequently replied more fully in writing as follows:

We are grateful for the extremely constructive and helpful comments that have been made and for the pinpointing of problems for further study. As Mr Healy has pointed out, there are a number of commonly used transformations of the binomial distribution which are for many purposes effectively linear functions of one another. The incomplete beta transformation studied in the paper does, however, give an appreciably more linear plot on probability paper than, for example, the inverse sine transformation, especially in the range $p = 0{\cdot}1 - 0{\cdot}3$ and with small n, for example $n \simeq 10$; in looking for systematic departures this could be an advantage. Mr Harrison has raised in particular the question of the behaviour of cusum charts plotted from residuals, and we agree that this deserves further study, especially of the effect on the plot of correlation between different residuals.

We have deliberately in the paper put the main emphasis on the plotting of residuals rather than on formal tests of significance. Professor Durbin and also Dr Priestley have stressed the need for caution in applying significance tests to residuals. That this is a very important point is clear from a consideration of the simple problem of examining in linear regression the possible importance of an omitted regressor variable, say z. It is entirely legitimate to make a graphical analysis by plotting residuals from the initial regression relation directly against z. If, however, the significance of the regression on z is to be tested by the usual formula the residuals of z must be used, as is in effect done in analysis of covariance. In the more general situations contemplated in our paper the expected value of linear or nearly linear test statistics calculated from residuals will be close to that calculated from the ϵ_i's, but the contribution of the covariance terms (28) to the variance of such test statistics will in general be non-negligible, because the number of covariance

terms will be of order n^2. A special case is Professor Durbin's example for the sum of residuals, where the correct variance he gives follows directly from (28). That is, in such applications it is essential to take into account the correction terms in (28) and with more complex test statistics to develop appropriate extensions of them.

Mr Walker has pointed out that the distribution of the ϵ_i's could by transformation be taken in any form, and he and Professor Tukey have enquired about the use of a log transformation in the exponential regression model (7), converting it into a model with additive error. It must be stressed that computationally there is no particular advantage in such a transformation, unless a log normal distribution of error is assumed. Our reason for keeping to the original scale was partly the simplicity of properties of the exponential distribution and more importantly that we felt that in this application the precise form of the distribution for small values is not important and that test statistics and plots that are sensitive to the very small values are not good. We think that similar arguments of simplicity and meaningfulness will often be applicable; see also the remarks at the beginning of Section 7 of the paper.

Dr Loynes and Dr Mallows have sketched theoretical arguments that should be an improvement on those that we have used and we look forward to seeing a detailed account of their work. In particular they should throw light on some of the points raised by Dr Pearce, with whom we agree that there is substantial scope for simulation in studying these problems.

We agree with Dr Barnett that it is not always necessary to take the distribution of the ϵ_i's in standardized form, nor always to include its parameters in fitting the model. Often however, as in our two examples, it will be useful to take the ϵ_i's as having a completely known distribution.

We accept the general points made by Dr Priestley; they make explicit some of the reservations about residuals discussed in the introduction of our paper.

Professor Tukey has made a considerable number of cogent points. We agree that the real usefulness of minor adjustments to the residuals is not established although, as Professor Durbin's contribution makes clear, second-order properties are important when it comes to tests. In the statement about Poisson variates that he queries, we had in mind that no *continuous* relation is available. If the R_i's are to be plotted against external variables we think that standardizing them to have equal means and variances is very reasonable to avoid spurious regularities in the plots but in examining distributional form the neglect of covariances is less easily defended, as indeed we indicated in the paper.

Finally we mention some relevant references that have come to our attention since we wrote the paper. Dr Mallows has pointed out that the work of David and Johnson (1948) has some bearing on the problem briefly mentioned at the end of Section 10. Also Theil (1965, 1968) has suggested that for normal theory linear model problems adjusted residuals with exactly the distribution of the true errors can be produced by attempting to find residuals only at a suitably limited set of observational points.

References in the Discussion

Benard, A. and Bos-Levenbach, E. C. (1953). The plotting of observations on probability paper. *Statistica*, **7**, 163–173.

Chernoff H. and Lieberman, G. J. (1954). Use of normal probability paper. *J. Amer. Statist. Ass.* **49**, 778–785.

David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, **35**, 182–190.

Iverson, K. E. (1962). *A Programming Language*. New York: Wiley.

Theil, H. (1965). The analysis of distributions in regression analysis. *J. Amer. Statist. Ass.*, **60**, 1067–1079.

—— (1968). A simplification of the BLUS procedure for analysing regression disturbances. *J. Amer. Statist. Ass.*, **63**, 242–251.