APPENDIX A

Multiple Regression Model

The following explanation of the multiple regression technique is based on chapter IV 'The General Linear Model' of J. Johnston, Econometric Methods, 1963. The need for and explanation of Durbin Watson test is based on pages 177-179 and 192 of the same book.

In this study the crop yield has been considered as the dependent variable Y while the climatic parameters as independent variables say $X_2, X_3, X_4$ etc. A linear relationship has been assumed to exist between the yield Y and the weather parameters $X_2X_3X_4$ etc. so that if we have data for a set of $n$ years we can write, the yield in the $i$th year

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i,$$

where $u_i$ is a disturbance term.
Thus the $\beta$ coefficients and the parameters of the $u$ distribution are unknown, and our problem is to obtain estimates of these unknowns. The $n$ equations can be set out compactly in matrix notation as

$$
\mathbf{y} = \mathbf{X} \mathbf{\beta} + \mathbf{u}
$$

when

$$
\mathbf{y} = \begin{bmatrix}
    Y_1 \\
    Y_2 \\
    \vdots \\
    Y_n
\end{bmatrix}
$$

$$
\mathbf{x} = \begin{bmatrix}
    1 & X_{21} & X_{31} & \cdots & X_{k1} \\
    1 & X_{22} & X_{32} & \cdots & X_{k2} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    1 & X_{2n} & X_{3n} & \cdots & X_{kn}
\end{bmatrix}
$$

$$
\mathbf{\beta} = \begin{bmatrix}
    \beta_1 \\
    \beta_2 \\
    \vdots \\
    \beta_k
\end{bmatrix}
$$

$$
\mathbf{u} = \begin{bmatrix}
    u_1 \\
    u_2 \\
    \vdots \\
    u_n
\end{bmatrix}
$$

The intercept term $\beta_1$ requires the insertion of a column of units in $\mathbf{X}$ matrix. The convention of using $X_{ki}$ to denote the $i$th observation on the variable $X_k$ means that the subscripts in the $\mathbf{X}$ matrix follow the reverse of the normal pattern, when the first subscript indicates the row and the second the column of the matrix.
In order to estimate the unknown vector of coefficients $\beta$ the following crucial assumptions must be fulfilled.

i) $E\left(\underline{u}_i\right) = 0$

ii) $E\left(\underline{uu}'\right) = \sigma^2 I_n$

iii) $X$ has rank $k < n$.

The first assumption states $E\left(\underline{u}_i\right) = 0$ for all $i$, that is, that the $\underline{u}_i$ are random variables with zero expectation. Assumption (ii) is a compact way of writing a very important double assumption. Since $\underline{u}$ is an $n \times 1$ column vector and $\underline{u}'$ is $1 \times n$ row vector, the product $\underline{u} \underline{u}'$ is a symmetric matrix of order $n$ and since the operation of taking expected values is applied to each element of the matrix, we have

$$E\left(\underline{uu}'\right) = \begin{bmatrix}
E\left(u_1^2\right) & E\left(u_1 u_2\right) & \ldots & E\left(u_1 u_n\right) \\
E\left(u_2 u_1\right) & E_2\left(\underline{u}\right) & \ldots & E\left(u_2 u_n\right) \\
\vdots & \vdots & \ddots & \vdots \\
E\left(u_n u_1\right) & E\left(u_n u_2\right) & \ldots & E\left(u_n^2\right)
\end{bmatrix}$$
\[
\begin{bmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & \sigma^2 \\
\end{bmatrix}
\]

The terms on the main diagonal show that \( E(\ u_i^2) = \sigma^2 \) for all \( i \); that is, the \( u_i \) have constant variance \( \sigma^2 \) which property is referred to as homocedasticity. The off-diagonal term \( E( u_t \ u_t + s ) = 0 \) for \( s \neq 0 \); that is the \( u_i \) values are pairwise uncorrelated.

The final assumption (iii) about the matrix \( X \) is that the number of observations exceed the number of parameters to be estimated and that no exact linear relations exist between any of the \( X \) variables, where it is convenient to extend the list of \( X \) variables to include \( X_1 \), whose value is unity, corresponding to the first column in \( X \). If, for example, our explanatory variable were a multiple of another, or one were an exact linear function of several others, then the rank of \( X \) would be less than \( k \), and likewise the rank of \( X'X \) would be less than \( k \). Since \( X'X \) is a symmetric matrix of order \( k \), this would mean that its inverse did not exist but this inverse \( (X'X)^{-1} \) plays a crucial role in the estimation procedure.
Under the above assumptions to estimate $\beta$ coefficients, we may write

$$
\hat{\beta} = \begin{bmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\hat{\beta}_3 \\
\vdots \\
\hat{\beta}_k
\end{bmatrix}
$$
denote a column vector of estimates of $\beta$.

Then we may write

$$
y = X \hat{\beta} + e
$$
where $e$ denotes the column vector of $n$ residuals $(y - X \hat{\beta})$. In the former equation the unknown disturbances $u$ appear, while in the latter we have a set of residuals $e$. From this second equation the sum of squared residuals is

$$
\sum_{c=1}^{n} e^2 = e' e = (y - X \hat{\beta})' (y - X \hat{\beta})
$$

$$
= y'y - 2 \hat{\beta}' X' y + \hat{\beta}' XX \hat{\beta}
$$

which follow from noting that $\hat{\beta}' X'y$ is a scalar and thus equal to its transpose $y'X \hat{\beta}$. To find the value of $\hat{\beta}$ which minimizes the sum of the squares of the
residuals, we differentiate the above equation

\[ \frac{\delta}{\delta \beta} (e' e) = -2 X'Y + 2 X'X \hat{\beta} \]

Equating to zero gives

\[ X'X \hat{\beta} = X'Y \]

i.e.

\[ \hat{\beta} = (X'X)^{-1} X'Y \]

This is the fundamental result for the least square estimates giving \( \hat{\beta} \) as a column vector.

The coefficient of multiple correlation

\( R_{1.23 \ldots k}^2 \) is defined by

\[ R_{1.23 \ldots k}^2 = \frac{\hat{\beta}' X'Y - \frac{1}{n} (\sum Y)^2}{Y'Y - \frac{1}{n} (\sum Y)^2} \]

To test the hypothesis that \( \beta_i = 0 \), that is \( X_i \) has no linear influence on \( Y \), we compute the test statistic

\[ t = \frac{\hat{\beta}_i}{\sqrt{\sum e_i^2/(n - k) \sqrt{a_{ii}}}} \]
which has the $t$ distribution with $n-k$ degrees of freedom, where $a_{ii}$ is the $i^{th}$ diagonal element in $(X'X)^{-1}$. To test the significance of the regression (or in other words, the multiple correlation coefficient) the value of the statistic $F$

$$F = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)}$$

was also provided by the computer output.

One of the crucial assumptions of the linear model is the serial independence of the disturbance term implied in $E\left( uu'\right) = \sigma^2 I$. where $I$ is identity matrix of order $n$ which gives

$$E\left( u_t u_t' + s \right) = 0$$

for all $t$ and all $s \neq 0$.

If there is any serial correlation in the $X$ values, then we will have serial correlation in the composite disturbance term. We include only certain important variables in a specified relation, and the disturbance term must then represent the influence of omitted variables. If the serial correlation in the omitted variables is pervasive
and if the omitted variables tend to move in phase, then there is a real possibility of an auto-correlated disturbance term. A disturbance term may also contain a component due to measurement error in the explained variable, here the yield. This too may be a source of serial correlation in the composite disturbance.

When there is such serial correlation, first, the sampling variances of the unbiased estimates of the regression coefficients may be unduly large and secondly the formulas for the sampling variances of the regression coefficients are likely to yield only serious under estimate of these variances. In any case these formulas are no longer valid, nor are the precise forms of the $t$ and $F$ tests derived for the linear model. Third, we shall obtain inefficient predictions, that is, predictions with needlessly large sample variances.

It is very important, therefore, to be able to test for the presence of autocorrelated disturbances for ensuring the validity of the linear model and a suitable test is available in the Durbin - Watson d statistic

Let $z_t \ (t = 1, 2 \ldots n)$ denote the residuals from a fitted-least square regression. We then define
\[
    d = \frac{\sum_{t=2}^{n} (z_t - z_{t-1})^2}{\sqrt{\sum_{t=1}^{n} z_t^2}}
\]

Exact significant levels for \(d\) are not available, but Durbin and Watson have tabulated lower and upper bounds \(d_L\) and \(d_U\) for the various values of \(n\) and \(k\)
(\(k = \) number of explanatory variables).

To conduct a one-sided test of positive auto-correlation, compute \(d\).

If \(d < d_L\) reject the hypothesis of random disturbances in favour of positive auto-correlation.

If \(d > d_U\) do not reject the hypothesis.

If \(d_L < d < d_U\) the test is inconclusive.

**Thiel - Nagar test:**

Thiel and Nagar (1961, *Journal of American statistical Association*, Vol. 56, pp. 793-806) presented a test for null hypothesis an improvement over Durbin-Watson test for the 'region of ignorance' that is the interval \((d_L, d_U)\). The test has been applied wherever necessary and the significance of the parameters tested at 1 or 5 percent level.
Von Neumann Ratio test:

Let $x_1 \ldots x_n$ be variables representing $n$ successive observations in a population with a normal distribution. The mean square successive difference

$$\delta^2 = \frac{1}{n-1} \sum_{\mu=1}^{n-1} (x_{\mu+1} - x_\mu)^2$$

The Von Neumann's ratio is

$$\eta = \frac{\delta^2}{s^2} \text{ where } s^2 \text{ is variance.}$$

A comparison of the observed value of $\eta$ with the distribution of $\eta$ is used as a basis of the judgement whether the observations $x_1 \ldots x_n$ are independent or whether a trend exists. (The Annals of Mathematical Statistics, XII, 1941, 367). If the calculated value exceeds the tabulated value, it may be assumed that there is no trend in the successive observations.
Appendix B

Computer Programming Used in the Study

The analysis of enormous data for developing suitable models was done with the help of IBM-1620 computer in the Indian Institute of Tropical Meteorology, Poona and the fast electronic computer at the Tata Institute of Fundamental Research at Bombay.

This involved designing of suitable computer programmes, recasting of the data in the laid down formats, its punching, verification and transfer on tape etc. A review of the computational programmes is presented here.

As a first step, correlation and multiple correlation coefficients between crop yield and each of the climatic and water balance parameters for the months of the growing season were obtained, and are presented in Table I.

The significant parameters obtained in Table I are filtered out for subjecting them to further computer dropping programme in which a total of 94 com-
binations are obtained dropping one or more variable at a time. The manner in which dropping takes place is shown in Table II. The programme accomodates nine independent and one dependent variable (here labeled $X_1$ instead of $Y$ of appendix A). The first independent parameter $X_2$ in the set of nine ($X_2 \ldots X_{10}$) is never dropped.

These 94 combinations are examined and the best one wherein the multiple correlation coefficient is high and also most of the partial coefficients are significant is selected in the first stage.

The final programme is designed to drop only one variable at a time. The combinations obtained by this dropping programme are examined and the best one which satisfies all statistical tests is selected as the final one for the model. The manner in which dropping is done is presented in Table III.

Computer aid was also taken for calculating water balance parameters according to Thornthwaite's formulae 1948,
### TABLE 1

**TOTAL CORRELATION COEFFICIENTS BETWEEN YIELD $X_1$ AND CLIMATIC PARAMETERS**

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>$T$</th>
<th>D. F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between $X_1$ and $X_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between $X_1$ and $X_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between $X_1$ \ldots $X_{10}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PARTIAL REGRESSION COEFFICIENTS**

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>$T$</th>
<th>D. F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between $X_1$ and $X_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between $X_1$ and $X_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between $X_1$ \ldots $X_{10}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where $X_1$ = Yield and $X_2$, $X_3$ \ldots $X_{10}$ are the months corresponding to growing season.

Multiple correlation coefficient = 

$F' = \ldots$
TABLE II

DROPPING PROGRAMME (FIRST STAGE).

TOTAL 94 COMBINATIONS ARE OBTAINED.

(1) First combination (Serial No.1) contains yield ($X_1$) and nine climatic parameters ($X_2$ to $X_{10}$) as dependent and independent variables respectively.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Variables Dropped</th>
<th>Sr. No.</th>
<th>Variables Dropped</th>
<th>Sr. No.</th>
<th>Variables Dropped</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$X_3$</td>
<td>3</td>
<td>$X_4$</td>
<td>4</td>
<td>$X_5$</td>
</tr>
<tr>
<td>5</td>
<td>$X_6$</td>
<td>6</td>
<td>$X_7$</td>
<td>7</td>
<td>$X_8$</td>
</tr>
<tr>
<td>8</td>
<td>$X_9$</td>
<td>9</td>
<td>$X_{10}$</td>
<td>10</td>
<td>$X_3$ $X_4$</td>
</tr>
<tr>
<td>11</td>
<td>$X_3$ $X_5$</td>
<td>12</td>
<td>$X_3$ $X_6$</td>
<td>13</td>
<td>$X_3$ $X_7$</td>
</tr>
<tr>
<td>14</td>
<td>$X_3$ $X_8$</td>
<td>15</td>
<td>$X_3$ $X_9$</td>
<td>16</td>
<td>$X_3$ $X_{10}$</td>
</tr>
<tr>
<td>17</td>
<td>$X_4$ $X_5$</td>
<td>18</td>
<td>$X_4$ $X_6$</td>
<td>19</td>
<td>$X_4$ $X_7$</td>
</tr>
</tbody>
</table>
20. $x_4 x_8$
21. $x_4 x_9$
22. $x_4 x_{10}$
23. $x_5 x_6$
24. $x_5 x_7$
25. $x_5 x_8$
26. $x_5 x_9$
27. $x_5 x_{10}$
28. $x_6 x_7$
29. $x_6 x_8$
30. $x_6 x_9$
31. $x_6 x_{10}$
32. $x_7 x_8$
33. $x_7 x_9$
34. $x_7 x_{10}$
35. $x_8 x_9$
36. $x_8 x_{10}$
37. $x_9 x_{10}$
38. $x_3 x_4 x_5$
39. $x_3 x_4 x_6$
40. $x_3 x_4 x_7$
41. $x_3 x_4 x_8$
42. $x_3 x_4 x_9$
43. $x_3 x_4 x_{10}$
44. $x_3 x_5 x_6$
45. $x_3 x_5 x_7$
46. $x_3 x_5 x_8$
47. $x_3 x_5 x_9$
48. $x_3 x_5 x_{10}$
49. $x_3 x_6 x_7$
50. $x_3 x_6 x_8$
51. $x_3 x_6 x_9$
52. $x_3 x_6 x_{10}$
53. $x_3 x_7 x_8$
54. $x_3 x_7 x_9$
55. $x_3 x_7 x_{10}$
56. $x_3 x_8 x_9$
57. $x_3 x_8 x_{10}$
58. $x_3 x_9 x_{10}$
59. $x_4 x_5 x_6$
60. $x_4 x_5 x_7$
61. $x_4 x_5 x_8$
62. $x_4 x_5 x_9$
63. $x_4 x_5 x_{10}$
64. $x_4 x_6 x_7$
65. \( x_4 x_6 x_8 \)  
66. \( x_4 x_6 x_9 \)  
67. \( x_4 x_6 x_{10} \)  
68. \( x_4 x_7 x_8 \)  
69. \( x_4 x_7 x_9 \)  
70. \( x_4 x_7 x_{10} \)  
71. \( x_4 x_8 x_9 \)  
72. \( x_4 x_8 x_{10} \)  
73. \( x_4 x_9 x_{10} \)  
74. \( x_5 x_6 x_7 \)  
75. \( x_5 x_6 x_8 \)  
76. \( x_5 x_6 x_9 \)  
77. \( x_5 x_6 x_{10} \)  
78. \( x_5 x_7 x_8 \)  
79. \( x_5 x_7 x_9 \)  
80. \( x_5 x_7 x_{10} \)  
81. \( x_5 x_8 x_9 \)  
82. \( x_5 x_8 x_{10} \)  
83. \( x_5 x_9 x_{10} \)  
84. \( x_6 x_7 x_8 \)  
85. \( x_6 x_7 x_9 \)  
86. \( x_6 x_7 x_{10} \)  
87. \( x_6 x_8 x_9 \)  
88. \( x_6 x_8 x_{10} \)  
89. \( x_6 x_9 x_{10} \)  
90. \( x_7 x_8 x_9 \)  
91. \( x_7 x_8 x_{10} \)  
92. \( x_7 x_9 x_{10} \)  
93. \( x_8 x_9 x_{10} \)  
94. \( x_3 x_4 x_5 x_6 \)
TABLE III

DROPPING PROGRAMME OF ONE VARIABLE
AT A TIME (FINAL STAGE)

\[ X_1 = \text{Yield and } X_2 \ldots \ldots X_{10} = \text{Climatic parameters} \]

Arranged as follows before dropping

\[
\begin{array}{cccccccccc}
X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \\
\hline
X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & - - \\
\hline
X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & - - & - - \\
\hline
X_1 & X_2 & - - & - - & - - & - - & - - & - - & - - \\
\end{array}
\]

(1) Each combination contains 'F' value and 't' value of simple and partial coefficients.

(2) Multiple correlation coefficient.

(3) Durbin value and Neumann's ratio.

The best combination which is statistically valid is selected.