THE ECONOMIC RESEARCH INSTITUTE

Memorandum No.16

Correcting for Seasonality: I The Regression Method, with an Application to Irish Data

by

R.C. Geary

Let the data be observed for T consecutive years with k(= 12 for months, = 4 for quarters) length of seasons. The observations are Y_t, $t = kt' + s$, $t' = 1,2,\ldots,T$, $s = 1,2,\ldots,k$. Set up the regression

$$(1)\quad y_t = b_1 x_{t1} + b_2 x_{t2} + \ldots + b_p x_{tp} + c_1 z_{t1} + c_2 z_{t2} + \ldots + c_k z_{tk} + u_t,$$

where $y_t = Y_t - \bar{Y}$. The x_{ti} are the trend (regarded as orthogonal) and the z_{tj} are the seasonality variables, u_t the random residual. The z_{tj} elements are all zeros or unities. Their Tk x k matrix Z is

$$(2)\quad Z = |z_{tj}| = \begin{vmatrix} I \\ I \\ \vdots \\ I \end{vmatrix}$$

where I is the unit k x k matrix and these are T such matrices. Since the x_{ti} are orthogonal,

$$(3)\quad \Sigma_t\, x_{ti}\, x_{ti'},\ (i \neq i') = 0.$$

Similarly for the z_{tj}:

$$(4)\quad \Sigma_t\, z_{tj}\, z_{tj'},\ (j \neq j') = 0$$

The regression solution of (1) in the b_i and c_j is found from

$$(5)\quad \begin{aligned} &(i)\ \Sigma_t\, x_{ti} y_t = b_i \Sigma x_{ti}^2 + \Sigma_j c_j \Sigma_t x_{ti} z_{tj},\ i = 1,2,\ldots,p \\ &(ii)\ \Sigma_t\, z_{tj} y_t = \Sigma_i b_i \Sigma_t z_{tj} x_{ti} + T c_j,\ j = 1,2,\ldots,k \end{aligned}$$

On account of the orthogonal property this

system is very easily set up and solved. It can readily be shown from (1) that

$$(6) \qquad \sum_{j=1}^{k} e_j = 0,$$

which is a desirable property in the seasonality components and is a check on the solution of (5).

To solve (5) the only sum-product computations required are

$$(7) \qquad S_i = \Sigma_t x_{ti} y_t = \Sigma x_{ti} Y_t, \quad i = 1,2,\ldots,p$$

All the other terms in (5) are summations the z_{tj} are zeros or unities. The method of solution involves substitution for the c_j from (5) (ii) into (5) (i), solving for the b_i and substituting back into (5) (ii) to obtain the c_j. The method will be illustrated by application to the data in Table 1.

Application to Raw Data

Table 1. Quarterly Output of Electricity in Five Years 1959-1963

KWHm.

Year	Quarter				Total
	I	II	III	IV	
1959	572	437	417	593	2,019
1960	646	470	464	658	2,238
1961	668	507	491	698	2,364
1962	754	563	538	756	2,611
1963	852	617	578	813	2,860
Sum	3,492 Z_1	2,594 Z_2	2,488 Z_3	3,518 Z_4	12,092

Source: ITJSB.

It is proposed to use the first 2(=p) orthogonal polynomials for trend. Also k = 4, T = 5, kT = 20, so that $\bar{Y}$ = 12,092/20 = 604.6. We also require the

orthogonal polynomial elements shown in Table 2 set out in 5 sets of 4.

Table 2. Orthogonal Polynomial Elements.

Set	x_{t1}				x_{t2}			
1-1959	-19	-17	-15	-13	57	39	23	9
2-1960	-11	- 9	- 7	- 5	- 3	-13	-21	-27
3-1961	- 3	- 1	1	3	-31	-33	-33	-31
4-1962	5	7	9	11	-27	-21	-13	- 3
5-1963	13	15	17	19	9	23	39	57
Sum	-15	- 5	5	15	5	- 5	- 5	5

Source: Statistical Tables by R.A. Fisher and F. Yates, Fifth Ed. page 91 ($n' = 20$)

The two sum-products are

$$\Sigma_t x_{t1} y_t = 16,412; \quad \Sigma_t x_{t2} y_t = 3,036.$$

The equation system (5) then is

$$\text{(i)} \quad \begin{aligned} 16,412 &= 2,660b_1 \overbrace{- 15c_1 - 5c_2 + 5c_3 + 15c_4}^{C_1} \\ 3,036 &= 17,556b_2 + \underbrace{5c_1 - 5c_2 - 5c_3 + 5c_4}_{C_2}, \end{aligned}$$

(The coefficients of b_1 and b_2 are tabled in the source)

$$\text{(ii)} \quad \begin{aligned} 3,492 - 5 \times 604.6 &= 469 = -15b_1 + 5b_2 + 5c_1 \\ 2,594 - \quad " &= -429 = - 5b_1 - 5b_2 + 5c_2 \\ 2,488 - \quad " &= -535 = \quad 5b_1 - 5b_2 + 5c_3 \\ 3,518 - \quad " &= 495 = 15b_1 + 5b_2 + 5c_4. \end{aligned}$$

(Note, by addition of (ii), $\Sigma c_j = 0$)

From (ii), noting definition of C_1 and C_2,

$$\begin{aligned} C_1 &= -28 - 10.0b_1 \\ C_2 &= 1,928 + 20b_2. \end{aligned}$$

Substituting in (i),

$16,412 = 2,660\ b_1 \quad -28-100b_1$ or $b_1 = 16,440/2,560 = 6.421875$

$3,036 = 17,556\ b_2 + 1,928 + 20b_2$ or $b_2 = 1,108/17,576 = 0.063041$

Finally, on substituting these values for b_1 and b_2 in (ii) we have

$5c_1 = 565.012$; $5c_2 = -396.575$; $5c_3 = -566.794$; $5c_4 = 398.357$

with check that the sum is zero. For computational purposes it will be convenient to set down the values of $c_j + \overline{Y}$. These are

$$c_1 + \overline{Y} = 717.6$$
$$c_2 + \overline{Y} = 525.3$$
$$c_3 + \overline{Y} = 491.2$$
$$c_4 + \overline{Y} = 684.3$$

Having determined the coefficients b_1, b_2, c_1, c_2, c_3, c_4, the value of R^2 can be found. This is .9800 which might be regarded as satisfactorily large. But is it the best we can do?

Application to Logarithmic Data

The answer is no. The implication of this arithmetical approach is that the actual production in any quarter in any year is to be found by adding the correction (i.e. the appropriate c_j) to the trend, this correction is the same for all years. This assumption is somewhat unreal in a rapidly increasing phase of expansion. As a rough test from Table 1 it will be observed that the ranges (all between quarters III and IV) in the successive five years were 176, 194, 207, 218, 235, or 8.7, 8.7, 8.8, 8.3, 8.2 of annual production. While there may be some slight tendency towards a declining amplitude there can be no doubt that the hypothesis of seasonality acting proportionately should yield more satisfactory results with this data than to assume additivity.

We accordingly repeat the foregoing calculations but now using log Y_t instead of Y_t. The results are as follows (using primed letters for corresponding symbols used in the foregoing arithmetical case).

$$\overline{Y}' = 2.7726$$

$$b_1' = .0^2 4609375$$

$$b_2' = .0^4 33910$$

$$c_1' + \overline{Y}' = 2.8536$$

$$c_2' + \overline{Y}' = 2.7162$$

$$c_3' + \overline{Y}' = 2.6896$$

$$c_4' + \overline{Y}' = 2.8310$$

The regression yields the satisfactorily high value of .9913 for R^2. There can be little doubt that the logarithmic approach is to be preferred. The results are compared in Table 3. The "calculated" values are those using formula (1) with $u_t = 0$. For the logarithmic regression the antilogs are, of course, shown. These resulted in a slight discrepancy in the five-year output of 12,092 in giving a figure of 12,088: the 4 units were distributed so as to give the correct total.

The actual series (Table 3, Column 2) and the two calculated series (Columns 3 and 4) are graphed on the appended Diagram.

Table 3. Actual and Calculated Values of Output of Electricity, Quarterly 1959-1963

KWHm.

Quarter	Actual	Calculated		Deviation (absolute value) from actual		Difference, cols. 5-6
		(i) Original data	(ii) Logarithmic	(i) Original data	(ii) Logarithmic	
1	2	3	4	5	6	7
1959 I	572	599	587	27	15	+12
II	437	419	436	18	1	+17
III	417	396	418	21	1	+20
IV	593	601	592	8	1	+ 7
1960 I	646	647	636	1	10	- 9
II	470	466	472	4	2	+ 2
III	464	445	454	19	10	+ 9
IV	658	651	641	7	17	-10
1961 I	668	696	690	28	22	+ 6
II	507	517	514	10	7	+ 3
III	491	495	493	4	2	+ 2
IV	698	702	698	4	0	+ 4
1962 I	754	748	751	6	3	+ 3
II	563	569	559	6	4	+ 2
III	538	548	538	10	0	+10
IV	756	755	761	1	5	- 4
1963 I	852	802	820	50	32	+18
II	617	623	611	6	6	+ 0
III	578	603	588	25	10	+15
IV	813	810	833	3	10	-17
Total	12,092	12,092	12,092	258	168	90

Testing for Relative Efficiency

Are the deviations in columns 5 and 6 significantly different? The two entries for each of the twenty units may be regarded as "treatments". The nul-hypothesis is that these are not significantly different. It will be noted from column 7 that in only 4 cases out of 20 are the column 5 deviations less than those in column 6. On the nul-hypothesis the number of - or + signs would have a probability distribution of

$$(\tfrac{1}{2} + \tfrac{1}{2})^{20},$$

when the probability of 4 or fewer - or + signs being found is .0118, so small that the nul-hypothesis is rejected.

Procedure

The seasonality factors c_1, c_2, c_3, c_4 based on the original data or some transform of these (e.g. the logarithm) would be calculated for the latest five calendar years and applied to correct the four quarters of the following year. As soon as the data for the four quarters of the current year become available the same procedure is adopted. The actual data displayed in the example worked out above will be used for different series and different sets of years. Only the following data are specific to each series: Z_1, Z_2, Z_3, Z_4 (see Table 1), $S_1 = \Sigma_t x_{t1} y_t$, $S_2 = \Sigma x_{t2} y_t$. The actual formulae for the c_j could be written down as linear expressions in these symbols and the "constant" symbols. The solution in arithmetical terms is so simple that there is not much point in doing so.

As to the actual procedure as applied to the data for 1964 we have

	log seasonality factors 1959-63	Antilog	Output Actual	Output Corrected
(1)	(2)	(3)	(4)	(5)
$c_1' =$	0.0810	1.2050	885	734
$c_2' =$	T.9436	0.8782	676	770
$c_3' =$	T.9170	0.8260		
$c_4' =$	0.0584	1.1440		

The product of the factors in column (3) is unity. Between the first and second quarter of 1964 seasonally corrected output increased by 36 KWHm. or by 4.9%.

Short term Forecasting

The logarithmic regression fit is so remarkably good that confidence might be reposed in using formula (1) (without u_t) for forecasting the quarterly output in the short term from I 1964. For this the formulae for x_{t1} and x_{t2} are required. They are as follows:-

$$x_{t1} \equiv 2t + 1$$

$$x_{t2} \equiv t^2 + t - 33$$

The serial number for I 1964 is 10. The actual values for these polynomial terms and the z_{tj} elements are shown in Table 4.

Table 4.

t	I 1964 10	II 1964 11	III 1964 12	IV 1964 13	I 1965 14	Coefficients (logarithmic)
x_{t1}	21	23	25	27	29	$0.0^{2}4609375$
x_{t2}	77	99	123	149	177	$-0.0^{4}33910$
z_{t1}	1	0	0	0	1	2.8536*
z_{t2}	0	1	0	0	0	2.7162*
z_{t3}	0	0	1	0	0	2.6896*
z_{t4}	0	0	0	1	0	2.8310*
Calculated						
Log Y_t	2.9478	2.8189	2.8007	2.9504	2.9813	
Y_t	887	659	632	892	958	KWHm.
Actual Y_t	885	676				"

* Including mean log Y = 2.7726

The entries at log Y_t are the sum products of the last column by the other columns. The correspondence is excellent in I 1964, actually better than is to be expected normally, as reference to column 6 of Table 3 will show. The mean deviation for that column is 8.4 and the standard deviation 9.7 so that the deviation of 17 KWHm for II 1964 while somewhat larger than might usually be found is not yet to be regarded as a change in the 1959-1963 trend.

Comparison of Seasonality Correction Divisors

The seasonality correction divisors are computed by three methods, the two methods described above and a third based on the adjusted moving average method. The results are shown in Table 5.

Table 5. Seasonality Correction Divisors Computed by Three Methods, for Quarterly Electricity Output 1959-1963

Quarter	Regression		Moving average
	Original data	Log orig. data	
I	1.208	1.205	1.208
II	0.878	0.878	0.880
III	0.825	0.826	0.82[illegible]
IV	1.143	1.144	1.140

In each case the four divisors have been adjusted so that their product is unity.

The results are almost identical. For the application of the moving average method (last column)

the actual moving average results are corrected for the well-known aberrations at high and low points on the moving average curve, to determine trend, using a method described in Memorandum No. 17 : estimates of divisors for each quarter are the five year geometric means of the quotient of original data by trend.

Concluding Remarks

It is not to be expected that the method described here will yield quite so satisfactory results when used with other major series: for example, in the Third French Plan the percentage increases (to the unit place) forecast were identical with the actual increases in each case in electricity, gas and petroleum production*. Quite a large amount of experimental work has been completed in the Institute on seasonality correction using the moving average method corrected for change in trend - to be described in another ERI Memorandum - and this work shows that

(1) major Irish time data fluctuate seasonally with large amplitude;

(2) at the quarterly level these fluctuations exhibit marked regularity;

(3) accordingly, any good method for seasonality correction, applied at the quarterly level is likely to yield reliable results.

---

*Lloyds Bank Review, June, 1964, p.24

The prospects of being able to appraise the trend of the Irish economy at quarterly intervals are distinctly promising.

August 1964.

DIAGRAM

Actual and Calculated Electricity Output

Quarterly 1959-1962