TESTING FOR SERIAL CORRELATION IN LEAST SQUARES REGRESSION. II

BY J. DURBIN,* *London School of Economics*
AND
G. S. WATSON, *Department of Applied Economics, University of Cambridge*

1. INTRODUCTION

In an earlier paper (Durbin & Watson, 1950) the authors investigated the problem of testing the error terms of a regression model for serial correlation. Test criteria were put forward, their moments calculated, and bounds to their distribution functions were obtained. In the present paper these bounds are tabulated and their use in practice is described. For cases in which the bounds do not settle the question of significance an approximate method is suggested. Expressions are given for the mean and variance of a test statistic for one- and two-way classifications and polynomial trends, leading to approximate tests for these cases. The procedures described should be capable of application by the practical worker without reference to the earlier paper (hereinafter referred to as Part I).

It should be emphasized that the tests described in this paper apply only to regression models in which the independent variables can be regarded as 'fixed variables'. They do not, therefore, apply to autoregressive schemes and similar models in which lagged values of the dependent variable occur as independent variables.

2. THE BOUNDS TEST

Throughout the paper the procedures suggested will be illustrated by numerical examples. We begin by considering some data from a demand analysis study.

Example 1. *Annual consumption of spirits from* 1870 *to* 1938. The data (given in Table 1) were compiled by A. R. Prest, to whose paper (1949) reference should be made for details of the source material. As is common in econometric work the original observations were transformed by taking logarithms:

y = log consumption of spirits per head;
x_1 = log real income per head;
x_2 = log relative price of spirits (i.e. price of spirits deflated by a cost-of-living index).

We suppose that the observations satisfy the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{1}$$

where β_0 is a constant, β_1 is the income elasticity, β_2 is the price elasticity, and ϵ is a random error with zero mean and constant variance.

To test the errors for serial correlation the following sums of squares and products are required:

$\Sigma(y-\bar{y})^2$	= 5·000123	$\Sigma(y-\bar{y})(x_2-\bar{x}_2)$	= −3·763579	$\Sigma(\Delta x_2)^2$	= 0·083559
$\Sigma(x_1-\bar{x}_1)^2$	= 0·632006	$\Sigma(x_1-\bar{x}_1)(x_2-\bar{x}_2)$	= 1·014984	$\Sigma\Delta y\Delta x_1$	= 0·014685
$\Sigma(x_2-\bar{x}_2)^2$	= 2·966354	$\Sigma(\Delta y)^2$	= 0·112592	$\Sigma\Delta y\Delta x_2$	= −0·076399
$\Sigma(y-\bar{y})(x_1-\bar{x}_1)$	= −1·321973	$\Sigma(\Delta x_1)^2$	= 0·023539	$\Sigma\Delta x_1\Delta x_2$	= 0·000527

* Junior research worker at the Department of Applied Economics, Cambridge, when this work was begun.

$\Sigma(\Delta y)^2$ stands for the sum of squares of the first differences of the y's, and $\Sigma \Delta y \Delta x$ stands for the sum of products of the first differences of the y's times the corresponding first differences of the x's, etc. Thus the first term in $\Sigma(\Delta y)^2$ is $(1{\cdot}9794 - 1{\cdot}9565)^2 = 0{\cdot}00052441$, and the first term in $\Sigma \Delta y \Delta x_1$ is $(1{\cdot}9794 - 1{\cdot}9565)(1{\cdot}7766 - 1{\cdot}7669) = 0{\cdot}00022213$. Since there are 69 observations there are 69 terms in each of the first six summations, and 68 in each of the second six summations.

Table 1. *Annual consumption of spirits from* 1870 *to* 1938

Year	Consumption y	Income x_1	Price x_2	Year	Consumption y	Income x_1	Price x_2
1870	1·9565	1·7669	1·9176	1905	1·9139	1·9924	1·9952
1871	1·9794	1·7766	1·9059	1906	1·9091	2·0117	1·9905
1872	2·0120	1·7764	1·8798	1907	1·9139	2·0204	1·9813
1873	2·0449	1·7942	1·8727	1908	1·8886	2·0018	1·9905
1874	2·0561	1·8156	1·8984	1909	1·7945	2·0038	1·9859
1875	2·0678	1·8083	1·9137	1910	1·7644	2·0099	2·0518
1876	2·0561	1·8083	1·9176	1911	1·7817	2·0174	2·0474
1877	2·0428	1·8067	1·9176	1912	1·7784	2·0279	2·0341
1878	2·0290	1·8166	1·9420	1913	1·7945	2·0359	2·0255
1879	1·9980	1·8041	1·9547	1914	1·7888	2·0216	2·0341
1880	1·9884	1·8053	1·9379	1915	1·8751	1·9896	1·9445
1881	1·9835	1·8242	1·9462	1916	1·7853	1·9843	1·9939
1882	1·9773	1·8395	1·9504	1917	1·6075	1·9764	2·2082
1883	1·9748	1·8464	1·9504	1918	1·5185	1·9965	2·2700
1884	1·9629	1·8492	1·9723	1919	1·6513	2·0652	2·2430
1885	1·9396	1·8668	2·0000	1920	1·6247	2·0369	2·2567
1886	1·9309	1·8783	2·0097	1921	1·5391	1·9723	2·2988
1887	1·9271	1·8914	2·0146	1922	1·4922	1·9797	2·3723
1888	1·9239	1·9166	2·0146	1923	1·4606	2·0136	2·4105
1889	1·9414	1·9363	2·0097	1924	1·4551	2·0165	2·4081
1890	1·9685	1·9548	2·0097	1925	1·4425	2·0213	2·4081
1891	1·9727	1·9453	2·0097	1926	1·4023	2·0206	2·4367
1892	1·9736	1·9292	2·0048	1927	1·3991	2·0563	2·4284
1893	1·9499	1·9209	2·0097	1928	1·3798	2·0579	2·4310
1894	1·9432	1·9510	2·0296	1929	1·3782	2·0649	2·4363
1895	1·9569	1·9776	2·0399	1930	1·3366	2·0582	2·4552
1896	1·9647	1·9814	2·0399	1931	1·3026	2·0517	2·4838
1897	1·9710	1·9819	2·0296	1932	1·2592	2·0491	2·4958
1898	1·9719	1·9828	2·0146	1933	1·2635	2·0766	2·5048
1899	1·9956	2·0076	2·0245	1934	1·2549	2·0890	2·5017
1900	2·0000	2·0000	2·0000	1935	1·2527	2·1059	2·4958
1901	1·9904	1·9939	2·0048	1936	1·2763	2·1205	2·4838
1902	1·9752	1·9933	2·0048	1937	1·2906	2·1205	2·4636
1903	1·9494	1·9797	2·0000	1938	1·2721	1·1182	2·4580
1904	1·9332	1·9772	1·9952				

The regression coefficients are calculated by inverting the matrix

$$\begin{bmatrix} \Sigma(x_1 - \bar{x}_1)^2 & \Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)^2 \\ \Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) & \Sigma(x_2 - \bar{x}_2)^2 \end{bmatrix},$$

giving for the estimates of β_1 and β_2

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 2{\cdot}966354 & -1{\cdot}014984 \\ -1{\cdot}014984 & 0{\cdot}632006 \end{bmatrix} \begin{bmatrix} -1{\cdot}321973 \\ -3{\cdot}763579 \end{bmatrix},$$

i.e.

$$b_1 = -0{\cdot}120142, \quad b_2 = -1{\cdot}227647.$$

Let z denote the residual from regression, i.e.

$$z = y - \bar{y} - b_1(x_1 - \bar{x}_1) - b_2(x_2 - \bar{x}_2).$$

Then
$$\Sigma z^2 = \Sigma(y-\bar{y})^2 - b_1\Sigma(y-\bar{y})(x_1-\bar{x}_1) - b_2\Sigma(y-\bar{y})(x_2-\bar{x}_2)$$
$$= 0{\cdot}22095.$$

The statistic to be used for testing for serial correlation is

$$d = \frac{\Sigma(\Delta z)^2}{\Sigma z^2}. \tag{2}$$

The reasons for choosing this statistic have been given in Part I and need not be discussed here. Now

$$\Delta z = \Delta y - b_1\Delta x_1 - b_2\Delta x_2,$$

so that

$$\Sigma(\Delta z)^2 = \Sigma(\Delta y)^2 + b_1^2\Sigma(\Delta x_1)^2 + b_2^2\Sigma(\Delta x_2)^2 - 2b_1\Sigma\Delta y\,\Delta x_1 - 2b_2\Sigma\Delta y\,\Delta x_2 + 2b_1b_2\Sigma\Delta x_1\,\Delta x_2$$
$$= 0{\cdot}054967.$$

Substituting in (2) we have $d = 0{\cdot}2488$.

We must now decide what departures from the null hypothesis of serial independence of the errors ϵ need be considered. Experience with econometric data such as the present indicates a test against the existence of positive serial correlation. If the errors were positively serially correlated, d would tend to be relatively small, while if the errors were negatively serially correlated d would tend to be large. We therefore require a critical value of d, say d^*, such that if the observed value of d is less than d^* we may infer that positive serial correlation is established at the significance level concerned.

It was shown in Part I that exact critical values of this kind cannot be obtained. However, it is possible to calculate upper and lower bounds to the critical values. These are denoted by d_U and d_L. If the observed d is less than d_L we conclude that the value is significant, while if the observed d is greater than d_U we conclude that the value is not significant at the significance level concerned. If d lies between d_L and d_U the test is inconclusive.

Significance points of d_L and d_U are tabulated for various levels in Tables 4, 5 and 6. In addition, a diagram is given to facilitate the test procedure in the most usual case of a test against positive serial correlation at the 5 % level (Fig. 1). k' is the number of independent variables.

In the present example $n = 69$ and $k' = 2$, so that at the 5 % level $d_L = 1{\cdot}54$ approximately. The observed value 0·25 is less than this and therefore indicates significant positive serial correlation at the 5 % level. In fact, the observed value is also significant at the 1 % level.

The procedure for other values of k' is exactly similar. In all cases the value of d given by (2) is calculated, the z's being the residuals from regression, and the appropriate table is consulted.

Tests against negative serial correlation and two-sided tests

Tests against negative serial correlation may sometimes be required. For instance, it is a common practice in econometric work to analyse the first differences of the observations rather than the observations themselves, on the ground that the serial correlation of the transformed errors is likely to be less than that of the original errors. We may wish to ensure that the transformation has not overcorrected, thus introducing negative serial correlation

into the transformed errors. To make a test against negative serial correlation, d is calculated as above and subtracted from 4. The quantity $4-d$ may now be treated as though it were the value of a d-statistic to be tested for positive serial correlation. Thus if $4-d$ is less than d_L, there is significant evidence of negative serial correlation, and if $4-d$ is greater than d_U, there is not significant evidence; otherwise the test is inconclusive.

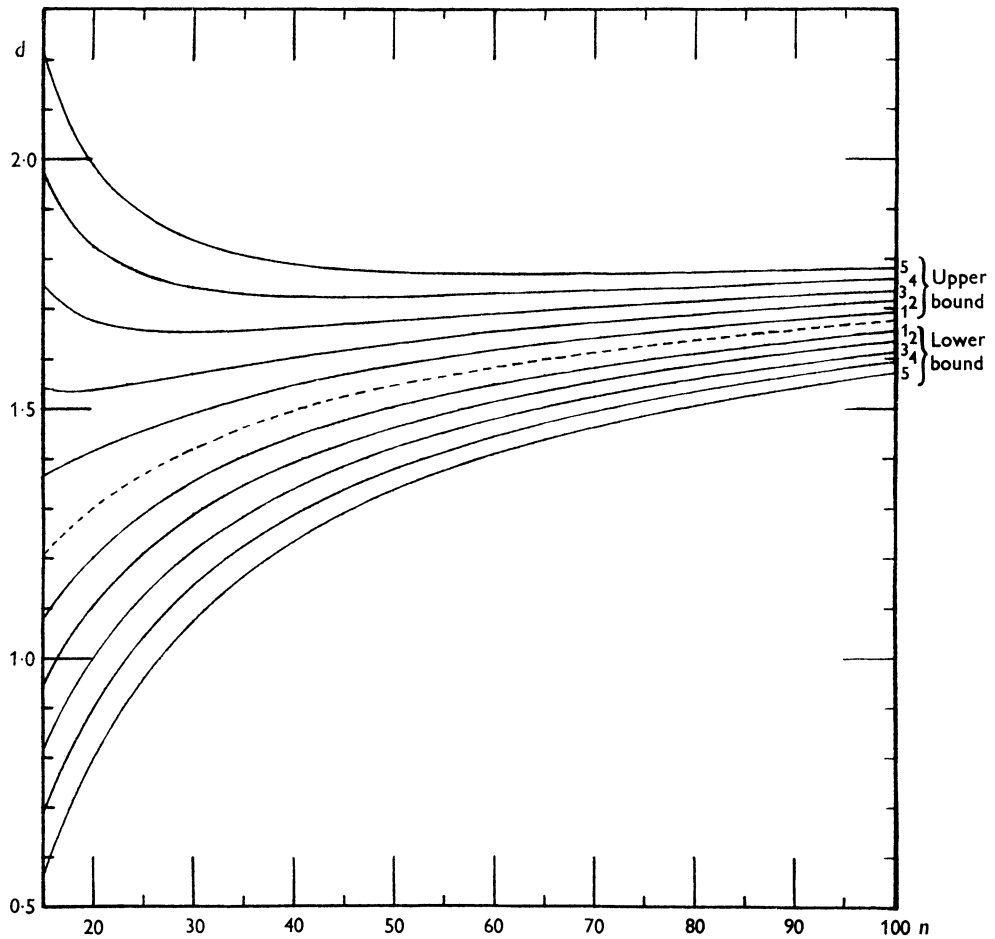

Fig. 1. Graphs of 5% values of d_L and d_U against n for $k'=1, 2, 3, 4, 5$.

When there is no prior knowledge of the sign of the serial correlation, two-sided tests may be made by combining single-tail tests. Using only equal tails, d will be significant at level α if either d is less than d_L or $4-d$ is less than d_L, non-significant if d lies between d_U and $4-d_U$ and inconclusive otherwise; the level α may be 2, 5 and 10 %. Thus, by using the 5 % values of d_L and d_U from Table 4, a two-sided test at the 10 % level is obtained.

3. Regression through the origin

The procedures described so far apply only to cases in which means have been fitted in the regressions, i.e. the fitted regression equations take the form

$$y-\bar{y} = b_1(x_1-\bar{x}_1)+\ldots+b_{k'}(x_{k'}-\bar{x}_{k'}). \tag{3}$$

These are the most common cases in practice. However, we occasionally require a fitted regression through the origin of the form

$$y = B_1 x_1 + \ldots + B_{k'} x_{k'}. \tag{4}$$

Tables 4, 5 and 6 do not apply directly to the residuals from regressions of this type. To test for serial correlation in such cases an equation of the form (3) must first be fitted. The residuals from the resulting regression may then be tested by the procedures of §§ 2 and 4. This gives a perfectly valid test even though (4) might be the more appropriate regression to fit for other purposes.

In order to avoid inverting more than one matrix the following method, due to Cochran (1938), should be used. First of all, the regression coefficients of equation (4) are determined. In this operation the inverse of the matrix of squares and cross-products of $x_1, \ldots, x_{k'}$ will be calculated; denote this matrix by $\mathbf{C} = \{c_{ij}\}$ and the means of the variables $y, x_1, \ldots, x_{k'}$ by $\bar{y}, \bar{x}_1, \ldots, \bar{x}_{k'}$. Then

$$\begin{bmatrix} b_1 \\ \vdots \\ b_{k'} \end{bmatrix} = \begin{bmatrix} B_1 \\ \vdots \\ B_{k'} \end{bmatrix} - nB_0 \mathbf{C} \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{k'} \end{bmatrix},$$

where

$$B_0 = \frac{\sum_{i=1}^{k'} B_i \bar{x}_i - \bar{y}}{\sum_{i,j=1}^{k'} \sum c_{ij} \bar{x}_i \bar{x}_j - 1}.$$

4. APPROXIMATE PROCEDURE WHEN THE BOUNDS TEST IS INCONCLUSIVE

No satisfactory procedure of general application seems to be available for cases in which the bounds test is inconclusive. However, an approximate test can be made, and this should be sufficiently accurate if the number of degrees of freedom is large enough, say greater than 40. For smaller numbers this test can only be regarded as giving a rough indication.

The method used is to transform d so that its range of variation is approximately from 0 to 1 and to fit a Beta distribution with the same mean and variance. The mean and variance of d vary according to the values of the independent variables, so the first step is to calculate them for the particular case concerned. The method of calculation will be illustrated by means of the data of Example 1, although in practice an approximate test would not be required for this case since the bounds test has given a definite answer.

The description of the computing procedure is greatly facilitated by the introduction of matrix notation. Thus the set $\{y_1, y_2, \ldots, y_n\}$ of observations of the independent variable is denoted by the column vector $\mathbf{y}$. In the same way the set

$$\begin{bmatrix} x_{11} & x_{21} & \ldots & x_{k'1} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{k'n} \end{bmatrix}$$

of observations of the independent variables is denoted by the matrix $\mathbf{X}$. We suppose that all these observations are measured from the sample means. The corresponding sets of first differences of the observations are denoted by $\Delta\mathbf{y}$ and $\Delta\mathbf{X}$. The numerator of (2) is the quadratic form

$$(\Delta\mathbf{z})'(\Delta\mathbf{z}) = \mathbf{z}'\mathbf{A}\mathbf{z},$$

where **A** is the real symmetric matrix

$$\begin{bmatrix} 1 & -1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 2 & -1 & \dots & \dots & \dots & \dots \\ 0 & -1 & 2 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & 2 & -1 \\ 0 & \dots & \dots & \dots & 0 & -1 & 1 \end{bmatrix}.$$

The moments of d are obtained by calculating the traces of certain matrices. The trace of a square matrix is simply the sum of the elements in the leading diagonal. For example, the trace of **A**, denoted by tr **A**, is $2(n-1)$, where n is the number of rows or columns in **A**.

It was shown in Part I that the mean and variance of d are given by

$$E(d) = \frac{P}{n-k'-1}, \tag{5}$$

$$\text{var}\,(d) = \frac{2}{(n-k'-1)(n-k'+1)}\{Q - PE(d)\}, \tag{6}$$

where
$$P = \text{tr}\,\mathbf{A} - \text{tr}\,\{\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\}, \tag{7}$$

and
$$Q = \text{tr}\,\mathbf{A}^2 - 2\,\text{tr}\,\{\mathbf{X}'\mathbf{A}^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} + \text{tr}\,[\{\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\}^2]. \tag{8}$$

The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ will have been obtained for the calculation of the regression coefficients, and the elements of $\mathbf{X}'\mathbf{A}\mathbf{X}$ for the calculation of d; for $\mathbf{X}'\mathbf{A}\mathbf{X} = (\boldsymbol{\Delta}\mathbf{X})'(\boldsymbol{\Delta}\mathbf{X})$, so that the (i,j)th element of $\mathbf{X}'\mathbf{A}\mathbf{X}$ is simply the sum of products $\Sigma\Delta x_i \Delta x_j$. Thus the only new matrix requiring calculation is $\mathbf{X}'\mathbf{A}^2\mathbf{X}$. Now $\mathbf{X}'\mathbf{A}^2\mathbf{X}$ is very nearly equal to $(\boldsymbol{\Delta}^2\mathbf{X})'(\boldsymbol{\Delta}^2\mathbf{X})$, where $\boldsymbol{\Delta}^2\mathbf{X}$ represents the matrix of second differences of the independent variables. Thus the (i,j)th element of $\mathbf{X}'\mathbf{A}^2\mathbf{X}$ will usually be given sufficiently closely by $\Sigma(\Delta^2 x_i)(\Delta^2 x_j)$, where $\Delta^2 x_i$ stands for the second difference of the ith independent variable. (More exactly

$$(\mathbf{X}'\mathbf{A}^2\mathbf{X})_{ij} = \Sigma(\Delta^2 x_i)(\Delta^2 x_j) + (x_{i1} - x_{i2})(x_{j1} - x_{j2}) + (x_{in-1} - x_{in})(x_{jn-1} - x_{jn}).)$$

The calculations will be exemplified by the data of Example 1. Referring to § 2 we see that

$$\mathbf{X}'\mathbf{A}\mathbf{X} = \begin{bmatrix} 0{\cdot}023539 & 0{\cdot}000527 \\ 0{\cdot}000527 & 0{\cdot}083559 \end{bmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 2{\cdot}966354 & -1{\cdot}014984 \\ -1{\cdot}014984 & 0{\cdot}632006 \end{bmatrix}.$$

Although $\text{tr}\,\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ is simply the sum of the two diagonal elements of the product of these two matrices, we shall need the remaining two elements below, so the whole matrix is computed giving

$$\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0{\cdot}069290 & -0{\cdot}023559 \\ -0{\cdot}083248 & 0{\cdot}052275 \end{bmatrix}.$$

Thus
$$\begin{aligned} \text{tr}\ \mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} &= 0{\cdot}069290 + 0{\cdot}052275 \\ &= 0{\cdot}121565. \end{aligned}$$

Substituting in (7) and (5) and remembering that $\operatorname{tr}\mathbf{A} = 2(n-1) = 136$ in the present case, we have

$$E(d) = \frac{135{\cdot}878435}{66} = 2{\cdot}05876.$$

The matrix of sums of squares and products of second differences is found to be

$$\mathbf{X}'\mathbf{A}^2\mathbf{X} = \begin{bmatrix} 0{\cdot}035867 & -0{\cdot}004495 \\ -0{\cdot}004495 & 0{\cdot}116368 \end{bmatrix}.$$

$\operatorname{tr}\mathbf{X}'\mathbf{A}^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ is obtained by multiplying the first column of $(\mathbf{X}'\mathbf{X})^{-1}$ into the first row of $\mathbf{X}'\mathbf{A}^2\mathbf{X}$ and adding the product of the second column of $(\mathbf{X}'\mathbf{X})^{-1}$ into the second row of $\mathbf{X}'\mathbf{A}^2\mathbf{X}$ giving $\operatorname{tr}\mathbf{X}'\mathbf{A}^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = 0{\cdot}189064$. $\operatorname{tr}[\{\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\}^2]$ is simply the sum of squares of the elements of the matrix $\mathbf{X}'\mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ i.e. 0·0150189. Also

$$\operatorname{tr}\mathbf{A}^2 = 2(3n-4) = 406.$$

Substituting in (8) and (6) we have

$$\begin{aligned} \operatorname{var} d &= \frac{2}{66\times 68}(405{\cdot}636891 - 135{\cdot}878435 \times 2{\cdot}05876) \\ &= 0{\cdot}0561033. \end{aligned}$$

We now assume that $\frac{1}{4}d$ is distributed in the Beta distribution with density

$$\frac{1}{B(p,q)}\left(\frac{d}{4}\right)^{p-1}\left(1-\frac{d}{4}\right)^{q-1}.$$

This distribution gives

$$E(d) = \frac{4p}{p+q},$$

$$\operatorname{var} d = \frac{16pq}{(p+q)^2(p+q+1)},$$

from which we find p and q by the equations

$$\left.\begin{aligned} p+q &= \frac{E(d)\{4-E(d)\}}{\operatorname{var} d} - 1, \\ p &= \tfrac{1}{4}(p+q)E(d). \end{aligned}\right\} \tag{9}$$

To test against positive serial correlation we require the critical value of $\frac{1}{4}d$ at the lower tail of the distribution. If $2p$ and $2q$ are integers, this can be obtained from Catherine Thompson's tables (1941), or indirectly from tables of the variance ratio or Fisher's z, such as those in the Fisher-Yates tables (1948); if $2p$ and $2q$ are not both integers, a first approximation may be found using the nearest integral values. Thus $F = \dfrac{p(4-d)}{qd}$ is distributed as the variance ratio and $z = \frac{1}{2}\log_e F$ is Fisher's z, both with $n_1 = 2q$, $n_2 = 2p$ degrees of freedom.

For moderately large numbers of observations a convenient way of finding the significance point when $2p$ and $2q$ are not integral is to use Carter's (1947) approximation to Fisher's z. This states that the critical value of z is approximately

$$\frac{\xi\sqrt{(h+\lambda)}}{h} - \left(\frac{1}{2q}-\frac{1}{2p}\right)\left(\lambda+\frac{5}{6}-\frac{s}{3}\right),$$

where

$$s = \frac{1}{2p}+\frac{1}{2q}, \quad h = \frac{2}{s}, \quad \lambda = \frac{\xi^2-3}{6}.$$

The values of ξ and λ to be used for 5 and 1 % tests against positive serial correlation are as follows:

	5%	1%
ξ	1·6449	2·3263
λ	0·0491	0·4020

Returning to the numerical example we find from (9)

$$p = 36{\cdot}1495, \quad q = 34{\cdot}0860, \qquad \text{whence} \quad F = 15{\cdot}99 \quad \text{and} \quad z = 1{\cdot}3848.$$

Carter's approximation gives a critical 1 % value of z of 0·278, which is less than the observed value, thus indicating significant serial correlation. Here the significance is so marked that it may be seen immediately by referring to a table of significant points of z or F around $n_1 = n_2 = 70$ (e.g. Snedecor, 1937).

For testing against negative serial correlation the same procedure is used except that d is replaced throughout by $4 - d$.

5. One- and two-way classifications

In any regression analysis where the independent variables assume the same values in all applications it would be theoretically possible to dispense with the bounds and tabulate the significance points of d once and for all. This could be done, for instance, for analysis of variance models such as one- and two-way classifications and for polynomial regressions with equally spaced variate values. The calculation of tables of this type is rather a formidable task and only one set has so far been published, namely, the significance points of the circular serial correlation coefficient of the least squares residuals from Fourier regressions, tabulated by R. L. & T. W. Anderson (1950). Pending the publication of further tables the bounds test of the present paper may be used, with the approximate procedure described in § 4 for cases in which the bounds do not give a decisive result, or when the n and k' are beyond the range of Tables 4, 5 and 6. In the present section the calculation of d is described for one- and two-way classifications and expressions are given for its mean and variance, while in the next section the same is done for polynomial regressions.

It is convenient to think of the observations as a time series consisting of monthly observations recorded for a number of years, though the results are of course of general application. We may fit constants for years only or for months only or for both months and years. An exact test of serial correlation in the 'months only' case may be made by means of R. L. & T. W. Anderson's tables (1950). For the 'years only' and the 'months and years' case no exact test is at present available. The tests described in this paper may, however, be used in all three cases.

If there are s 'years' each of t 'months', the mean and variance of d for the three models are as follows:

'Years' only:
$$\left.\begin{aligned} E(d) &= 2\left(1 + \frac{1}{t} - \frac{1}{st}\right), \\ \operatorname{var} d &= \frac{4}{s(t-1)(st-s+2)}\left\{st - 2s - \frac{2s}{t} + \frac{4}{t} + \frac{5s}{t^2} - \frac{8}{t^2} - \frac{2}{st} + \frac{2}{st^2}\right\}. \end{aligned}\right\} \tag{10}$$

'Months' only:
$$\left.\begin{aligned} E(d) &= 2\left(1 - \frac{1}{st}\right), \\ \operatorname{var} d &= \frac{4}{t(s-1)(st-t+2)}\left\{st - t - 1 + \frac{1}{s^2} - \frac{2}{st} + \frac{2}{s^2 t}\right\}. \end{aligned}\right\} \tag{11}$$

'Years and months':

$$\left.\begin{aligned} E(d) &= 2\left\{1+\frac{1}{t}-\frac{1}{s(t-1)}\right\}, \\ \operatorname{var} d &= \frac{4}{(s-1)(t-1)^2(st-t-s+3)} \\ &\quad \left\{st^2-t^2-3st+3t+4-\frac{2t}{s}+\frac{7s}{t}-\frac{12}{t}-\frac{5s}{t^2}+\frac{2t}{s^2}+\frac{6}{t^2}\right\}. \end{aligned}\right\} \tag{12}$$

These formulae were found by substituting the appropriate matrices into the general formulae given in Part I. The subsequent reductions were straightforward but extremely tedious.

Example 2. To illustrate the test procedures for models of this kind we shall consider the two-way classification of the data in Table 2 on the receipts of butter (in units of 1,000,000 lb.) at five markets (Boston, Chicago, San Francisco, Milwaukee and St Louis). The same data, for 1935, 1936 and 1937 only, were used by R. L. & T. W. Anderson (1950) for illustrating their procedure for testing serial correlation in the 'months only' case. The figures in parentheses are residuals from the monthly averages.

Table 2. *Receipts of butter (millions of lb. weight) at five U.S.A. markets*

Month	Year					Average
	1933	1934	1935	1936	1937	
Jan.	58·3 (+8·20)	52·6 (+2·50)	48·9 (−1·20)	48·3 (−1·80)	42·4 (−7·70)	50·10
Feb.	51·3 (+5·28)	46·9 (+0·88)	43·4 (−2·62)	47·1 (+1·08)	41·4 (−4·62)	46·02
March	58·1 (+5·86)	57·9 (+5·66)	43·8 (−8·44)	52·4 (+0·16)	49·0 (−3·24)	52·24
April	55·1 (+1·86)	54·2 (+0·96)	50·8 (−2·44)	55·3 (+2·06)	50·8 (−2·44)	53·24
May	74·6 (+5·94)	70·6 (+1·94)	67·6 (−1·06)	64·7 (−3·96)	65·8 (−2·86)	68·66
June	83·9 (+2·64)	73·3 (−7·96)	83·7 (+2·44)	79·5 (−1·76)	85·9 (+4·64)	81·26
July	73·5 (+1·56)	70·3 (−1·64)	82·7 (+10·76)	62·6 (−9·34)	70·6 (−1·34)	71·94
Aug.	73·3 (+11·78)	66·4 (+4·88)	60·8 (−0·72)	51·3 (−10·22)	55·8 (−5·72)	61·52
Sept.	63·0 (+7·96)	56·7 (+1·66)	55·4 (+0·36)	51·0 (−4·04)	49·1 (−5·94)	55·04
Oct.	58·3 (+5·58)	57·2 (+4·48)	48·4 (−4·32)	54·0 (+1·28)	45·7 (−7·02)	52·72
Nov.	55·1 (+9·20)	47·7 (+1·80)	37·7 (−8·20)	45·2 (−0·70)	43·8 (−2·10)	45·90
Dec.	56·5 (+9·70)	44·9 (−1·90)	41·0 (−5·80)	44·9 (−1·90)	46·7 (−0·10)	46·80
Average	63·42	58·22	55·35	54·69	53·92	57·12
Total	761·0	698·7	664·2	656·3	647·0	3427·2

Source: *Agricultural Statistics*, United States Government Printing Office, Washington, D.C., 1939, p. 390.

We require to test for serial correlation in the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (i = 1, 2, \ldots, 5; j = 1, 2, \ldots, 12),$$

where y_{ij} is the observation in the jth month of the ith year, μ, α_i and β_j are constants and ϵ_{ij} is the error term.

The least squares estimates of μ, α_i and β_j are m, $a_i - m$ and $b_j - m$, where m is the sample mean of all the observations and a_i and b_j are the means of observations in the ith year and the jth month respectively. Thus the residuals are given by

$$z_{ij} = y_{ij} - a_i - b_j + m.$$

The test is made as before by calculating

$$d = \frac{\Sigma(\Delta z_{ij})^2}{\Sigma z_{ij}^2},$$

where the Δz_{ij}'s are the first differences of the residuals when arranged as a single time series.

$\Sigma(\Delta z)^2$ may be calculated by working out the individual residuals from the monthly averages and finding their first differences. The difference between the December value of one year and the January value of the succeeding year then needs to be adjusted to take account of the difference between the two yearly averages. For instance, the difference between the 'months only' residuals for December 1933 and January 1934 is

$$2{\cdot}50 - 9{\cdot}70 = -7{\cdot}20.$$

From this must be subtracted the difference between the 1934 and 1933 yearly averages, i.e. $58{\cdot}22 - 63{\cdot}42 = -5{\cdot}20$. The net difference is therefore $-7{\cdot}20 + 5{\cdot}20 = -2{\cdot}00$. The sum of the resulting differences squared is $\Sigma(\Delta z)^2$. For the calculation of Σz^2 the normal method for the residual sum of squares may be used, i.e. find the sum of squares of the original observations and subtract the sums of squares due to the fitted constants. Alternatively, with the above method of calculating $\Sigma(\Delta z)^2$, the sum of squares of the residuals from the monthly averages may be calculated directly, from which it only remains to subtract the sum of squares due to years.

For the data in the table we find $\Sigma(\Delta z)^2 = 1191{\cdot}2454$ and $\Sigma z^2 = 850{\cdot}8890$, giving $d = \dfrac{1191{\cdot}2454}{850{\cdot}8890} = 1{\cdot}4000$. It remains to test the significance of this value of d by the method of § 4. For this purpose the formulae (12) may be evaluated with $s = 5$ and $t = 12$ to give

$$E(d) = 2{\cdot}1303, \quad \operatorname{var} d = 0{\cdot}077964.$$

If $\frac{1}{4}d$ is assumed to have a Beta-distribution with parameters p, q, then values of $E(d)$ and $\operatorname{var} d$ may be substituted in the formulae (9) to give

$$p = 26{\cdot}6758, \quad q = 23{\cdot}4125.$$

The observed value of d is 1·4000, so that

$$F = \frac{p(4-d)}{pd} = 2{\cdot}16,$$

with $n_1 \sim 47$, $n_2 \sim 53$ degrees of freedom. A cursory examination of the significant points of F shows that the 5 % point is certainly less than 1·63, while the 1 % point is less than 2·00. Thus our value of F is significant at the 1 % level, and the hypothesis of serial independence of the errors in the above model may be considered untenable.

6. Polynomial regressions

An important application of the least squares method in time-series analysis is in the fitting of polynomial trend lines. When the values of the time variate (or its equivalent in other applications) are spaced at equal intervals, the fitting is carried out most expeditiously by

means of orthogonal polynomials. In what follows we shall assume that the ξ' polynomials tabulated by Fisher & Yates (1948), and in a more extended form by Anderson & Houseman (1942), have been used. The regression model is

$$y = \beta_0 + \beta_1 \xi_1' + \beta_2 \xi_2' + \ldots + \beta_{k'} \xi_{k'}' + \epsilon,$$

where ξ_1' is the polynomial of ith degree in x, the independent variable, which we suppose takes the values $1, 2, \ldots, n$, $\beta_0, \beta_1, \ldots, \beta_{k'}$, are constants and ϵ is the error term. We require to test the error term for serial correlation.

The test procedure is a good deal less laborious in this application than in the ordinary regression case described in § 2. We shall illustrate it by considering the following example. The data were taken by Anderson & Houseman (1942) from Schultz's demand studies (1938) to illustrate the routine procedure of fitting the polynomials. We shall not, therefore, give details of the initial calculations but refer the reader instead to Anderson & Houseman's bulletin.

Table 3. *Price of sugar*, 1875–1936

Year	Price	Year	Price	Year	Price	Year	Price
1875	67	1891	6	1907	6	1923	44
1876	65	1892	3	1908	10	1924	35
1877	73	1893	8	1909	8	1925	15
1878	55	1894	1	1910	10	1926	15
1879	48	1895	2	1911	13	1927	18
1880	56	1896	5	1912	10	1928	15
1881	57	1897	5	1913	3	1929	10
1882	52	1898	10	1914	7	1930	6
1883	45	1899	9	1915	16	1931	4
1884	28	1900	13	1916	29	1932	0
1885	24	1901	10	1917	37	1933	3
1886	21	1902	5	1918	38	1934	1
1887	20	1903	6	1919	50	1935	3
1888	30	1904	8	1920	74	1936	7
1889	36	1905	13	1921	22		
1890	22	1906	5	1922	19		

Example 3. A polynomial time trend is to be fitted to 62 annual sugar prices (1875–1936) given in Table 3. The prices are in terms of mills (tenths of a cent) coded by subtracting 40 mills from each price.

Anderson & Houseman find the following values for the sums of squares and products, giving the regression coefficients shown:

$$\begin{aligned}
\Sigma y &= 1{,}336 & b_0 &= 21{\cdot}548 \\
\Sigma(y-\bar{y})^2 &= 25{,}250 & & \\
\Sigma(\xi_1')^2 &= 79{,}422 & \Sigma(\xi_2')^2 &= 1{,}270{,}752 \\
\Sigma(\xi_3')^2 &= 139{,}238{,}112 & \Sigma(\xi_4')^2 &= 103{,}639{,}568{,}032 \\
\Sigma(y\xi_1') &= -20{,}286 & b_1 &= -0{\cdot}2554 \\
\Sigma(y\xi_2') &= 72{,}775 & b_2 &= 0{\cdot}05727 \\
\Sigma(y\xi_3') &= -1{,}080{,}557 & b_3 &= -0{\cdot}0077605 \\
\Sigma(y\xi_4') &= -7{,}599{,}201 & b_4 &= -0{\cdot}00007332
\end{aligned}$$

To test for serial correlation we must calculate

$$d = \frac{\Sigma(\Delta z)^2}{\Sigma z^2}.$$

Taking first the case in which terms only as far as ξ_3' are fitted,

$$z = y - b_0 - b_1\xi_1' - b_2\xi_2' - b_3\xi_3'$$

and

$$\begin{aligned}\Sigma z^2 &= 25{,}250 - (-0{\cdot}2554)(-20{,}286) - (0{\cdot}05727)(72{,}775) - (-0{\cdot}0077605)(-1{,}080{,}557)\\ &= 7515,\end{aligned} \tag{13}$$

$$\Sigma(\Delta z)^2 = \Sigma(\Delta y)^2 - 2b_1\Sigma\Delta y\Delta\xi_1' - 2b_2\Sigma\Delta y\Delta\xi_2' - 2b_3\Sigma\Delta y\Delta\xi_3' + \sum_{i=1}^{3}\sum_{j=1}^{3} b_ib_j\Sigma\Delta\xi_i'\Delta\xi_j'. \tag{14}$$

$\Sigma(\Delta y)^2$ is calculated directly as the sum of the squares of the first differences of the series of observations of y; its value here is 2590. For the remaining terms, indirect methods are much quicker. It may be verified that

$$\left.\begin{aligned}
\Sigma\Delta y\Delta\xi_1' &= (y_n - y_1)\Delta\xi_1'(0),\\
\Sigma\Delta y\Delta\xi_2' &= -2\lambda_2\Sigma y - (y_n + y_1)\Delta\xi_2'(0),\\
\Sigma\Delta y\Delta\xi_3' &= -\frac{6\lambda_3}{\lambda_1}\Sigma y\xi_1' + (y_n - y_1)\Delta\xi_3'(0),\\
\Sigma\Delta y\Delta\xi_4' &= -\lambda_4\left[\frac{12}{\lambda_2}\Sigma y\xi_2' + \left\{\frac{1}{\lambda_2}(n^2-1) - \frac{3n^2-13}{7}\right\}\Sigma y\right] - (y_n + y_1)\Delta\xi_4'(0),\\
\Sigma\Delta y\Delta\xi_5' &= \lambda_5\left[\frac{20}{\lambda_3}\Sigma y\xi_3' + \left\{\frac{1}{\lambda_3}(3n^2-7) - \frac{10}{\lambda_1}\frac{n^2-13}{3}\right\}\Sigma y\xi_1'\right] + (y_n - y_1)\Delta\xi_5'(0),
\end{aligned}\right\} \tag{15}$$

and that, with $(i \leqslant j)$,

$$\left.\begin{aligned}
\Sigma\Delta\xi_i'\Delta\xi_j' &= -2\xi_j'(1)\Delta\xi_i'(0) \quad (i+j \text{ even}),\\
&= 0 \quad (i+j \text{ odd}).
\end{aligned}\right\} \tag{16}$$

In these expressions it is assumed that the original time variate x takes the values $1, 2, \ldots, n$. $\xi_j'(1)$ denotes the value of ξ_j' for $x = 1$. Similarly, $\Delta\xi_j'(0)$ denotes the value of $\Delta\xi_j'$ for $x = 0$, i.e. $\xi_j'(1) - \xi_j'(0)$. y_1 and y_n are the first and last observations in the series of values of the dependent variate. The λ_i is given for each n at the foot of the appropriate column of ξ_i' values in the published tables.

The values of $\Delta\xi_i'(0)$ are obtained by writing down the first few terms of the series $\xi_i'(x)$ for $x = 1, 2, \ldots$, and preparing a small table of differences. $\Delta\xi_i'(0)$ is then found by simple addition. It should be noted that the values of $\xi_i'(x)$ should be read upwards starting at the bottom of the published table, and that the signs should be reversed for polynomials of odd degree. Example 3, with $n = 62$, gives the following lay-out. The values of $\Delta\xi_i'(0)$ required are printed in italics:

x	ξ_1'	$\Delta\xi_1'$	x	ξ_2'	$\Delta\xi_2'$	$\Delta^2\xi_2'$	x	ξ_3'	$\Delta\xi_3'$	$\Delta^2\xi_3'$	$\Delta^3\xi_3'$
		2			*−31*				*769*		
1	−61	2	1	305	−30	1	1	−3599	708	−61	2
2	−59	2	2	275	−29	1	2	−2891	649	−59	2
3	−57		3	246	−28	1	3	−2242	592	−57	2
			4	218			4	−1650	437	−55	
							5	−1113			

Substituting in formulae (15, 16) we have

$$\begin{aligned}
\Sigma\Delta y\,\Delta\xi_1' &= (7-67)(2) = -120,\\
\Sigma\Delta y\,\Delta\xi_2' &= (-2)(\tfrac{1}{2})(1336)-(7+67)(-31) = 958,\\
\Sigma\Delta y\,\Delta\xi_3' &= -\frac{6}{2\cdot 3}(-20{,}286)+(7-67)\,769 = -25{,}854,\\
\Sigma(\Delta\xi_1')^2 &= -2(-61)(2) = 244,\\
\Sigma(\Delta\xi_2')^2 &= -2(305)(-31) = 18{,}910,\\
\Sigma(\Delta\xi_3')^2 &= -2(-3599)(769) = 5{,}535{,}262,\\
\Sigma\Delta\xi_1'\Delta\xi_2' &= \Sigma\Delta\xi_2'\Delta\xi_3' = 0,\\
\Sigma\Delta\xi_1'\Delta\xi_3' &= -2(-3599)(2) = 14{,}396.
\end{aligned}$$

We now have all the quantities necessary for the calculation of $\Sigma(\Delta z)^2$. Substituting in (14) we have

$$\begin{aligned}
\Sigma(\Delta z)^2 = {} & 2590-2(-0{\cdot}2554)(120)-2(0{\cdot}05727)(958)\\
& -2(-0{\cdot}0077605)(-25{,}854)+2(-0{\cdot}2554)(-0{\cdot}0077605)(14{,}396)\\
& +(0{\cdot}2554)^2(244)+(0{\cdot}05727)^2(18{,}910)+(0{\cdot}0077605)^2(5{,}535{,}262)\\
= {} & 2486{\cdot}06.
\end{aligned}$$

Thus
$$d = \frac{2486{\cdot}06}{7515} = 0{\cdot}3308.$$

Reference to Table 6 shows that this value is significant at the 1 % level and therefore provides significant evidence of the existence of positive serial correlation.

When a polynomial of the fourth degree is fitted to the data a value of d of 0·3603 is obtained. This still remains highly significant at the 1 % level.

Mean and variance of d

The quickest way of calculating the mean and variance of d for polynomial regressions is to use the numerical procedure described below. It is, however, possible to obtain explicit formulae, and these have been calculated for polynomials up to the fifth degree. Owing to the complexity of the resulting expressions we shall only give them for $k' = 1, 2, 3$.

Taking first the numerical procedure, let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ denote the column vectors of values of the polynomials (the usual prime is omitted in the vector form since we wish to use the same sign for the matrix operation of transposition). The mean and variance of d are given by (5) and (6) with

$$P = 2(n-1)-\sum_{i=1}^{k'}\frac{\boldsymbol{\xi}_i'\mathbf{A}\boldsymbol{\xi}_i}{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i},$$

$$Q = 2(3n-4)-2\sum_{i=1}^{k'}\frac{\boldsymbol{\xi}_i'\mathbf{A}^2\boldsymbol{\xi}_i}{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}+\sum_{i=1}^{k'}\left(\frac{\boldsymbol{\xi}_i'\mathbf{A}\boldsymbol{\xi}_i}{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i}\right)^2+2\sum_{j=2}^{k'}\sum_{i=1}^{j-1}\frac{(\boldsymbol{\xi}_i'\mathbf{A}\boldsymbol{\xi}_i)^2}{\boldsymbol{\xi}_i'\boldsymbol{\xi}_i\,\boldsymbol{\xi}_j'\boldsymbol{\xi}_j},$$

as was shown in Part I, **A** being the matrix defined in § 4. The quantities $\boldsymbol{\xi}_i'\boldsymbol{\xi}_i$ are the sums of squares of the values of ξ_i', given at the foot of the ξ' tables. The quantities $\boldsymbol{\xi}_i'\mathbf{A}\boldsymbol{\xi}_j$ and $\boldsymbol{\xi}_i'\mathbf{A}^2\boldsymbol{\xi}_i$ may be found from the expressions

$$\boldsymbol{\xi}_i'\mathbf{A}\boldsymbol{\xi}_j = \begin{cases} -2\xi_j'(1)\,\Delta\xi_i'(0) & (i+j \text{ even}),\\ 0 & (i+j \text{ odd}),\end{cases}$$

$$\boldsymbol{\xi}_i'\mathbf{A}^2\boldsymbol{\xi}_i = 2\xi_i'(1)\,\Delta^3\xi_i'(-1)+2\Delta\xi_i'(0)\,\Delta\xi_i'(1).$$

The values of $\xi_i'(x)$ and its differences may easily be found from the published tables.

Applying the method to Example 3 we find

$$\begin{aligned} P &= 122-\frac{244}{79,422}-\frac{18,910}{1,270,752}-\frac{5,535,262}{139,238,112} \\ &= 121{\cdot}9423, \\ Q &= 368-2\left(\frac{8}{79,422}+\frac{1860}{1,270,752}+\frac{1,074,508}{139,238,112}\right) \\ &\quad +\left(\frac{244}{79,422}\right)^2+\left(\frac{18,910}{1,270,752}\right)^2+\left(\frac{5,535,262}{139,238,112}\right)^2 \\ &\quad +2\left(\frac{14,396}{79,422\times 139,238,112}\right) \\ &= 367{\cdot}9832. \end{aligned}$$

The explicit formulae for P and Q are as follows:

$$P = 2\left(n-1-\sum_{i=1}^{k'} p_i\right),$$

where $$p_1 = \frac{6}{n(n+1)}, \quad p_2 = \frac{30}{(n+1)(n+2)}, \quad p_3 = \frac{84(n^2+1)}{n(n+1)(n+2)(n+3)},$$

and $$Q = 2\left(3n-4+\sum_{i=1}^{k'} q_i\right),$$

where $$q_1 = 2p_1^2-\frac{24}{n(n^2-1)}, \quad q_2 = 2p_2^2-\frac{360}{(n^2-1)(n+2)},$$

$$q_3 = 2p_3^2+\frac{336(n-2)(n-3)}{n^2(n+1)^2(n+2)(n+3)}-\frac{336(3n^2+10n-7)}{n(n^2-1)(n+2)(n+3)}.$$

Appendix on the calculation of the tables

1. *Tables* 4, 5 *and* 6. The exact distribitions of d_L and d_U, whose significance points are required for Tables 4, 5 and 6, are not known. When transformed to the range (0, 1), their probability densities may, however, be represented by the series

$$g(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}\left\{1+\sum_{s=1}^{\infty} a_s G_s(x)\right\}, \tag{17}$$

where the G's are the polynomials (Jacobi) which are orthogonal on the range (0, 1) with respect to the weight function $\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$. These polynomials are defined by (see Courant & Hilbert,* 1931)

$$\begin{aligned} G_s(x) &= 1-\frac{p+q-1+s}{p}x+\frac{(p+q-1+s)(p+q+s)}{p(p+1)}x^2-\ldots \\ &\quad +(-1)^s\frac{(p+q-1+s)(p+q+s)\ldots(p+q+2-2s)}{p(p+1)\ldots(p+s-1)}x^s. \end{aligned}$$

The coefficients a_s may be determined by the method of moments; the distribution of d_L and d_U is then a series of Incomplete Beta Functions.

* Note, however, the misprint: $x^q(1-x)^{p-q}$ should read $x^{q-1}(1-x)^{p-q}$.

The weight function was chosen to be the density of the Beta-distribution with the correct mean and variance. (An alternative weight function giving the right order of vanishing of (17) at $x = 0$ and $x = 1$ was also tried but it was found to be less satisfactory.) With this weight function, the coefficients a_1 and a_2 in (17) were zero. Terms as far as $G_4(x)$ were used.

Table 4. *Significance points of d_L and d_U: 5 %*

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1·08	1·36	0·95	1·54	0·82	1·75	0·69	1·97	0·56	2·21
16	1·10	1·37	0·98	1·54	0·86	1·73	0·74	1·93	0·62	2·15
17	1·13	1·38	1·02	1·54	0·90	1·71	0·78	1·90	0·67	2·10
18	1·16	1·39	1·05	1·53	0·93	1·69	0·82	1·87	0·71	2·06
19	1·18	1·40	1·08	1·53	0·97	1·68	0·86	1·85	0·75	2·02
20	1·20	1·41	1·10	1·54	1·00	1·68	0·90	1·83	0·79	1·99
21	1·22	1·42	1·13	1·54	1·03	1·67	0·93	1·81	0·83	1·96
22	1·24	1·43	1·15	1·54	1·05	1·66	0·96	1·80	0·86	1·94
23	1·26	1·44	1·17	1·54	1·08	1·66	0·99	1·79	0·90	1·92
24	1·27	1·45	1·19	1·55	1·10	1·66	1·01	1·78	0·93	1·90
25	1·29	1·45	1·21	1·55	1·12	1·66	1·04	1·77	0·95	1·89
26	1·30	1·46	1·22	1·55	1·14	1·65	1·06	1·76	0·98	1·88
27	1·32	1·47	1·24	1·56	1·16	1·65	1·08	1·76	1·01	1·86
28	1·33	1·48	1·26	1·56	1·18	1·65	1·10	1·75	1·03	1·85
29	1·34	1·48	1·27	1·56	1·20	1·65	1·12	1·74	1·05	1·84
30	1·35	1·49	1·28	1·57	1·21	1·65	1·14	1·74	1·07	1·83
31	1·36	1·50	1·30	1·57	1·23	1·65	1·16	1·74	1·09	1·83
32	1·37	1·50	1·31	1·57	1·24	1·65	1·18	1·73	1·11	1·82
33	1·38	1·51	1·32	1·58	1·26	1·65	1·19	1·73	1·13	1·81
34	1·39	1·51	1·33	1·58	1·27	1·65	1·21	1·73	1·15	1·81
35	1·40	1·52	1·34	1·58	1·28	1·65	1·22	1·73	1·16	1·80
36	1·41	1·52	1·35	1·59	1·29	1·65	1·24	1·73	1·18	1·80
37	1·42	1·53	1·36	1·59	1·31	1·66	1·25	1·72	1·19	1·80
38	1·43	1·54	1·37	1·59	1·32	1·66	1·26	1·72	1·21	1·79
39	1·43	1·54	1·38	1·60	1·33	1·66	1·27	1·72	1·22	1·79
40	1·44	1·54	1·39	1·60	1·34	1·66	1·29	1·72	1·23	1·79
45	1·48	1·57	1·43	1·62	1·38	1·67	1·34	1·72	1·29	1·78
50	1·50	1·59	1·46	1·63	1·42	1·67	1·38	1·72	1·34	1·77
55	1·53	1·60	1·49	1·64	1·45	1·68	1·41	1·72	1·38	1·77
60	1·55	1·62	1·51	1·65	1·48	1·69	1·44	1·73	1·41	1·77
65	1·57	1·63	1·54	1·66	1·50	1·70	1·47	1·73	1·44	1·77
70	1·58	1·64	1·55	1·67	1·52	1·70	1·49	1·74	1·46	1·77
75	1·60	1·65	1·57	1·68	1·54	1·71	1·51	1·74	1·49	1·77
80	1·61	1·66	1·59	1·69	1·56	1·72	1·53	1·74	1·51	1·77
85	1·62	1·67	1·60	1·70	1·57	1·72	1·55	1·75	1·52	1·77
90	1·63	1·68	1·61	1·70	1·59	1·73	1·57	1·75	1·54	1·78
95	1·64	1·69	1·62	1·71	1·60	1·73	1·58	1·75	1·56	1·78
100	1·65	1·69	1·63	1·72	1·61	1·74	1·59	1·76	1·57	1·78

A first set of significance points was obtained using the weight function as a first approximation; these values were then adjusted using the higher terms of the series. The first set of values was calculated partly by Wise's* (1950) method and partly by means of Carter's (1947) approximation. The adjustments necessary were found to be small and to vary very slowly with p and q; they could therefore be calculated by the following method which reduces to a minimum interpolation in the *Tables of the Incomplete Beta Function* (1948).

* We are indebted to Mr Wise for some helpful correspondence on his method.

First, p and q may be replaced by the integers nearest them. For these integers an exact significance point may be found by quadratic inverse interpolation. The difference of this exact point and the first approximation to it is the required adjustment. The adjustment required for the fourth moment turned out to be negligible to the order of accuracy aimed at and could have been ignored. The adjustments were so small and regular that it was only

Table 5. *Significance points of* d_L *and* d_U: 2·5 %

n	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0·95	1·23	0·83	1·40	0·71	1·61	0·59	1·84	0·48	2·09
16	0·98	1·24	0·86	1·40	0·75	1·59	0·64	1·80	0·53	2·03
17	1·01	1·25	0·90	1·40	0·79	1·58	0·68	1·77	0·57	1·98
18	1·03	1·26	0·93	1·40	0·82	1·56	0·72	1·74	0·62	1·93
19	1·06	1·28	0·96	1·41	0·86	1·55	0·76	1·72	0·66	1·90
20	1·08	1·28	0·99	1·41	0·89	1·55	0·79	1·70	0·70	1·87
21	1·10	1·30	1·01	1·41	0·92	1·54	0·83	1·69	0·73	1·84
22	1·12	1·31	1·04	1·42	0·95	1·54	0·86	1·68	0·77	1·82
23	1·14	1·32	1·06	1·42	0·97	1·54	0·89	1·67	0·80	1·80
24	1·16	1·33	1·08	1·43	1·00	1·54	0·91	1·66	0·83	1·79
25	1·18	1·34	1·10	1·43	1·02	1·54	0·94	1·65	0·86	1·77
26	1·19	1·35	1·12	1·44	1·04	1·54	0·96	1·65	0·88	1·76
27	1·21	1·36	1·13	1·44	1·06	1·54	0·99	1·64	0·91	1·75
28	1·22	1·37	1·15	1·45	1·08	1·54	1·01	1·64	0·93	1·74
29	1·24	1·38	1·17	1·45	1·10	1·54	1·03	1·63	0·96	1·73
30	1·25	1·38	1·18	1·46	1·12	1·54	1·05	1·63	0·98	1·73
31	1·26	1·39	1·20	1·47	1·13	1·55	1·07	1·63	1·00	1·72
32	1·27	1·40	1·21	1·47	1·15	1·55	1·08	1·63	1·02	1·71
33	1·28	1·41	1·22	1·48	1·16	1·55	1·10	1·63	1·04	1·71
34	1·29	1·41	1·24	1·48	1·17	1·55	1·12	1·63	1·06	1·70
35	1·30	1·42	1·25	1·48	1·19	1·55	1·13	1·63	1·07	1·70
36	1·31	1·43	1·26	1·49	1·20	1·56	1·15	1·63	1·09	1·70
37	1·32	1·43	1·27	1·49	1·21	1·56	1·16	1·62	1·10	1·70
38	1·33	1·44	1·28	1·50	1·23	1·56	1·17	1·62	1·12	1·70
39	1·34	1·44	1·29	1·50	1·24	1·56	1·19	1·63	1·13	1·69
40	1·35	1·45	1·30	1·51	1·25	1·57	1·20	1·63	1·15	1·69
45	1·39	1·48	1·34	1·53	1·30	1·58	1·25	1·63	1·21	1·69
50	1·42	1·50	1·38	1·54	1·34	1·59	1·30	1·64	1·26	1·69
55	1·45	1·52	1·41	1·56	1·37	1·60	1·33	1·64	1·30	1·69
60	1·47	1·54	1·44	1·57	1·40	1·61	1·37	1·65	1·33	1·69
65	1·49	1·55	1·46	1·59	1·43	1·62	1·40	1·66	1·36	1·69
70	1·51	1·57	1·48	1·60	1·45	1·63	1·42	1·66	1·39	1·70
75	1·53	1·58	1·50	1·61	1·47	1·64	1·45	1·67	1·42	1·70
80	1·54	1·59	1·52	1·62	1·49	1·65	1·47	1·67	1·44	1·70
85	1·56	1·60	1·53	1·63	1·51	1·65	1·49	1·68	1·46	1·71
90	1·57	1·61	1·55	1·64	1·53	1·66	1·50	1·69	1·48	1·71
95	1·58	1·62	1·56	1·65	1·54	1·67	1·52	1·69	1·50	1·71
100	1·59	1·63	1·57	1·65	1·55	1·67	1·53	1·70	1·51	1·72

necessary to calculate them at 39 places in the entire set of tables, the remainder being obtained by linear interpolation. The adjustments were negligible for numbers of observations greater than 40.

As a partial check on the calculating procedure it was applied to the calculation of the significance points for a related distribution for which exact significance points were available. To make the circumstances as unfavourable as possible the case $n = 16$, $k' = 5$ at the

Table 6. *Significance points of d_L and d_U*: 1 %

	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0·81	1·07	0·70	1·25	0·59	1·46	0·49	1·70	0·39	1·96
16	0·84	1·09	0·74	1·25	0·63	1·44	0·53	1·66	0·44	1·90
17	0·87	1·10	0·77	1·25	0·67	1·43	0·57	1·63	0·48	1·85
18	0·90	1·12	0·80	1·26	0·71	1·42	0·61	1·60	0·52	1·80
19	0·93	1·13	0·83	1·26	0·74	1·41	0·65	1·58	0·56	1·77
20	0·95	1·15	0·86	1·27	0·77	1·41	0·68	1·57	0·60	1·74
21	0·97	1·16	0·89	1·27	0·80	1·41	0·72	1·55	0·63	1·71
22	1·00	1·17	0·91	1·28	0·83	1·40	0·75	1·54	0·66	1·69
23	1·02	1·19	0·94	1·29	0·86	1·40	0·77	1·53	0·70	1·67
24	1·04	1·20	0·96	1·30	0·88	1·41	0·80	1·53	0·72	1·66
25	1·05	1·21	0·98	1·30	0·90	1·41	0·83	1·52	0·75	1·65
26	1·07	1·22	1·00	1·31	0·93	1·41	0·85	1·52	0·78	1·64
27	1·09	1·23	1·02	1·32	0·95	1·41	0·88	1·51	0·81	1·63
28	1·10	1·24	1·04	1·32	0·97	1·41	0·90	1·51	0·83	1·62
29	1·12	1·25	1·05	1·33	0·99	1·42	0·92	1·51	0·85	1·61
30	1·13	1·26	1·07	1·34	1·01	1·42	0·94	1·51	0·88	1·61
31	1·15	1·27	1·08	1·34	1·02	1·42	0·96	1·51	0·90	1·60
32	1·16	1·28	1·10	1·35	1·04	1·43	0·98	1·51	0·92	1·60
33	1·17	1·29	1·11	1·36	1·05	1·43	1·00	1·51	0·94	1·59
34	1·18	1·30	1·13	1·36	1·07	1·43	1·01	1·51	0·95	1·59
35	1·19	1·31	1·14	1·37	1·08	1·44	1·03	1·51	0·97	1·59
36	1·21	1·32	1·15	1·38	1·10	1·44	1·04	1·51	0·99	1·59
37	1·22	1·32	1·16	1·38	1·11	1·45	1·06	1·51	1·00	1·59
38	1·23	1·33	1·18	1·39	1·12	1·45	1·07	1·52	1·02	1·58
39	1·24	1·34	1·19	1·39	1·14	1·45	1·09	1·52	1·03	1·58
40	1·25	1·34	1·20	1·40	1·15	1·46	1·10	1·52	1·05	1·58
45	1·29	1·38	1·24	1·42	1·20	1·48	1·16	1·53	1·11	1·58
50	1·32	1·40	1·28	1·45	1·24	1·49	1·20	1·54	1·16	1·59
55	1·36	1·43	1·32	1·47	1·28	1·51	1·25	1·55	1·21	1·59
60	1·38	1·45	1·35	1·48	1·32	1·52	1·28	1·56	1·25	1·60
65	1·41	1·47	1·38	1·50	1·35	1·53	1·31	1·57	1·28	1·61
70	1·43	1·49	1·40	1·52	1·37	1·55	1·34	1·58	1·31	1·61
75	1·45	1·50	1·42	1·53	1·39	1·56	1·37	1·59	1·34	1·62
80	1·47	1·52	1·44	1·54	1·42	1·57	1·39	1·60	1·36	1·62
85	1·48	1·53	1·46	1·55	1·43	1·58	1·41	1·60	1·39	1·63
90	1·50	1·54	1·47	1·56	1·45	1·59	1·43	1·61	1·41	1·64
95	1·51	1·55	1·49	1·57	1·47	1·60	1·45	1·62	1·42	1·64
100	1·52	1·56	1·50	1·58	1·48	1·60	1·46	1·63	1·44	1·65

extreme of the tabulated values was examined. The distributions of d_L and d_U for these values were modified to give latent roots occurring in equal pairs, i.e. five pairs in all. The new roots were chosen to lie midway between the roots of the original distribution, thus preserving its asymmetry and general character. By pairing the roots in this way the exact significance points could be determined using results given by R. L. Anderson (1942). The significance points obtained by the approximate procedure agreed with these exact significance points to the order of accuracy required here.

7. AN EXACT BOUNDS TEST

We have given elsewhere (Watson & Durbin, 1951) a general method of constructing exact tests of serial independence which do not require the use of circular definitions of the serial correlation coefficient. The method can be used to obtain the exact distributions of bounding

statistics similar to d_L and d_U. The advantage of having exact distributions is obtained at the cost of throwing away a certain amount of relevant information.

For testing the independence of the errors in a regression model a statistic d' is defined which is a slight modification of d. If the number of observations is even, $2m$ say, then

$$d' = \frac{(z_1 - z_2)^2 + \ldots + (z_{m-1} - z_m)^2 + (z_{m+1} - z_{m+2})^2 + \ldots + (z_{2m-1} - z_{2m})^2}{\sum_1^{2m} z_i^2},$$

and if it is odd, $2m + 1$ say, then

$$d' = \frac{(z_1 - z_2)^2 + \ldots + (z_{m-1} - z_m)^2 + (z_{m+2} - z_{m+3})^2 + \ldots + (z_{2m} - z_{2m+1})^2}{\sum_1^{2m+1} z_i^2},$$

where the z's are the least-squares residuals. The only difference from d is that one or two of the squared differences are omitted from the numerator of d'. Thus when m is small a substantial fraction of the relevant information is sacrificed.

The theory developed in Part I can be applied to show that d' lies between two values d'_L and d'_U. In contrast to d, exact significance points can be calculated for d'_L and d'_U using the results of R. L. Anderson (1942), except for the case of an even number of observations and an odd number of independent variables, for which the exact distribution of d'_L has not been found. A short table of such values for odd numbers of observations is given in Table 7. This table may be used for testing the significance of an observed value of d' in exactly the same way as Table 4 is used for testing the significance of an observed value of d.

Table 7. *Significance points of d'_L and d'_U*: 5 %

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
	d'_L	d'_U	d'_L	d'_U	d'_L	d'_U	d'_L	d'_U	d'_L	d'_U
13	0·69	0·97	0·56	1·15	—	—	—	—	—	—
15	0·80	1·04	0·67	1·20	0·55	1·36	—	—	—	—
17	0·89	1·11	0·77	1·24	0·65	1·38	0·54	1·57	0·44	1·74
19	0·96	1·16	0·85	1·27	0·75	1·40	0·64	1·56	0·54	1·71
21	1·02	1·20	0·92	1·30	0·82	1·42	0·72	1·56	0·63	1·69
23	1·07	1·24	0·98	1·33	0·89	1·44	0·80	1·56	0·71	1·68

The calculations required even for such a short table were very heavy, and our chief motive for including it is the satisfaction of demonstrating that the problem has an exact solution. In practice we ourselves would prefer to use Table 4 owing to the greater power and simplicity of the statistic d.

We wish to express our most grateful thanks to Mrs E. G. Chambers and the computing staff of the Department of Applied Economics, Cambridge, for carying out most of the calculations required for Tables 4–7; also to Miss J. Graham of the Division of Research Techniques, London School of Economics, for assisting with Table 7.

REFERENCES

ANDERSON, R. L. (1942). *Ann. Math. Statist.* **13**, 1.
ANDERSON, R. L. & ANDERSON, T. W. (1950). *Ann. Math. Statist.* **21**, 59.
ANDERSON, R. L. & HOUSEMAN, E. E. (1942). Tables of orthogonal polynomial values extended to $N = 104$. *Res. Bull. Iowa St. Coll.* no. 297.
CARTER, A. H. (1947). *Biometrika*, **34**, 352.
COCHRAN, W. G. (1938). *J.R. Statist. Soc., Suppl.*, **5**, 171.
COURANT, R. & HILBERT, D. (1931). *Methoden der Mathematischen Physik.* Berlin: Julius Springer.
DURBIN, J. & WATSON, G. S. (1950). *Biometrika*, **37**, 409.
FISHER, R. A. & YATES, F. (1948). *Statistical Tables.* Edinburgh: Oliver and Boyd.
PEARSON, K. (1948). *Tables of the Incomplete Beta Function.* Cambridge University Press.
PREST, A. R. (1949). *Rev. Econ. Statist.* **31**, 33.
SCHULTZ, HENRY (1938). *Theory and Measurement of Demand*, pp. 674–7. University of Chicago Press.
SNEDECOR, G. W. (1937). *Statistical Methods.* Collegiate Press.
THOMPSON, CATHERINE (1941). *Biometrika*, **32**, 151.
WATSON, G. S. & DURBIN, J. (1951). Exact tests of serial correlation using non-circular statistics. (To be published).
WISE, M. E. (1950). *Biometrika*, **37**, 208.

CORRECTIONS TO PART I. (Durbin & Watson, 1950)

We are grateful to Prof. T. W. Anderson for pointing out an error in the section headed 'Inequalities on $\nu_1, \nu_2, \ldots, \nu_{n-k}$'. Part (*b*) of the lemma and its corollary have been correctly applied in the remaining parts of the paper, but are incorrectly stated.

The necessary corrections are as follows:

p. 415. The second paragraph beginning 'We therefore seek...' should read:

'We therefore seek inequalities on $\nu_1, \nu_2, \ldots, \nu_{n-k}$. For the sake of generality we suppose that certain of the regression vectors, say s of them, coincide with latent vectors of **A** (or are linear combinations of them). From the results of the previous section it follows that the problem is reduced to the consideration of $k-s$ arbitrary regression vectors, while **A** may be supposed to have s zero roots together with the roots of **A** not associated with the s latent vectors mentioned above. These roots may be renumbered so that

$$\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_{n-s}.$$

We proceed to show that $\lambda_i \leqslant \nu_i \leqslant \lambda_{i+k-s} \quad (i = 1, 2, \ldots, n-k).$' (3)

p. 415. The next to the last sentence should read:

'Allowing for cases in which regression vectors coincide with latent vectors of **A** we have (3).'

p. 416. Lemma (*b*) should read:

'If s of the columns of **X** are linear combinations of s of the latent vectors of **A**, and if the roots of **A** associated with the remaining $n-s$ latent vectors of **A** are renumbered so that

$$\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_{n-s},$$

then $$\lambda_i \leqslant \nu_i \leqslant \lambda_{i+k-s} \quad (i = 1, 2, \ldots, n-k).$'

p. 416. Corollary should read: '$r_L \leqslant r \leqslant r_U$,

where $$r_L = \frac{\sum_{i=1}^{n-k} \lambda_i \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2}$$

and $$r_U = \frac{\sum_{i=1}^{n-k} \lambda_{i+k-s} \zeta_i^2}{\sum_{i=1}^{n-k} \zeta_i^2}.$'