Table 13. - North Carolina cotton yields 1927-1941.

Year	Acreage harvested	Reported Condition August 1	Reported weevil infestation August 1	Yield Per acre
	thousand acres	percent	percent	pounds
1927	1,565	78	19.1	262
1928	1,620	73	25.5	245
1929	1,635	68	30.3	217
1930	1,448	74	25.4	254
1931	1,206	78	16.9	298
1932	1,251	65	21.0	252
1933	1,072	79	13.6	305
1934	970	77	13.2	311
1935	930	77	17.7	294
1936	957	60	9.0	298
1937	1,103	85	15.8	338
1938	857	68	30.0	216
1939	737	83	25.1	296
1940	829	84	6.5	427
1941	795	74	20.5	333

The data in table 13 can be used to derive a regression equation for forecasting the cotton yield from information available to the statistician on August 1. The relationship between yield and reported condition will be considered first. Final yield is plotted against reported August condition in figure 19. This relationship is based on data taken in different years whereas the proceding regressions were based on data taken at the same time, but there is no difference in fundamental concepts.

A regression equation could be derived from these data by the method described previously, but with so few observations it is better to use a more accurate method. The most accurate method that can be used is known as the method of least squares. This method leads to a regression equation such that the sum of the squares of the deviations of the observed cotton yields from the corresponding computed yields will have the smallest possible numerical value. The equation can be fitted by this method more easily if it is written in a slightly different form than the one given previously. It can be written,

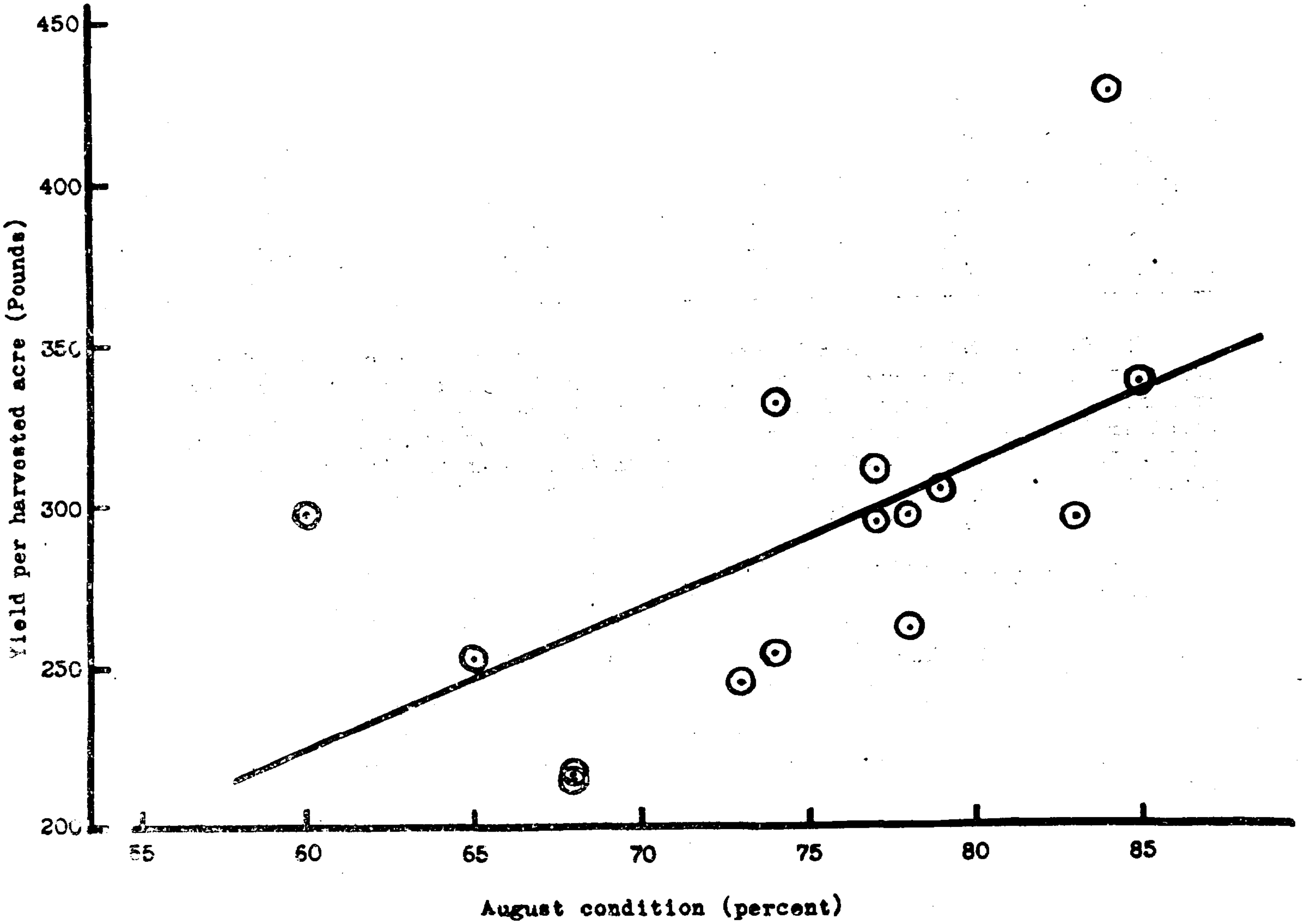
$$Y = \bar{y} + b(X - \bar{x}) \qquad \text{- - - - - - - - - - -} \qquad (75)$$

In this equation the constant b has the same meaning as before, but the averages $\bar{x}$ and $\bar{y}$ take the place of the constant a that appeared in equation (69). Equation (75) is equivalent to equation (69). It can be reduced to that form by letting

$$a = \bar{y} - b\bar{x} \qquad \text{- - - - - - - - - - -} \qquad (76)$$

and leaving the value of b unchanged.

Figure 19. Relation between North Carolina cotton yields and August condition 1927-1941



In the present example $\bar{y}$ is the average cotton yield for the 15-year period and $\bar{x}$ is the average reported August condition.

$$\bar{x} = 74.867$$

$$\bar{y} = 289.73$$

This leaves only the constant b to be evaluated. It is given by the equation,

$$b = \frac{S\left[(X - \bar{x})(Y - \bar{y})\right]}{S\left[(X - \bar{x})^2\right]} \qquad \text{- - - - - - - - -} \qquad (77)$$

in which X is the reported August condition for any year and Y is the observed cotton yield for the same year. If x represents the deviation of the reported condition for any year from the 15-year average and y represents the deviation of the observed yield for the same year from the 15-year average, equation (77) can be written in the form

$$b = \frac{S(xy)}{S(x^2)} \qquad \text{- - - - - - - - - - - - - - -} \qquad (78)$$

The numerical values of $S(xy)$ and $S(x^2)$ can be computed most conveniently from the relations,

$$S(xy) = S(XY) - \frac{[S(X)]\,[S(Y)]}{n} \qquad \text{- - - - - - - -} \qquad (79)$$

$$S(x^2) = S(X^2) - \frac{[S(X)]^2}{n} \qquad \text{- - - - - - - - - - -} \qquad (80)$$

in which n represents the number of years. For the data at hand,

$$S(xy) = 328553.0 - 325370.5 = 3182.5$$

$$S(x^2) = 84791.00 - 84075.27 = 715.73$$

$$b = 3182.5/715.73 = 4.4465$$

The complete regression equation is thus,

$$Y = 289.73 + 4.4465(X - 74.867) \qquad \text{- - - - - - - -} \qquad (81)$$

or

$$Y = -\,43.17 + 4.4465X \qquad \text{- - - - - - - - - - - -} \qquad (82)$$

The values of Y computed from equation (81) or equation (82) will fall on the straight line shown in figure 19. These equations enable one to forecast the final cotton yield from the reported August condition.

An inspection of figure 19 indicates that the simple regression equation just derived will not forecase cotton yields very accurately. The observed yields fluctuate over a wide range about the regression line. To measure the accuracy with which the equation predicts cotton yields, some additional computations would be necessary. The sum of squares of the deviations of the observed yields from the 15-year average is given by the equation,

$$S(y^2) = S(Y^2) - \frac{\left[S(Y)\right]^2}{n} \qquad \text{- - - - - - - - - -} \quad (83)$$

For the data at hand,

$$S(y^2) = 1299302 - 1259181 = 40121$$

This quantity represents the sum of the squares of the deviations of the observed yields in figure 19 from a horizontal line drawn at the level of the 15-year average yield. This line would represent a forecase of cotton yield for any year under the assumption that there was no relation between yield and August condition. Any reduction of this sum of squares would be contributed by the constant b in the regression equation. Giving b a value different from zero involves nothing more than tilting the regression line, using the point defined by $\bar{x}$ and $\bar{y}$ as a pivot. The sum of squares contributed to the total by the slope of the regression line is $\left[S(xy)\right]^2/S(x^2) = (3182.5)^2/715.73 = 14151.$ The residual sum of squares left by this quantity is 40,121 - 14,151 = 25,970. This means that, of a total sum of squares equal to 40,121, an amount equal to 14,151 was accounted for by the constant b. The remainder, 25,970, represents the residual sum of squares of the deviations of the observed cotton yields from the computed values given by the regression equation. These results can be summarized by an analysis of variance, as indicated in table 14.

Table 14. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of Squares	Mean square
Regression on August condition	1	14,151	14,151
Error	13	25,970	1,998
Total	14	40,121	2,866

In this table the total degrees of freedom is one less than the number of observed points plotted in figure 19. The mean square 2,866, computed from these 14 degrees of freedom measures the scatter of the observed points in the chart about a horizontal line drawn at the level of the average cotton yield for the 15-year period. The 14 degrees of freedom are broken down into 1 degree of freedom corresponding to the part of the total variability in yield that is associated with August condition, and 13 degrees of freedom corresponding to the residual variability that was not accounted for by the August condition. The success with which the regression equation predicts the cotton yields is indicated by the mean square 1,998. The improvement brought about by using the regression equation to predict yields, instead of using only the 15-year average, can be measured by comparing the residual or error mean square 1,998 with the total mean square 2,866.

Statisticians sometimes compute a number called the <u>correlation coefficient</u> to measure the degree of relationship between two variables like observed cotton yield and reported condition. This quantity is usually computed from the equation,

$$r = \frac{S(xy)}{\sqrt{\left[S(x^2)\right]\left[S(y^2)\right]}} \qquad \text{- - - - - - - - -} \quad (84)$$

This quantity can be computed from an analysis of variance like that in table 14 very easily when such a table is available. The coefficient of correlation defined by equation (84) is equal to the square root of the quotient obtained when one divides the sum of squares for regression by the total sum of squares. For the data in table 14,

$$r = \sqrt{\frac{14151}{40121}} = \sqrt{0.3527} = 0.5939$$

The square of the correlation coefficient thus measures the relative amount which the regression on condition contributes to the total sum of squares.

A discussion of the correlation coefficient and its many properties will not be attempted here. The subject is covered so thoroughly in almost every textbook on statistics that the reader will have no difficulty in pursuing the subject if he desires to do so. For the present it is sufficient to show how the correlation coefficient can be computed from an analysis of variance like that in table 14.

But in problems of this kind it is unnecessary to compute the correlation coefficient, as the analysis of variance itself gives more information than could be gotten from the correlation coefficient. Anyone who is interested in computing a correlation coefficient, however, can easily obtain it from the analysis of variance. As a matter of fact, there is some objection to using the correlation coefficient as a measure of the success with which cotton yields can be predicted from August condition. The numerical value of the correlation coefficient is computed from two sums of squares. This is not a fair comparison because no allowance is made for the number of degrees of freedom entering into each sum of squares. A more legitimate comparison can be made between the error mean square and the total mean square in table 14. When the error mean square is much smaller than the total, it can be concluded that the cotton yields are being predicted accurately. When the error mean square is almost as large as the total mean square, however, the yields are not being predicted accurately from the condition figures.

Sometimes interest may be in learning whether the regression coefficient, b, is significantly different from zero. Especially when working with a fairly small number of observations, a good relationship between two variables is sometimes nothing but an accident. For that reason it is usually a good idea to test a relationship like the one given above to see how often it would occur by chance if there were no actual relationship between reported August condition and final yield. At one time it was common practice to compute the standard error of the regression coefficient, b, and to compare the numerical value of b with its standard error. An analysis of variance like the one in table 14 makes this procedure unnecessary. The significance of the regression can be tested by computing the F-ratio of the regression mean square to the error mean square. For the results in table 14, F = 14151/1998 = 7.08. Reference to a table of F values shows that this value of F is significantly greater than 1.00. Thus it may be concluded that the observed regression of cotton yield on reported August condition is not an accident.

Exercise 28.-A State contains 150,000 farms and 12,500,000 acres of farm land. The relationship between acreage of potatoes and size of farm is given by the equation

$$Y = 5.42 + .0234X$$

A sample of 1,000 farms, averaging 200 acres in size, shows an average of 10.10 acres of potatoes. Estimate the potato acreage for the State:
(a) From the number of farms and per farm average.
(b) From the farm land and farm-land average.
(c) From the regression equation.
Explain the differences between these three estimates.

Exercise 29.-A sample of 1,200 farms is divided into two equal parts on the basis of size. 600 farms are above 150 acres while the other 600 are less than 150 acres. The average size and average wheat acreage for each group is:

	Average land in farm (acres)	Average wheat acreage (acres)
Large farms	250	75
Small farms	50	25

Compute the linear regression equation that gives the relationship between size of farm and wheat acreage. If the average farm in the State contains 125 acres of farm land, what would you expect the average wheat acreage to be? What is the average size of farm in the entire sample of 1,200 farms?

Exercise 30.-Suppose the 1,200 farms in Exercise 29 were from a State containing 100,000 farms and 15,000,000 acres of farm land. Estimate the wheat acreage for the State by the three methods used in Exercise 28 and explain the results.

Multiple Regression and Multiple Correlation

Methods for studying the relationship between two variables were described in the preceding section. The relationship between North Carolina cotton yield and reported August condition was worked out to show how August condition can be used to forecast final yield. As indicated in figure 19, the August condition alone is not an accurate indicator of the final yield. The observed yields shown in the chart often differ widely from the forecasts given by the regression line. The actual yields are compared with the forecasts in table 15.

Table 15. - North Carolina cotton yields compared with yields estimated from August condition.

Year	Yield per acre		Error
	Observed	Estimated	
	pounds	pounds	pounds
1927	262	304	– 42
1928	245	281	– 36
1929	217	259	– 42
1930	254	286	– 32
1931	298	304	– 6
1932	252	246	+ 6
1933	305	308	– 3
1934	311	299	+ 12
1935	294	299	– 5
1936	298	224	+ 74
1937	338	335	+ 3
1938	216	259	– 43
1939	296	326	– 30
1940	427	330	+ 97
1941	333	286	+ 47

The reader would naturally be interested in learning how much a forecase of final yield could be improved by making use of additional information like the figures on weevil infestation and harvested acreage in table 13. Reported weevil infestation seems to vary from year to year and the data in table 13 indicate that the final yield is correlated with this factor. Furthermore, the acreage harvested declined considerably during the 15-year period. There is reason to believe that the reduction in acreage had some effect on the relationship between yield and reported condition.

Additional factors of this kind can be included in a forecasting equation. Such an equation is called a multiple regression equation because the final yield is predicted from more than one factor. The effect of reported weevil infestation will be considered first. A multiple regression equation for forecasting yield from reported August condition and reported weevil infestation may be written in the form,

$$Y = a_0 + a_1X_1 + a_2X_2 \quad - - - - - - - - - \quad (85)$$

or

$$Y = \bar{y} + a_1(X_1 - \bar{x}_1) + a_2(X_2 - \bar{x}_2) \quad - - - - - - - \quad (86)$$

Equation (85) can be derived from equation (86) by letting

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 \quad - - - - - - - - - \quad (87)$$

and leaving a_1 and a_2 unchanged. X_1 represents the reported August condition and X_2 represents reported weevil infestation. Representing $(X_1 - \bar{x}_1)$ by

x_1, $(X_2 - \overline{x}_2)$ by x_2, and $(Y - \overline{y})$ by y, equation (86) can be written in the form

$$y = a_1x_1 + a_2x_2 \qquad - - - - - - - \qquad (88)$$

The net regression coefficients a_1 and a_2 are the only quantities that need to be evaluated. These can be obtained by solving the <u>normal equations,</u>

$$\begin{aligned} a_1S(x_1^{2}) + a_2S(x_1x_2) &= S(x_1y) \\ a_1S(x_1x_2) + a_2S(x_2^{2}) &= S(x_2y) \qquad - - - - - - \qquad (89) \end{aligned}$$

The quantities entering into these equations can be computed most conveniently from relations similar to those used previously:

$$\begin{aligned} S(x_1^{2}) &= S(X_1^{2}) - \frac{\left[S(X_1)\right]^2}{n} \\ S(x_1x_2) &= S(X_1X_2) - \frac{\left[S(X_1)\right]\left[S(X_2)\right]}{n} \\ S(x_2^{2}) &= S(X_2^{2}) - \frac{\left[S(X_2)\right]^2}{n} \\ S(x_1y) &= S(X_1Y) - \frac{\left[S(X_1)\right]\left[S(Y)\right]}{n} \\ S(X_2y) &= S(X_2Y) - \frac{\left[S(X_2)\right]\left[S(Y)\right]}{n} \qquad - - - - - - \qquad (90) \end{aligned}$$

For the data at hand, the normal equations are,

$$\begin{aligned} + 715.73a_1 - 183.89a_2 &= + 3182.5 \\ - 183.89a_1 + 709.35a_2 &= -.4309.2 \qquad - - - - - \qquad (91) \end{aligned}$$

The numerical values of a_1 and a_2 obtained by solving these simultaneous equations are,

$$\begin{aligned} a_1 &= + 3.0916 \\ a_2 &= - 5.2734 \end{aligned}$$

The complete regression equation, in the form indicated by equation (86), may thus be written,

$$Y = 289.73 + 3.0916(X_1 - 74.867) - 5.2734(X_2 - 19.307) \qquad - - - - - \qquad (92)$$

This equation can easily be written in the form indicated by equation (85),

$$Y = 160.08 + 3.0916X_1 - 5.2734X_2 \qquad - - - - - - - - - - - - - - - - - - \qquad (93)$$

Equation (93) can be used to forecast North Carolina cotton yields from reported August condition and reported weevil infestation. The coefficient of X_1 is positive, indicating that yield increases as reported condition increases. The coefficient of X_2 is negative, indicating that yield decreases as weevil infestation increases. The forecast of final yield is the net result of these opposing influences. The success with which equation (93) forecasts final yield is shown in table 16.

Table 16.- North Carolina cotton yields compared with yields estimated from August condition and weevil infestation.

Year	Yield per acre		Error
	Observed	Estimated	
	pounds	pounds	pounds
1927	262	300	- 38
1928	245	251	- 6
1929	217	211	+ 6
1930	254	255	- 1
1931	298	312	- 14
1932	252	250	+ 2
1933	305	333	- 28
1934	311	329	- 18
1935	294	305	- 11
1936	298	298	0
1937	338	340	- 2
1938	216	212	+ 4
1939	296	284	+ 12
1940	427	385	+ 42
1941	333	281	+ 52

A comparison of tables 15 and 16 shows that equation (93) gives better results than equation (82). The errors in the estimates are much smaller when weevil infestation is considered along with reported August condition in making a forecast of final yield. This conclusion is strengthened by constructing an analysis of variance table, similar to table 14, for the data in table 16. The total sum of squares is equal to 40,121 as before. The sum of squares for regression is $a_1S(x_1y) + a_2S(x_2y) = (+ 3.0916)(+ 3182.5)+(- 5.2734)(- 4309.2) = 9839 + 22724 = 32563$. The residual error sum of squares is $40121 - 32563 = 7558$. As there are two regression coefficients in the forecasting equation, there are 2 degrees of freedom for regression and 12 degrees of freedom for estimating residual error. The complete analysis is summarized in table 17.

Table 17. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of squares	Mean square
Regression on August condition and weevil infestation	2	32,563	16,282
Error	12	7,558	630
Total	14	40,121	2,866

A comparison of tables 14 and 17 shows that allowance for weevil infestation improved the yield forecasts materially. The error mean square was reduced from 1998 to 630. Thus it is possible to compute a multiple correlation coefficient from the data in table 17.

$$R = \sqrt{\frac{32563}{40121}} = \sqrt{0.8116} = 0.9009$$

The multiple correlation coefficient is much greater than the simple correlation coefficient, $r = 0.5939$, obtained when reported August condition was used alone to forecast yields. The multiple correlation coefficient given above is interpreted in the same way as the simple correlation coefficient computed previously. R^2 represents the fraction of the total sum of squares that is associated with the regression of yield on reported August condition and reported weevil infestation. But as stated previously, the correlation coefficient makes no allowance for the degrees of freedom entering into the sums of squares from which it was computed. The values of R and r are not strictly comparable because one additional degree of freedom was taken out of the sum of squares for error and transferred to the regression sum of squares. This in itself would make R greater than r, even if reported weevil infestation had no significant effect in improving the forecasts. This at once raises the question of whether R is significantly larger than r.

The significance of the improvement in the forecasts can be tested more conveniently by an analysis of variance than by comparing the two correlation coefficients. When reported August condition was used by itself to forecast yields, the error sum of squares was 25,970 as shown in table 14. When weevil infestation was included as a second factor in the forecasting equation, the error sum of squares was 7,558 as shown in table 17. The reduction in the error sum of squares, 25,970 - 7,558 = 18,412, was effected by the weevil infestation factor. The regression sum of squares 32,563 in table 17 can thus be broken down into two components. The first represents the sum of squares contributed by the simple regression of yield on reported August condition. This is 14,151 as shown in table 14. The second component represents the additional sum of squares contributed by the second factor, weevil infestation. This is equal to 18,412 as indicated above. The analysis of variance in table 17 can thus be presented in more detais, as shown in table 18.

Table 18. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of squares	Mean square
Regression on August condition	1	14,151	14,151
Regression on weevil infestation	1	18,412	18,412
Error	12	7,558	630
Total	14	40,121	2,866

The improvement in the forecasts, brought about by using weevil infestation as a second factor in the forecasting equation, can be tested by computing F = 18412/630 = 29.2. This value is much larger than unity and reference to a table of F values shows that it is highly significant. Reported weevil infestation is thus demonstrated to have a highly significant effect on yield. A forecast of yield is improved tremendously when this factor is included in the forecasting equation along with reported August condition.

When working with an analysis of variance like the one in table 18, the reader should be careful to interpret the table correctly. The sum of squares ascribed to weevil infestation represents the reduction in the error sum of squares brought about by this factor after the simple regression of yield on reported August condition has exerted its effect. The student should not be misled into thinking that the two sums of squares for regression represent the independent net effects of August condition and weevil infestation. At one time statisticians were interested in measuring the net effect of each factor in a multiple regression equation. Various formulas were devised for this purpose, but they have not proved to be completely satisfactory. They are rigorously correct only in certain special cases. In general, there is no theoretically sound method by which the net effects of the various factors in a multiple regression equation can be measured. The procedure described above is theoretically sound, but it does not represent an attempt to measure the net effects of condition and weevil infestation. The effect of condition is measured without regard to weevil infestation. After this effect has been measured, the improvement brought about by including weevil infestation as an additional factor is measured. This differs from an attempt to measure net effects of the two factors, and the difference in viewpoint should be noted carefully.

The importance of this distinction can be demonstrated most forcefully by first measuring the effect of weevil infestation alone and then measuring the improvement brought about by including August condition in the forecasting equation. When this is done, the analysis of variance given in table 19 is obtained. A forecasting equation based only on weevil infestation accounts for 26,178 of the total sum of squares. The improvement brought about by adding the August condition is only 6,385. The analyses given in tables 18 and 19 are both correct, but they must be interpreted correctly to avoid confusion. It may be observed that reported weevil infestation, considered alone, is a better indicator of final yield than reported August condition when that variable is used alone.

Table 19. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of squares	Mean square
Regression on weevil infestation	1	26,178	26,178
Regression on August condition	1	6,385	6,385
Error	12	7,558	630
Total	14	40,121	2,866

This method of analysis may be extended to include harvested acreage. The final forecasting equation will then include three variables. The equation may be written in any of the following forms, using the same notation as before:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 \qquad ---------------- (94)$$

$$Y = \bar{y} + a_1(X_1 - \bar{x}_1) + a_2(X_2 - \bar{x}_2) + a_3(X_3 - \bar{x}_3) \qquad -------- (95)$$

$$y = a_1x_1 + a_2x_2 + a_3x_3 \qquad -------------------- (96)$$

in these equations X_1 represents reported August condition, X_2 represents reported weevil infestation, and X_3 represents harvested acreage. Small letters denote deviations of the variables from their respective arithmetic means as before,

The normal equations for evaluating a_1, a_2, and a_3 are

$$\begin{aligned} a_1S(x_1^2) + a_2S(x_1x_2) + a_3S(x_1x_3) &= S(x_1y) \\ a_1S(x_1x_2) + a_2S(x_2^2) + a_3S(x_2x_3) &= S(x_2y) \\ a_1S(x_1x_3) + a_2S(x_2x_3) + a_3S(x_3^2) &= S(x_3y) \qquad --------- (97) \end{aligned}$$

The quantities entering into these equations may be computed by methods described previously. For the data at hand,

$$\begin{aligned} +\ 715.73a_1 - 183.89a_2 - 6741a_3 &= +\ 3182.5 \\ -\ 183.89a_1 + 709.35a_2 - 12222a_3 &= -\ 4309.2 \\ -\ 6741\ a_1 - 12222a_2 + 1337300a_3 &= -\ 132030 \qquad ------- (98) \end{aligned}$$

Solving these equations, one obtains

$$a_1 = +\ 2.8660$$

$$a_2 = -\ 4.6048$$

$$a_3 = -\ 0.042197$$

The regression equation may then be written,

$$Y = 289.73 + 2.8660(X_1-74.867) - 4,6048(X_2- 19.307) - 0.042197(X_3- 1131.7) \quad ---- (99)$$

or

$$Y = 211.81 + 2.8660X_1 - 4.6048X_2 - 0.042197X_3 \quad ------------(100)$$

The observed yields are compared with the yields estimated from this equation in table 20.

Table 20. - North Carolina cotton yields compared with yields estimated from August condition, weevil infestation, and harvested acreage.

Year	Yield per acre		Error
	Observed	Computed	
	pounds	pounds	pounds
1927	262	281	- 19
1928	245	235	+ 10
1929	217	198	+ 19
1930	254	246	+ 8
1931	298	307	- 9
1932	252	249	+ 3
1933	305	330	- 25
1934	311	331	- 20
1935	294	312	- 18
1936	298	302	- 4
1937	338	336	+ 2
1938	216	232	- 16
1939	296	303	- 7
1940	427	388	+ 39
1941	333	296	+ 37

These results may be summarized in an analysis of variance as before. The sum of squares associated with the regression is,

$$a_1S(x_1y) + a_2S(x_2y) + a_3S(x_3y) =$$

$$(+ 2.8660)(+ 3182.5) + (- 4.6048)(- 4309.2) + (- 0.042197)(-132030) = 34535.$$

The residual error sum of squares is 40121 - 34535 = 5586. Table 21 shows the analysis of variance.

Table 21. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of squares	Mean square
Regression on August condition, weevil infestation, and acreage	3	34,535	11,512
Error	11	5,586	508
Total	14	40,121	2,866

The multiple correlation coefficient is,

$$R= \sqrt{\frac{34535}{40121}} = \sqrt{0.8608} = 0.9278$$

The improvement in the forecasts brought about by using harvested acreage as a third variable in the regression equation can be tested by computing the reduction in the error sum of squares. August condition and weevil infestation left an error sum of squares equal to 7,558 as shown in tables 17, 18, and 19. The additional reduction due to the acreage is 7558 - 5586 = 1972.

Maintaining the same order of analysis as in table 18, the 3 degrees of freedom for regression shown in table 21 can be broken down to obtain the more detailed analysis given in table 22.

Table 22. - Analysis of variance of North Carolina cotton yields, 1927-1941.

Source of variability	Degrees of freedom	Sum of squares	Mean square
Regression on August condition	1	14,151	14,151
Regression on weevil infestation	1	18,412	18,412
Regression on harvested acreage	1	1,972	1,972
Error	11	5,586	508
Total	14	40,121	2,866

The significance of the acreage effect may be tested by computing F = 1972/508 = 3.88. Reference to a table of F values shows that this ratio is not significantly greater than unity. The apparent improvement in the yield forecasts brought about by including harvested acreage as a third variable in the regression equation is thus not great enough to conclude that acreage exerts a real effect on yield. But to decide this question more definitely it would be desirable to work with a longer series of observations. If acreage really does exert an effect, and it might logically be expected to do so, a longer series of observations would lead to a significant value of F.

The discussion of multiple regression presented in this section covers methods of analysis that have been developed in recent years. For several reasons, no attempt has been made to review the more familiar procedures that may be found in any textbook. Many of the older methods seem decidedly cumbersome when compared with the techniques described here. Furthermore, they do not contribute any additional information essential to a statistical analysis. Such topics as partial correlation, standard errors of net regression coefficients, coefficients of determination, and similar subjects add little to what can be learned from a set of data by applying the methods of analysis given above. If the reader wishes to pursue the subject, he will find it instructive to study systematic methods for solving normal equations, such as the so-called Doolittle method. The standard error of a forecast made from a regression equation could also be studied to good advantage. These topics are covered so well in many excellent available textbooks that it is not necessary to duplicate the material here.

When working with forecasting equations, the reader should bear in mind that such equations will give most accurate results at the point defined by the arithmetic means of the observed data. The accuracy of a forecast diminishes at an increasing rate as one gets farther away from the means of the observed data. Forecasts much beyond the range of the observed data used in deriving the forecasting equation are usually subject to such high standard errors that the forecasts are of little practical value. The accuracy of a forecast also depends upon the size of the error mean square such as the values given in the preceding analyses of variance. The forecasts are more accurate as the error mean square is decreased.

Joint Regression equations

The joint effects of two or more variables are often important in developing a forecasting equation. Methods for measuring such effects and making proper allowance for them are described in some textbooks, but these methods do not seem to be used as widely as they might be. The application of these methods to a few problems in forecasting crop yields is described in this section. The examples given are sufficient to illustrate the procedure. Many other forecasting equations can be worked out by the same method.

Consider the ordinary multiple regression of North Carolina cotton yield on reported August condition and harvested acreage. The necessary data are given in table 13. If X_1 represents reported August condition and X_3 represents harvested acreage, the regression equation is of the same form as equation (85) in the preceding section. Working with deviations from the various arithmetic means, as before, the normal equations are

$$\begin{aligned} +\ 715.73a_1 - 6741a_2 &= 3182.5 \\ -\ 6741\ a_1 + 1337300a_2 &= -\ 132030 \qquad \text{- - - - - - - -} \quad (101) \end{aligned}$$

Solving these equations gives

$$a_1 = +\ 3.6919$$

$$a_2 = -\ 0.080119$$

The final regression equation is

$$Y = 289.73 + 3.6919(X_1-74.867) - 0.080119(X_3-1131.7) \qquad \text{- - - -} \quad (102)$$

or

$$Y = 104.00 + 3.6919X_1 - 0.080119X_3 \qquad \text{- - - - - - - - - - - - - -} \quad (103)$$

Equation (103) enables the forecasting of yield from reported August condition and harvested acreage. When the acreage is given, equation (103) can be reduced to a form that involves only the reported August condition as the indicator of final yield. By assigning different values to X_3, a family of regression lines for forcasting yield from condition is obtained. To make a forecast for any year, one is in effect choosing the particular regression line specified by the acreage. If the harvested acreage is 1,600, for example, equation (103) may be written,

$$Y = 104.00 + 3.6919X_1 - (0.080119)\ (1600)$$

or

$$Y = -24.19 + 3.6919X_1 \text{ - - - - - - - - - - - - - - -} (104)$$

For an acreage of 1200, equation (103) reduces to the form,

$$Y = +\ 7.86 + 3.6919X_1 \text{ - - - - - - - - - - - - - - -} (105)$$

For an acreage of 800, equation (103) becomes,

$$Y = +\ 39.91 + 3.6919X_1 \text{ - - - - - - - - - - - - - - -} (106)$$

For an acreage of 400, equation (103) becomes,

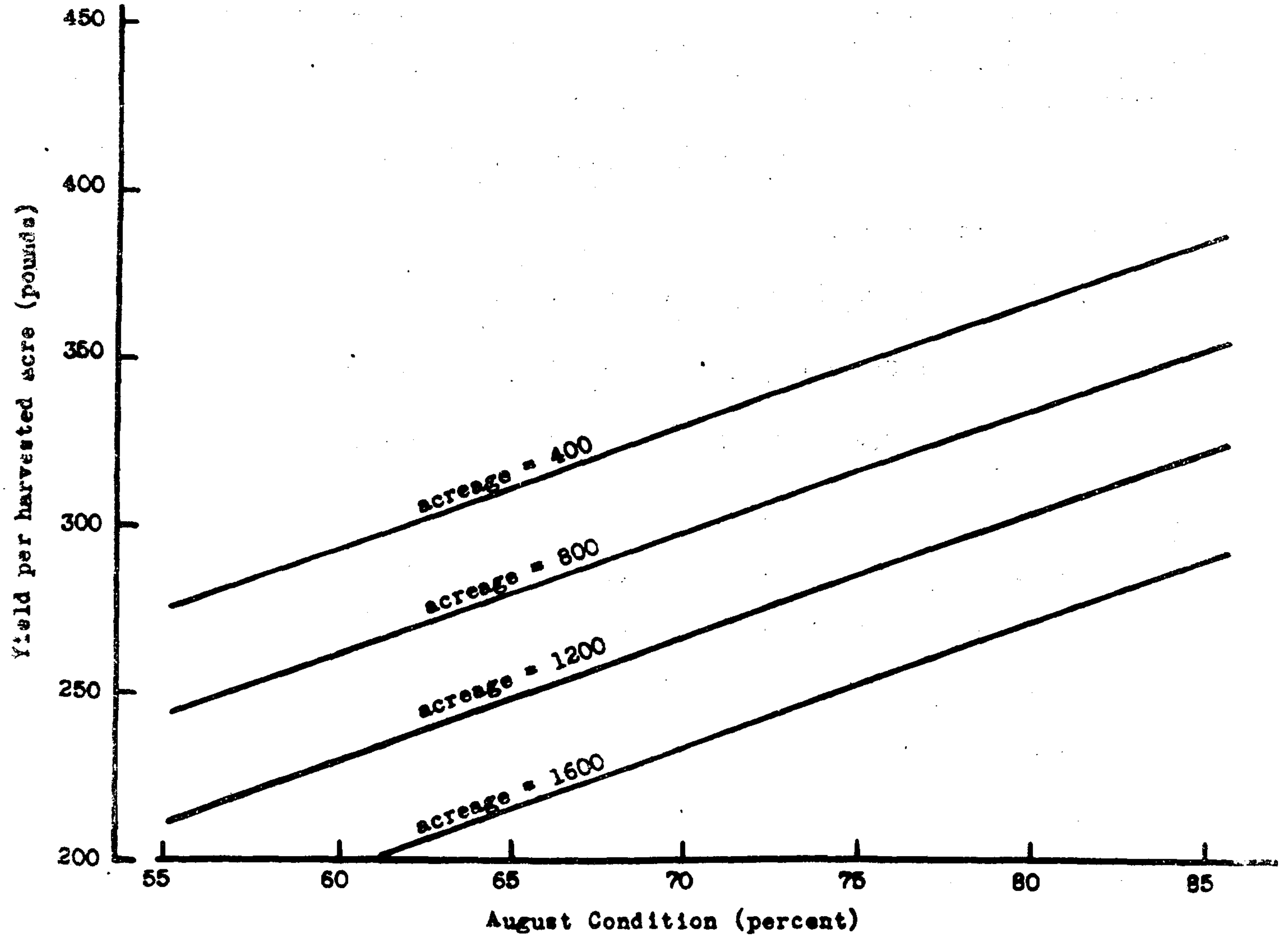
$$Y = +\ 71.95 + 3.6919X_1 \text{ - - - - - - - - - - - - - - -} (107)$$

The regression lines corresponding to equations (104) , (105), (106), and (107) are shown in figure 20.

Figure 20 illustrates a fundamental feature of the ordinary multiple regression analysis. The net regression of final yield on reported August condition is a line of constant slope, regardless of the acreage harvested. This line is raised or lowered as the harvested acreage is decreased or increased. It seems logical that the line should be raised as the acreage is decreased in this case. Most of the reduction in acreage during the 15-year period was effected by Government crop-control programs. When a farmer's cotton acreage is restricted to a smaller allotment than would ordinarily be harvested, it seems reasonable to suppose that the poorer land would be the first to be taken out of production. This should result in a higher yield per unit of reported condition because the condition figure is not a measure of probable yield in itself. It is supposed to represent a percentage of a normal crop and can be used as a measure of probable yield only when the normal yield for every locality is specified. The normal yield of the better land remains fairly constant and a reduction in acreage could easily result in a higher yield for the State without a corresponding increase in the reported condition.

The ordinary multiple regression equation makes no allowance for a possible change in the slope of the line representing the relation between yield and condition as the acreage changes. If a higher yield per unit of reported condition were a consequence of reduced acreage under the crop-control program, the slope of the line might be expected to change also. Instead of being parallel, the lines shown in figure 20 should be steeper when the acreage is low than when the acreage is high. The necessary flexibility that permits the net regression lines to have this property can be introduced into a multiple regression equation by writing it in the form,

Figure 20. Forecasts of North Carolina cotton yield from reported August condition for different harvested acreages. (Ordinary multiple regression.)



$$Y = a_0 + a_1X_1 + a_2X_3 + a_3X_1X_3 \quad ------ \quad (108)$$

X_1 represents reported August condition and X_3 represents harvested acreage as before. The third variable X_1X_3 is the product of August condition and harvested acreage. When this equation is fitted to the data, the product X_1X_3 is used as a third independent variable. To avoid confusion, the equation can be written in the form,

$$Y = a_0 + a_1W_1 + a_2W_2 + a_3W_3 \quad ------ \quad (109)$$

in which $W_1 = X_1$, $W_2 = X_3$, and $W_3 = X_1X_3$. In this form, the equation is just an ordinary multiple regression equation with three independent variables that can be fitted to the data by methods described in the preceding section. Using small letters to represent deviations from arithmetic means as before, equation (109) may be written in the form,

$$y = a_1w_1 + a_2w_2 + a_3w_3 \quad -------- \quad (110)$$

The normal equations are,

$$\begin{aligned} a_1S(w_1^{\,2}) + a_2S(w_1w_2) + a_3S(w_1w_3) &= S(w_1y) \\ a_1S(w_1w_2) + a_2S(w_2^{\,2}) + a_3S(w_2w_3) &= S(w_2y) \\ a_1S(w_1w_3) + a_2S(w_2w_3) + a_3S(w_3^{\,2}) &= S(w_3y) \quad --- \quad (111) \end{aligned}$$

For the data in table 13, these equations are,

$$\begin{aligned} +\ 715.73a_1 - 6741a_2 + 239500a_3 &= +\ 3182.5 \\ -\ 6741a_1 + 1337300a_2 + 91940000a_3 &= -\ 132030 \\ +\ 239500a_1 + 91940000a_2 + 7081000000a_3 &= -\ 6515000 \quad --- \quad (112) \end{aligned}$$

Solving these equations, one obtains

$$a_1 = +\ 4.6543$$

$$a_2 = -\ 0.011094$$

$$a_3 = -\ 0.00093343$$

The final regression equation is

$$Y = 289.73 + 4.6543(W_1-74.867) - 0.011094(W_2-1131.7) - 0.00093343(W_3-84290) \qquad ----- (113)$$

or

$$Y = 32.52 + 4.6543W_1 - 0.011094W_2 - 0.00093343W_3 \qquad ----- (114)$$

Changing back to the original variables, this equation may be written in the form,

$$Y = 32.52 + 4.6543X_1 - 0.011094X_2 - 0.00093343X_1X_3 \qquad ----- (115)$$

When X_3 is successively given the values 1600,1200,800, and 400, as in equation (103), one obtains the four net regression equations,

$$Y = + 14.77 + 3.1608X_1 \qquad ------------- (116)$$

$$Y = + 19.21 + 3.5342X_1 \qquad ------------- (117)$$

$$Y = + 23.64 + 3.9076X_1 \qquad ------------- (118)$$

$$Y = + 28.08 + 4.2809X_1 \qquad ------------- (119)$$

These equations have the properties that they would be expected to have. The slope increases with a reduction in harvested acreage. The regression lines corresponding to the equations are shown in figure 21.

Although the slopes of these lines differ, the differences are not great enough to produce much improvement in the accuracy of the forecasts. The lines tend to spread out in the form of a fan as the acreage changes, but the general picture does not deviate much from that shown in figure 20.,

Now consider the effects of weevil infestation. As in the case of acreage, one could expect a different net relationship between final yield and reported August condition for every different degree of weevil infestation. This relationship could also be modified by the total acreage. A joint regression equation that can make allowance for all of these interactions may be written in the form,

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_1X_2 + a_5X_1X_3 + a_6X_2X_3 + a_7X_1X_2X_3 \qquad ----- (120)$$

This equation contains terms for all possible joint effects of reported August condition, reported weevil infestation, and harvested acreage. When fitted to the data in table 13 by methods described previously, one obtains

Figure 21. Forecasts of North Carolina cotton yield from reported August condition for different harvested acreages. (Joint multiple regression.)

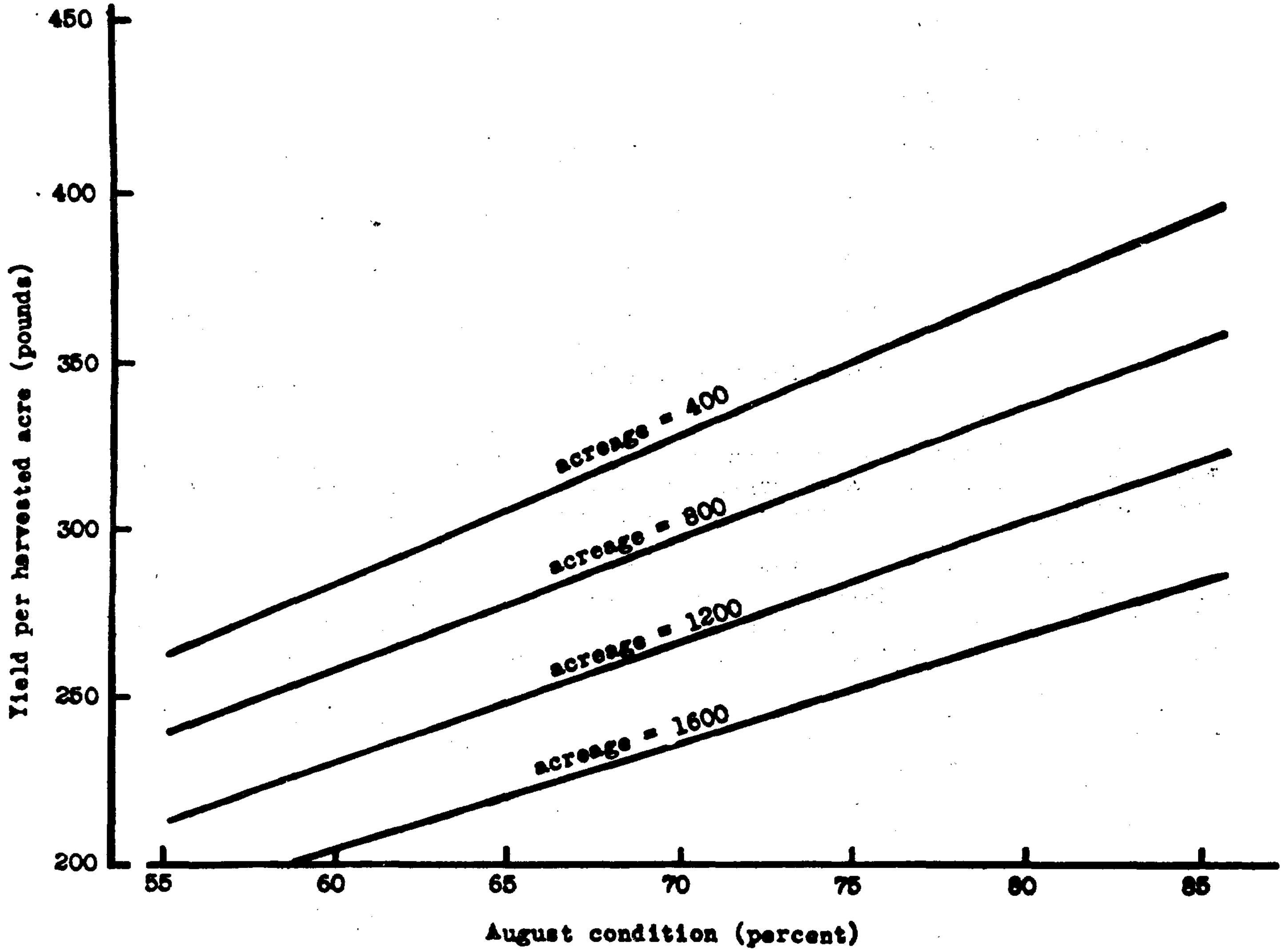
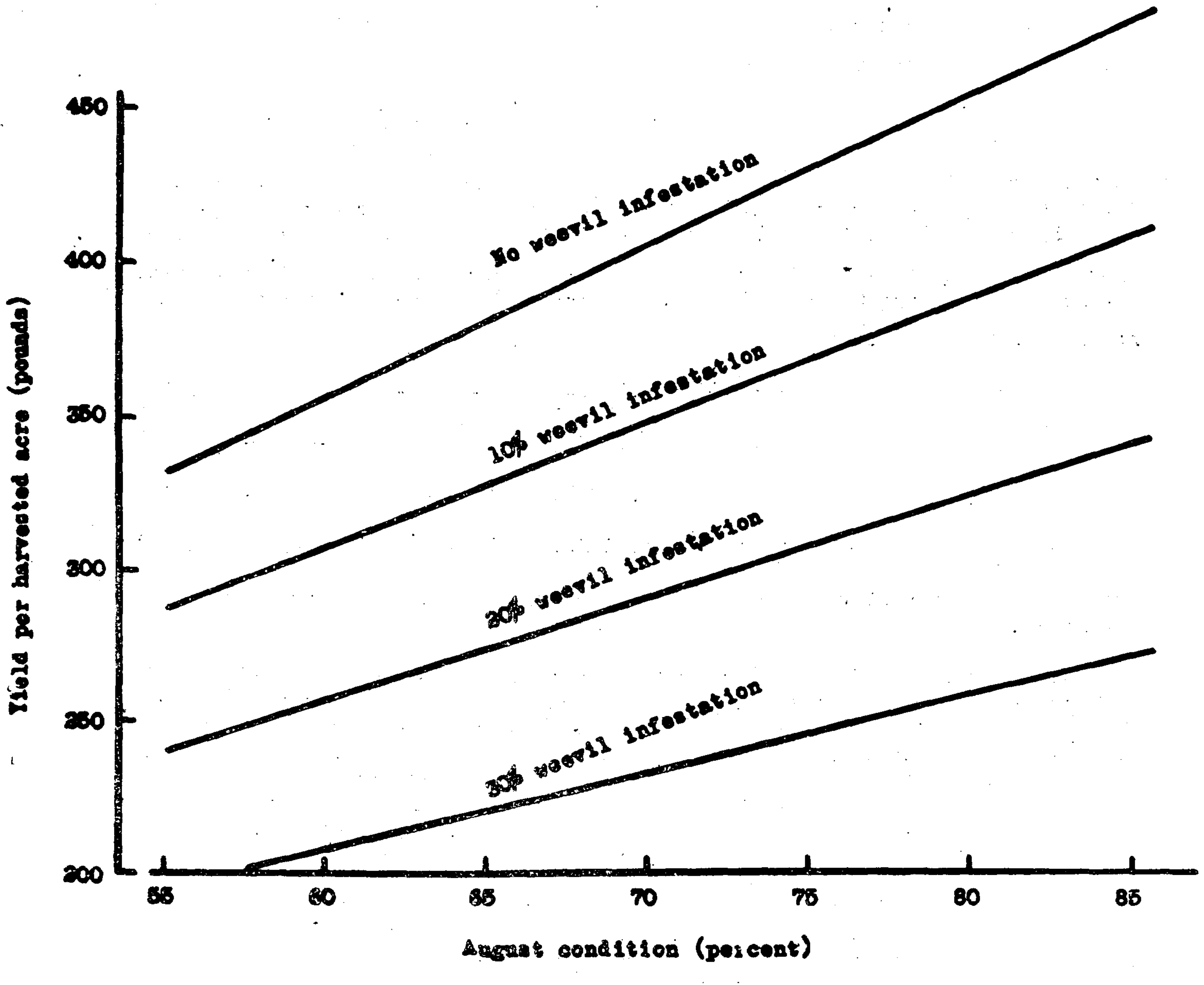


Figure 22. Forecasts of North Carolina cotton yield from reported August condition for different degrees of weevil infestation when harvested acreage = 800. (Joint multiple regression.)



$$\begin{array}{llll} a_0 = -\ 331.34 & & a_4 = -\ 0.461300 \\ a_1 = +\ 12.8708 & & a_5 = -\ 0.01009476 \\ a_2 = +\ 21.0550 & & a_6 = -\ 0.0268086 \\ a_3 = +\ 0.499854 & & a_7 = +\ 0.000481324 \end{array}$$

As the harvested cotton acreage for North Carolina has been fairly stable at about 800 in recent years, the properties of the regression equation given above will be considered only for the special case when $X_3 = 800$. For that value of X_3, the equation reduces to

$$Y = +\ 68.54 + 4.7950X_1 - 0.3919X_2 - 0.076241X_1X_2 \quad - - - - \ (121)$$

This is the joint multiple regression of cotton yield on reported August condition and reported weevil infestation when the harvested acreage is equal to 800. By assigning different values to X_2 in the equation, a family of equations for forecasting final yield from reported August condition is obtained. Giving X_2 the values 0, 10, 20, and 30 in succession, leads to the following net regression equations:

$$Y = 68.54 + 4.7950X_1 \quad - - - - - - - - - - - - - \ (122)$$

$$Y = 64.62 + 4.0326X_1 \quad - - - - - - - - - - - - - \ (123)$$

$$Y = 60.70 + 3.2702X_1 \quad - - - - - - - - - - - - - \ (124)$$

$$Y = 56.78 + 2.5078X_1 \quad - - - - - - - - - - - - - \ (125)$$

The net regression lines corresponding to these equations are shown in figure 22.

The regression lines in figure 22 show more marked changes in slope with differences in weevil infestation than the changes effected by acreage differences. But these changes in slope do not appear large enough to improve the yield forecasts very much. The results of the entire joint multiple regression analysis on the North Carolina cotton-yield data indicate that equation (100), developed in the preceding section, will give just as satisfactory results as an equation that allows for joint effects. The joint effects in this problem do not seem to be worth taking into consideration.

In many multiple regression studies, joint effects are important. The possibility of the existence of such effects should be investigated more frequently than some statisticians consider necessary. A striking example of such effects was found in an attempt to predict yields of corn in Ohio from temperature and rainfall during June, July, and August. For any year in which June and August weather conditions are about average, the relationship between corn yield in Ohio and July temperature and rainfall is given by the equation.

$Y = 78.69 - 0.6700X_1 - 4.6908X_2 + 0.0981X_1X_2$ - - - - - - - - - (126)

In this equation, Y is the corn yield for the State (bushels per acre) , X_1 is the average temperature in July (degrees Fahrenheit), and X_2 is the July rainfall (inches). By letting X_2 take the values 0, 2, 4, 6, 8, and 10 in succession, one obtains the following equations for forecasting Ohio corn yields from July temperature:

$Y = 78.69 - 0.6700X_1$ - - - - - - - - - - - - - - (127)

$Y = 69.31 - 0.4738X_1$ - - - - - - - - - - - - - - (128)

$Y = 59.93 - 0.2776X_1$ - - - - - - - - - - - - - - (129)

$Y = 50.54 - 0.0814X_1$ - - - - - - - - - - - - - - (130)

$Y = 41.16 + 0.1148X_1$ - - - - - - - - - - - - - - (131)

$Y = 31.78 + 0.3110X_1$ - - - - - - - - - - - - - - (132)

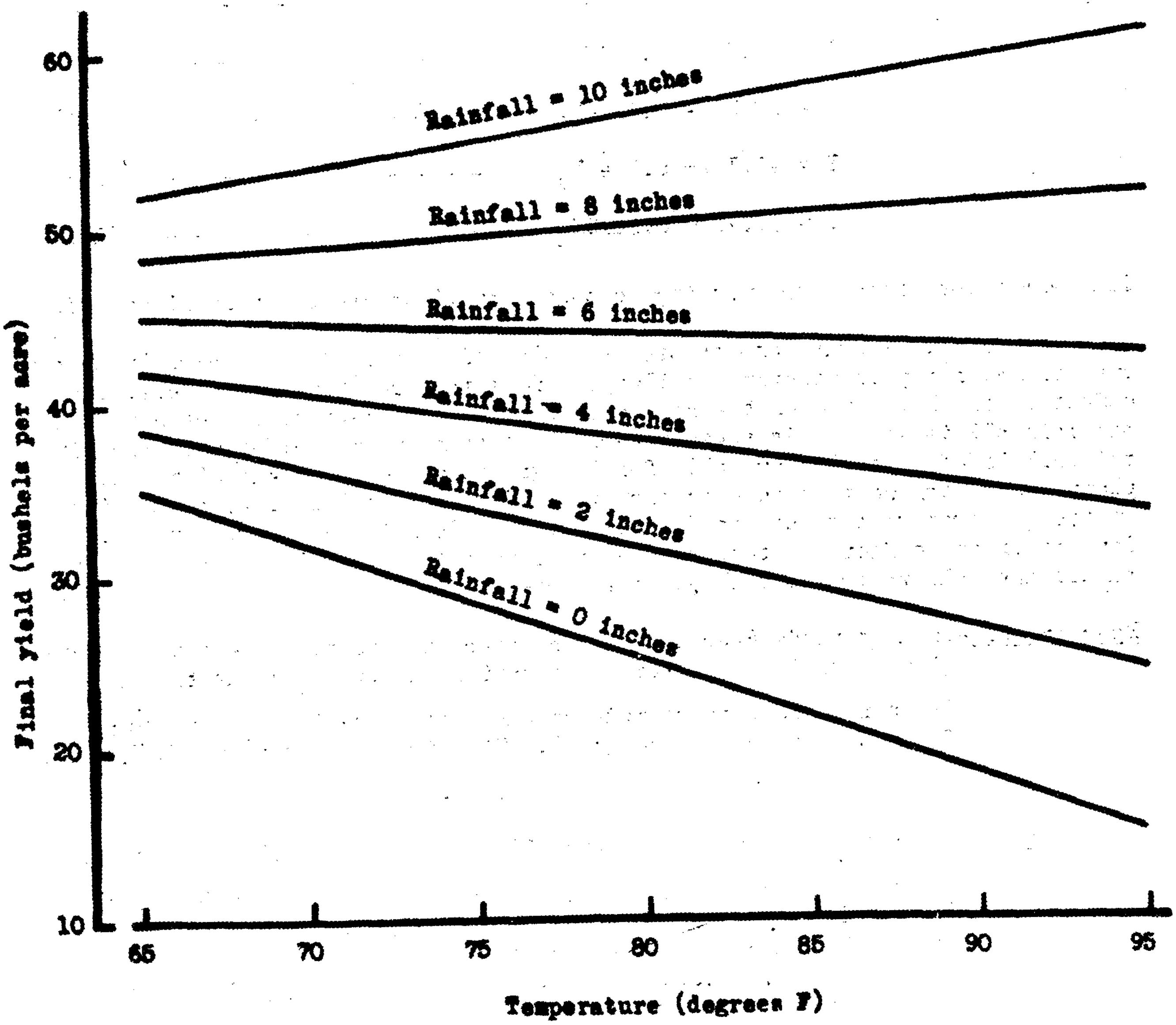
The regression lines corresponding to these equations are shown in figure 23.

The effect of July temperature on yields of corn in Ohio apparently depends upon the quantity of moisture available. High temperatures have a beneficial effect on final yield when the rainfall is high and a detrimental effect when the rainfall is low. This seems reasonable because warm weather with plentiful rainfall is known to be favorable for the growth and development of the corn plant. Hot, dry weather injures the crop. When rainfall is low, a lowering of the temperature compensates for the moisture deficiency to some extent. Studies of the effects of temperature and rainfall on the yields of all crops should make allowance for such joint effects.

<u>Exercise 31.</u>-Using methods described in the preceding section, test the significance of the X_1X_3 term in equation (115). To apply the test it is necessary to compute an analysis of variance separately for equations (103) and (115). The significance of the X_1X_3 term can be determined from those results.

<u>Exercise 32.</u>-By letting X_3 take the value 1200 in equation (121), compute the equations corresponding to equations (122), (123), (124), and (125). Draw the graphs of these equations in a chart like figure 22 and compare the results with the regression lines in figure 22. How would you account for the differences in the two sets of regression lines?

Figure 23. Forecasts of Ohio corn yield from July temperature for different amounts of July rainfall. (Joint multiple regression.)


Rainfall = 10 inches
Rainfall = 8 inches
Rainfall = 6 inches
Rainfall = 4 inches
Rainfall = 2 inches
Rainfall = 0 inches
Final yield (bushels per acre)
Temperature (degrees F)