Reprinted from JOURNAL OF FARM ECONOMICS, Vol. XVIII, No. 3, August, 1936

COVARIANCE USED TO ANALYZE THE RELATION BETWEEN CORN YIELD AND ACREAGE[1]

GERTRUDE M. COX AND GEORGE W. SNEDECOR
STATISTICAL LABORATORY
IOWA STATE COLLEGE

The relation between corn yield and the number of acres in corn on Iowa farms is a matter of considerable interest. If technical and economic conditions were uniform, the presumption would be that corn acreage would vary with the productive capacity of the soil. The very existence of different type-of-farming areas in the state, however, indicates non-uniform conditions. The questions arise as to whether any relation of yield to corn acreage on individual farms exists within these type-of-farming areas, and if so, whether the relation is the same in one area as in another. Recent findings indicating that there is greater erosion on small farms lend additional interest to the inquiry.[2]

This problem was first attacked by Schultz[3] using the technique of analysis of variance. At the time, however, he was handicapped by the fact that no method was available for handling disproportionate subclass numbers. He was compelled to make a random choice of only part of the available data in order to keep his subclass numbers proportional. Later, Snedecor and Cox[4] examined a similar set of data, using the newly developed techniques appropriate to disproportionate subclass numbers. The solution was still inconclusive because in dealing with disproportionate subclass numbers there must be set up some hypothesis about the population from which the sample is drawn. Usually, the exact form of the hypothesis is unessential, but in this case the result obtained seemed to be a consequence of the hypothesis chosen.

Meanwhile, the analysis of covariance as originally presented by Fisher has been expanded and perfected by him and his coworkers.[5] In its extended form, it lends itself admirably to the solution of the yield-acreage problem. In addition, further sur-

[1] The authors wish to express their appreciation of the counsel and assistance rendered by Professor T. W. Schultz, head of the Economics and Sociology Department.

[2] Iowa State Planning Board. The second report. Part I. Land. April 1935.

[3] Schultz, T. W. Testing the mean values drawn from stratified samples. JOURNAL OF FARM ECONOMICS, Vol. XV, No. 3, 452-475 (1933).

[4] Snedecor, George W., and Cox, Gertrude M. Disproportionate subclass numbers in tables of multiple classification. Iowa Agr. Exp. Station. Res. Bul. 180 (1934).

[5] Fisher, R. A. Statistical Methods for research workers. Oliver and Boyd, Edinburgh. Fourth and Fifth Edition (1932, 1934).

Fisher, R. A. The design of experiments. Oliver and Boyd, Edinburgh (1935).

Bartlett, M. S. The problem in statistics of testing several variances. Proc. of the Cambridge Philosophical Soc., XXX, 2:164 (1934).

Welch, B. L. Some problems in the analysis of regression among *k* samples of two variables. Biometrika 27:145 (1935).

veys of Iowa conditions have resulted in a division of the state into more homogeneous type-of-farming areas, seven in number, which are based upon selected physical and economic criteria, such as the proportion of the farm's income obtained from the sales of livestock, livestock products and crops, and upon the typical organization of the farms. It seems appropriate, therefore, to present the new solution of the problem both for its economic contribution and as an illustration of the use of the powerful new group of statistical methods known as analysis of covariance.

The data for 1933 were made available through the courtesy

TABLE 1. NUMBER OF FARMS, CORN ACREAGE AND CORN YIELD IN SEVEN TYPE-OF-FARMING AREAS IN IOWA, 1933

Type-of-farming area	Number of farms	Mean corn acreage	Mean corn yield in bu. per acre
1. Dairy	372	42.8	43.1
2. Northern cash grain	159	75.0	45.4
3. Western meat	403	85.8	41.2
4. North Central cash grain	421	78.6	48.3
5. Eastern meat	396	57.9	47.4
6. South Central pasture	176	59.0	42.5
7. Southern pasture	278	43.3	34.6
Total	2205	63.9	43.5

of L. M. Carl, Agricultural Statistician of the Division of Crop and Livestock Estimates, Bureau of Agricultural Economics, U. S. D. A. A summary for the 2205 farms surveyed is presented in table 1. The original data for each farm were punched in a card, and were tabulated mechanically.[6] This accounts for the large size of some of the numbers appearing in later tables. When using machine tabulation, it is usually easier to carry the large numbers than it is to code the original entries. Looking at the means given in table 1, it is clear that the farms in the several areas are characterized by markedly different corn acreages as well as by widely varying yields. It is not evident that acreage and yield are related in any way. The Western Meat area has the largest corn acreage per farm, but less than average yield. The Southern pasture area, however, is low in both average acreage and yield. This table furnishes no evidence about the relationships among the farms within the type-of-farming areas, nor about the variability of the data for the individual farms. For the isolation of such information, the techniques of analysis of variance and covariance were used. The data for the analyses which

[6] Brandt, A. E. Uses of the progressive digit method. In Baehn, G. W. ed. Practical applications of the punched card method in colleges and universities. Columbia Univ. Press, New York, N.Y. p. 423 (1935).

follow are given in table 2. In this table are given the number of farms in each type-of-farming area, the sums, sums of squares and sums of products of the original entries. These sub-totals were secured from the machine tabulations. The totals for the 2205

TABLE 2. NUMBER OF FARMS, SUMS, SUMS OF SQUARES AND PRODUCTS OF THE ORIGINAL ENTRIES FOR CORN ACREAGE AND CORN YIELD

Area	Number of farms	Acreage or Yield	Sum	Sum of squares of entries	Sum of products of entries
1. Dairy	372	Acreage Yield	15,911 16,027	957,169 738,849	695,449
2. Northern cash grain	159	Acreage Yield	11,931 7,213	1,209,333 346,397	524,768
3. Western meat	403	Acreage Yield	34,561 16,602	4,145,423 737,280	1,422,889
4. North Central cash grain	421	Acreage Yield	33,111 20,325	3,660,707 1,024,781	1,602,352
5. Eastern meat	396	Acreage Yield	22,946 18,756	1,847,742 958,520	1,090,666
6. South Central pasture	176	Acreage Yield	10,376 7,484	947,938 343,926	455,172
7. Southern pasture	278	Acreage Yield	12,032 9,606	744,744 379,858	420,339
Total	2205	Acreage Yield	140,868 96,013	13,513,056 4,529,611	6,211,635

farms, derived by summing the seven sub-totals, are recorded at the bottom of the table.

The analysis of the variance of corn yields on the farms, carried out in the same manner as described in this Journal by Schultz,[7] is recorded in table 3. The large value,

$$F = 6{,}781.8/140.2 = 48.4$$

leaves no doubt of the significance of the differences between the mean area yields in table 1. The questions to be answered are

TABLE 3. ANALYSIS OF VARIANCE OF CORN YIELD IN SEVEN TYPE-OF-FARMING AREAS IN IOWA, 1933

Source of variation	Degrees of freedom	Sum of squares	Mean square
Total	2,204	348,887	
Between area means	6	40,691	6,781.8**
Within areas	2,198	308,196	140.2

** Highly significant.

to what extent and in what manner these yields are related to the acreage devoted to corn on the several farms. To answer these questions, analysis of covariance was used. The computations are fundamentally of the same kind as those in the analysis

[7] Schultz, T. W., *loc. cit.*

of variance, the difference being the introduction of the sum of products associated with the concepts of correlation and regression.

TABLE 4. ANALYSIS OF COVARIANCE OF CORN YIELD AND CORN ACREAGE ON 2205 IOWA FARMS, 1933

Source of variation	Degrees of freedom	Sum of squares and products			Correlation coefficients	Regression coefficients
		Corn acreage Sx^2	Corn Yield Sy^2	Products Sxy		
Total	2,204	4,513,603	348,887	77,776	0.0620	0.0172
Between area means	6	606,472	40,691	58,968	0.3754	0.0972
Within areas	2,198	3,907,131	308,196	18,808	0.0171	0.0048

The sums of squares and products in table 4 were computed from the original data given in table 2 in the manner described below and also as given by Snedecor[8] (examples 2 and 10).

The correction terms:

$$\text{for acreage, } \frac{(140{,}868)^2}{2205} = 8{,}999{,}453$$

$$\text{for yield, } \frac{(96{,}013)^2}{2205} = 4{,}180{,}724$$

$$\text{for products, } \frac{(140{,}868)\ (96{,}013)}{2205} = 6{,}133{,}859$$

The last correction term has a numerator which is the product of two sums instead of the square of a sum.

The *total* sums of squares and products:

for acreage, $13{,}513{,}056 - 8{,}999{,}453 = 4{,}513{,}603$
for yield, $4{,}529{,}611 - 4{,}180{,}724 = 348{,}887$
for product, $6{,}211{,}635 - 6{,}133{,}859 = 77{,}776$

Between area means sums of squares and products:

for acreage,

$$\frac{(15{,}911)^2}{372} + \frac{(11{,}931)^2}{159} + \cdots \frac{(12{,}032)^2}{278} - 8{,}999{,}453 = 606{,}472$$

for yield,

$$\frac{(16{,}027)^2}{372} + \frac{(7{,}213)^2}{159} + \cdots \frac{(9{,}606)^2}{278} - 4{,}180{,}724 = 40{,}691$$

[8] Snedecor, George W. Calculation and interpretation of analysis of variance and covariance. Collegiate Press, Inc., Ames, Iowa (1934).

for products,

$$\frac{(15{,}911)\ (16{,}027)}{372} + \frac{(11{,}931)\ (7{,}213)}{159} + \cdots \frac{(12{,}032)\ (9{,}606)}{278} -$$

$$6{,}133{,}859 = 58{,}968.$$

The sums of squares and sum of products *within areas* are then obtained by subtraction.

for acreage, 4,513,603 — 606,472 = 3,907,131
for yield, 348,887 — 40,691 = 308,196
for products, 77,776 — 58,968 = 18,808

The correlation and regression coefficients[9] are calculated as usual. As examples, the correlation between yield and acreage *between means of areas* is

$$\frac{58{,}968}{\sqrt{(606{,}472)\ (40{,}691)}} = 0.3754,$$

while the regression coefficient of yield on acreage *between means of areas* is

$$58{,}968/606{,}472 = 0.0972.$$

The only significant correlation (and, therefore, regression coefficient) is the one for *total.* The larger correlation *between means of areas* is based on too few degrees of freedom for significance. The one *within areas* is an average of the seven individual area coefficients which will be segregated later. All this indicates that there is no very striking relation between yield and acreage in the state as a whole.

It is now necessary to introduce into our area comparisons a correction which makes allowance for such variations in yield as may be due directly to variation in corn acreage. When this is done, there remains the sum of squares of errors of estimate of corn yield for each of the lines of table 4. Since the correlations are small, these sums of squares will be little less than those under the caption, Sy^2. The formula for the computation is

$$Sy^2 - (Sxy)^2/Sx^2,$$

the second term being the indicated correction.

Applying this to the first line in table 4,

$$348{,}887 - (77{,}776)^2/4{,}513{,}603 = 347{,}547.$$

[9] Wallace, H. A., and Snedecor, George W. Correlation and machine calculation. Iowa State College Official Publication. 30: No. 4. Revised edition (1931).

In a similar manner, the sums of squares of errors of estimate are computed for the other two groups, all the results being recorded in table 5. The degrees of freedom in each line have been reduced by one, corresponding to the regression coefficient (or correlation coefficient) implicitly used in the adjustment.

Owing to the slight decrease in the *between means of areas* sum of squares, coupled with the relatively great reduction of the degrees of freedom, the mean square of errors of estimate *between area means* is actually greater than the unadjusted

TABLE 5. SUMS OF SQUARES OF ERRORS OF ESTIMATE, MEAN SQUARES, AND DEGREES OF FREEDOM FOR CORN YIELD AS ESTIMATED FROM CORN ACREAGE

Source of variation	Degrees of freedom	Sum of squares of errors of estimate	"Adjusted" Mean square
Total	2,203	347,547	
Within areas	2,197	308,105	140.2
Between area means	5	34,958	6,991.6
Remainder	1	4,484	4,484.0

mean square in table 3. This indicates the failure of the attempt to associate any of the variation in yield with the corresponding area acreage. The adjusted mean square for *within areas* turns out to be the same as the corresponding mean square in table 3. If there is any association between yield and acreage within the individual areas, the effect is lost in the average *within areas* of this table.

After adjustments are completed, there remains a single degree of freedom together with an associated sum of squares,

TABLE 6. ANALYSIS OF COVARIANCE AND ERRORS OF ESTIMATE OF CORN YIELD AND CORN ACREAGE IN SEVEN TYPE-OF-FARMING AREAS IN IOWA

Type of farming area	Degrees of freedom	Sum of squares and products			Correlation Coefficient	Regression Coefficient	Degrees of freedom	Sum of squares of errors of estimate $Sy^2 - (Sxy)^2/Sx^2$
		Corn acreage Sx^2	Corn Yield Sy^2	Product Sxy				
1. Dairy	371	276,632	48,352	9,950	0.0860	0.0360	370	47,994
2. Northern cash grain	158	314,058	19,181	−16,479	−0.2123**	−0.0525**	157	18,316
3. Western meat	402	1,181,496	53,344	− 887	−0.0035	−0.0008	401	53,343
4. North Central cash grain	420	1,056,578	43,532	3,822	0.0178	0.0036	419	43,518
5. Eastern meat	395	518,149	70,168	3,860	0.0202	0.0074	394	70,139
6. South Central pasture	175	336,226	25,686	13,956	0.1502*	0.0415*	174	25,107
7. Southern pasture	277	223,992	47,933	4,586	0.0443	0.0205	276	47,839
Sum							2191	306,256

* Significant.
** Highly significant.

4,484. This remainder furnishes a test of the significance of the difference between the regression coefficients in the last two lines of table 4.

As has been intimated, the real interest in this problem lies in the correlations *within* the seven areas of the state. The necessary calculations are summarized in table 6. The sums of squares and products of deviations from means in each area were computed from the original data given in table 2. For example,

Dairy area correction terms:

$$\text{for acreage, } \frac{(15{,}911)^2}{372} = 680{,}537$$

$$\text{for yield, } \frac{(16{,}027)^2}{372} = 690{,}497$$

$$\text{for products, } \frac{(15{,}911)\ (16{,}027)}{372} = 685{,}499$$

Dairy area sums of squares and products:

for acreage,	957,169 — 680,537 =	276,632
for yield,	738,849 — 690,497 =	48,352
for products,	695,449 — 685,499 =	9,950

Each correction term is subtracted from the corresponding sum of squares or products of entries from table 2 to determine the sums of squares and products in line 1 of table 6. The correlations and regressions are computed as previously illustrated.

Since an examination of the results recorded in table 6 show that there are differences among the correlations and regression coefficients in the seven areas, it is desirable to find out whether they vary only as much as might be expected in random sampling from a homogeneous population, or whether the areas differ significantly among themselves in the relation between acreage and corn yield. As a first step in testing the differences among the area regression coefficients, the sum of squares of corn yield in each area is adjusted for its own regression on corn acreage, that is, the sum of squares of errors of estimate is computed for each area in the manner already explained. The results for the seven areas are recorded in the last column of table 6. Each source of variation loses a degree of freedom corresponding to the regression coefficient used in the adjustment. The sums of squares and degrees of freedom for the seven areas are added

to give the *within areas* sum of squares, adjusted for individual regressions. This *within areas* sum of squares (306,256) is the pooled sum of squares of the deviations of the individual observations, each measured from its own area regression. Contrast with this, the *within areas* sum of squares of errors of estimate (308,105) from table 5, which is adjusted for the average *within area* regression: that is, instead of taking the deviation of each observation from its area regression, the deviation is measured from an average regression common to all areas.

The data for these two sums of squares of errors of estimate *within areas* are entered in table 7. The failure of the individual area regressions to estimate the corn yield within the several areas is a valid estimate of error, so that other mean squares may be tested for significance against the value, 139.8. It happens that, in table 7, both the *within areas* mean squares are

TABLE 7. ANALYSIS OF ERRORS OF ESTIMATE "WITHIN AREAS"

Source of variation	Degrees of freedom	Sum of squares	Mean square
Within areas, from average regression, table 5	2,197	308,105	140.2
Within areas, from individual regressions, table 6	2,191	306,256	*139.8*
Remainder, between area regression coefficients	6	1,849	308.2*

* Significant.

approximately 140. These estimates usually differ somewhat but not much.

The six degrees of freedom associated with the seven area regressions yield the mean square, 308.2. This furnishes the desired test of significance of the differences among the area regression coefficients of table 6. The value,

$$F = 308.2/139.8 = 2.2,$$

just beyond the 5 per cent level, is not convincingly significant. Nevertheless, it will be assumed in what follows that some of the type-of-farming areas are characterized by differing regressions of yield on acreage. This assumption is supported also by the fact that of the two significant regressions in table 6 one is positive and the other negative. On the other hand, it is clear that the other five regressions may exhibit only sampling variation from a homogeneous population.

Before attempting to interpret the results already obtained, it is enlightening to consider the graphical representation of the regressions under discussion. Regression equations[10] may be

[10] Wallace, H. A., and Snedecor, George W., *loc. cit.*

calculated from the data in tables 1, 4, and 6. For example, that for the dairy area is

$$\begin{aligned}\overline{Y} &= 43.1 + 0.0360(X - 42.8)\\ &= 0.0360X + 41.6\end{aligned}$$

where X indicates the farm acreage in corn, and $\overline{Y}$, the estimated corn yield in bushels per acre. The average regression of yield on acreage *within areas* is

$$\begin{aligned}\overline{Y} &= 43.5 + 0.0048(X - 63.9)\\ &= 0.0048X + 43.2\end{aligned}$$

and the *between area means* regression is,

$$\begin{aligned}\overline{Y} &= 43.5 + .0972(X - 63.9)\\ \overline{Y} &= .0972X + 37.3\end{aligned}$$

These and the other area regressions are plotted in figure 1. The variation in the slopes of these area regression lines has been previously demonstrated to be significant. The slope of the line for the Northern Cash Grain area is significantly negative, while

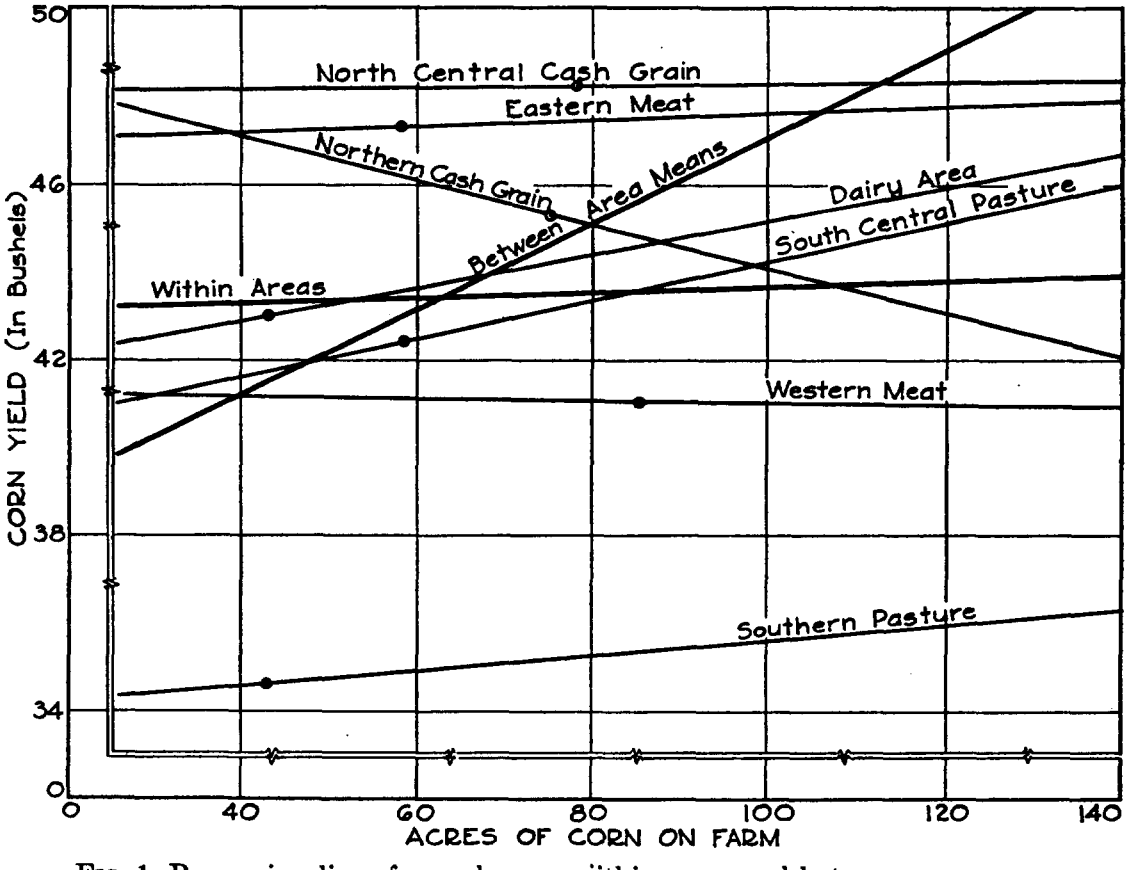

FIG. 1. Regression lines for each area, within areas and between area means

that for South Central Pasture area is significantly positive.

The negative regression in the Northern Cash Grain area is apparently associated with varying management practices of the individual farmers. Those who raise corn on too great a portion of their land, seeking immediate cash returns, fail to preserve the soil fertility, the result being a decrease in yield per acre.

The situation in the South Central Pasture area is more complicated. Many of the farms in this region are hilly and unsuited

TABLE 8. ANALYSIS OF ERRORS OF ESTIMATE OF CORN YIELD BY TYPE-OF-FARMING AREAS

Source of variation	Degrees of freedom	Sum of squares	Mean square
Total	2,203	347,547	
Within areas, from individual regressions	2,191	306,256	139.8
Between area regression coefficients	6	1,849	308.2*
Between area means	5	34,958	6,991.6**
Remainder	1	4,484	4,484.0**

* Significant.
** Highly significant.

to corn production. The small corn acreages are associated with small farms on which erosion has in many instances progressed so far as to cause serious impairment of the fertility. The farms with large corn acreages are usually on level ground where the fertility has been maintained. In this area, therefore, there is a significant positive correlation between acreage and yield. To a less extent, this proved true in the Southern Pasture area. The non-significant correlation in this area may be due to some peculiarity of the sample of farms. It is believed that much the same relations exist in these two areas, and also that the yields associated with small acreages have been tending to become smaller for a number of years.

Two observations may now be made. (i) The adjusted mean square *between area means*, in table 8, measures the deviations of these means from their own regression line. The graph appears in figure 1. The circled points represent the area means. The somewhat contradictory statement may be made that the most notable feature of this area mean regression is its failure to account for the mean area yields. The large deviations explain the non-significant correlation *between means of areas* in table 4. It is questionable whether this regression represents any population relationship or is merely a sampling accident. (ii) The *remainder* mean square, 4,484, furnishes a test of the significance of the difference between the last two regression coefficients in table 4, those *between area means* and *within areas* from average

regression. The lines are indicated in figure 1. The F value, $4{,}484/139.8 = 32.1$, leaves no doubt of the significance of this difference. Nevertheless, the doubtful objectivity of the regression for area means makes the interpretation questionable. Subject to this limitation, quantitatively at least, the relation between acreage and yield on the farms in the state is different from that relation for the means of the type-of-farming areas.

Summary

1. Analysis of covariance has been used to study the relation between corn yield and corn acreage within and between the seven types-of-farming areas of Iowa in 1933. The relationships are quantitatively small.

2. In the Northern Cash Grain area a significant negative correlation is interpreted as indicating decreased fertility on farms where corn is raised so persistently as to impair fertility.

3. In the South Central Pasture area the significant positive correlation reflects the tendency for operators of small farms to keep land in corn despite the effects of erosion.

4. In the other five areas of the state there is little relation between corn acreage and corn yield.

5. There is a highly significant difference in the mean area corn yields of the seven type-of-farming areas. In this 1933 sample, the North Central cash grain farms had a larger average corn yield than the other areas.