# PREDICTING APPLE SIZE: PRELIMINARY RESULT

BU-201-M O. Ladipo and A. Hedayat July, 1965

## Abstract

The final size of an apple, that is to say the size of an apple at harvest, does not depend only on its size four weeks before harvest, but it also depends on a number of factors which, acting simultaneously, determine how big or small the apple is. In order to predict apple production more accurately several factors must be considered.

---

Biometrics Unit, Cornell University, Ithaca, New York

PREDICTING APPLE SIZE: PRELIMINARY RESULT

BU-201-M O. Ladipo and A. Hedayat July, 1965

Nature of the Preliminary Problem:

The question here is "Can we predict final apple size on the basis of one factor, namely the size of the apple in mid-August?". Our first reaction is that final apple size is a response which is affected by a number of quantitative factors. Which of these factors are the most efficient in predicting final apple size? From here on we will refer to final apple size as Y; the other factors will be as follows:

$X_1$ = Apple size as of (or near) August 10

$X_2$ = Apple size as of (or near) August 21

$X_3$ = Nitrogen level as of August 15

$X_4$ = Total number of dry days

$X_5$ = Number of dry days before August 15

$X_6$ = Crop-load

The ultimate purpose of this study is to predict total production of apples in New York State as early as possible before harvesting time. At this stage we want to show that several factors affect production. After finding those factors that are most efficient in predicting apple size we will then aim at constructing a formula (a model) and designing a sampling procedure both of which, we hope, can be used annually before harvesting to predict what the year's total apple production will be.

---

Biometrics Unit, Cornell University, Ithaca, New York

The Data:

The data we are working with in this proliminary stage were collected at the Geneva Experimental Station. The set of data consists of observations in six orchards. Each observation is "fruit volume in cubic inches". The measurements were taken at intervals of approximately ten days starting July 2nd and over a period of four years -- 1961-1964.

It is our feeling that the data were collected as a matter of routine. By this we mean they lacked any design as to the purpose of the collection. We do not know whether the value 6.13 (for August 21st, orchard No. 1, for 1962) is the average volume of all apples for that orchard or a sample mean volume. We, for the purpose of the present problem, regard it as a mean volume of the apples in a particular orchard. Under "Soil Moisture" it is difficult to decide what "Dry days" really mean. Here we assume, following the footnotes on the data sheet, that the figures represent numbers of days with at least minimum amount of moisture. We scored crop-load as light = 1, moderate = 2, and heavy = 3.

We do feel that temperature is an important variable which whould have been included. Maybe the pomologists can help us with an explanation for its omission. Other things we would have wanted to see are rainfall (in addition to dry days), thinning of apples and age of orchard. With sufficient information we can find the yield per acre and knowing the number of acres per orchard we will be able to establish the production of apples in an orchard.

Analysis:

There is no gainsaying the fact that final apple size (i.e., its size in mid-September) is dependent on its size in mid-August. This is justified by the high correlation coefficient $r_{yx_1}$, which is .88. The question, however, is: Can we do better in the prediction of Y if we consider other factors?. In answer to this we resorted to multiple linear regression (see diagrams I-V, to each of which a regression line can be fitted). Setting up the equation we have:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

We claim that adding more variables may give us more information. To check this we present Table 1.

Table 1

| Variables added in order | Meaning of variable | R | $R^2$ | Additional information gained | Remarks |
|---|---|---|---|---|---|
| $X_1$ | size as of 8/15 | .88 | .776 | - | Almost 23% of variability in Y is not accounted for by regression |
| $X_4$ | Total dry days | .9129 | .8334 | .0574 | Additional information gained by considering these three factors is 13.85% and only 8.6% of variability is not accounted for by regression |
| $X_3$ | Nitrogen level | .9440 | .8912 | .0578 | |
| $X_6$ | Crop-load | .9563 | .9145 | .0233 | |
| $X_2$ | Size as of 8/21 | .9614 | .9242 | .0097 | Additional gain is only 1.09% |
| $X_5$ | Dry days before 8/15 | .9620 | .9254 | .0012 | |

From the above table we observe that not all factors contribute a lot of information when added. The bulk of the extra information comes from fixing $X_4$ and $X_3$ which yields 11.5% more information while if $X_2$ and $X_5$ are fitted along with $X_1$, $X_3$, $X_4$ and $X_6$ we only gained 1.01% more information. It seems, therefore, not worth while to include $X_2$ and $X_5$. Our equation then becomes

$$Y = \alpha + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6$$

If, however, we can take two measurements of fruit volume the addition of $X_2$ will give us a little additional information and our regression equation will then be

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6$$

Based on the crude data at our disposal we regard this as our optimal equation. The inclusion of $X_5$ adds only 0.12% to the amount of information. Some of the combinations with the amount of information are given in Table 2. In Table 3 we try to show that though 92.54% of variability in Y may be accounted for by regression, yet the individual coefficient of regression ($b_i$) need not be significant at 5% level. In Table 3 we find that, using a t-test not all of the $b_i$'s are significant at 5% level.

In order to test for the relative importance of each of the independent variables we considered $\beta_i'$ "since each $\beta'$ is independent of the original units of measurement, a comparison of any two indicates the relative importance of the independent variables involved". (Steel and Torrie, 1960, p. 284.) In Table 3 we showed the relative importnace of each factor in predicting Y.

Table 2

| Variables in equation | R | $R^2$ | Variables left out |
|---|---|---|---|
| $X_1, X_2, X_4, X_5, X_6$ | .94 | .88 | $X_3$ |
| $X_1, X_2, X_3, X_4, X_5$ | .9565 | .9149 | $X_6$ |
| $X_1, X_2, X_3, X_5, X_6$ | .9558 | .91355 | $X_4$ |
| $X_1, X_2, X_3, X_4, X_6$ | .9614 | .9242 | $X_5$ |
| $X_1, X_2, X_3, X_6$ | .9155 | .84 | $X_4, X_5$ |
| $X_1, X_2, X_4, X_5$ | .9373 | .8785 | $X_3, X_6$ |
| $X_1, X_2, X_4$ | .9138 | .8350 | $X_3, X_5, X_6$ |
| $X_1, X_2, X_5$ | .8967 | .8041 | $X_3, X_4, X_6$ |
| $X_1, X_2, X_6$ | .8816 | .7772 | $X_3, X_4, X_5$ |
| $X_1, X_2, X_3$ | .9106 | .8292 | $X_4, X_5, X_6$ |
| $X_1, X_2, X_3, X_5$ | .9424 | .8882 | $X_4, X_6$ |
| $X_1, X_2, X_3, X_4$ | .95235 | .90696 | $X_5, X_6$ |
| $X_1, X_2, X_5, X_6$ | .8983 | .8069 | $X_3, X_4$ |
| $X_1, X_2, X_4, X_6$ | .91503 | .83729 | $X_3, X_5$ |

Nitrogen level is the most consistent. The importance of the other factors (as shown by their ranks) changes with respect to the combination. If we leave out $X_4$ and replace it with $X_5$ we lose only 1.06% of the information. We will be forced to do this because, in predicting final apple size, the total number of dry days will not be available to us, and we will,therefore, use number of dry days before August $15^{th}$. It has been stated in the letter of March 11, 1965 from Professor Forshey to Professor Federer that "at least 30 days of deficient soil moisture are required for a significant reduction in fruit size". If this

Table 3

| Combination of variables | $b_i$ | Value of $t_i$ for testing $b_i$ | Level of $t_i$ being significant | $b_i'$ | Rank of $X_i$ in predicting Y | R | $R^2$ | Remarks |
|---|---|---|---|---|---|---|---|---|
| 1 Size at 8/10 | +0.384042 | 0.355 | > 50% | +0.232962 | 4 | 0.962 | 0.9254 | |
| 2 Size at 8/21 | +0.476965 | 0.553 | > 50% | +0.321627 | 3 | | | |
| 3 Nitrogen | +3.50273 | 2.379 | 3% | +0.423022 | 2 | | | |
| 4 Total days | -0.020606 | 1.195 | 25% | -0.443585 | 1 | | | |
| 5 Day to 8/15 | +0.010286 | 0.373 | > 50% | +0.144365 | 5 | | | |
| 6 Crop-load | -0.20787 | 1.127 | 25% | -0.143572 | 6 | | | |
| Constant term | -2.10978 | | | | | | | |
| 1 | +0.889263 | 4.074 | 1% | +0.059346 | 5 | 0.961 | 0.9242 | Deleting variable five or total dry days before 8/15 does not decrease the amount of information for predicting Y. |
| 2 | +0.693389 | 1.135 | 25% | +0.467566 | 1 | | | |
| 3 | +3.383050 | 3.387 | 1% | +0.462607 | 2 | | | |
| 4 | -0.014381 | 3.371 | 1% | -0.309577 | 3 | | | |
| 6 | -0.238530 | 1.510 | 15% | -0.164748 | 4 | | | |
| Constant term | -6.70803 | | | | | | | |
| 1 | -0.499068 | 0.619 | > 50% | -0.302737 | 3 | 0.956 | 0.914 | Here $b_1$ has negative sign that should be interpreted, because we expect it should have always positive sign. |
| 2 | +1.162220 | 1.767 | 15% | +0.783708 | 1 | | | |
| 3 | +4.450220 | 3.512 | 1% | +0.537449 | 2 | | | |
| 5 | -0.021419 | 2.954 | 1.5% | -0.301488 | 4 | | | |
| 6 | -0.295910 | 1.713 | 15% | -0.204421 | 5 | | | |
| Constant term | -3.63809 | | | | | | | |
| 1 | +0.889263 | 4.074 | 1% | +0.53931 | 1 | 0.956 | 0.914 | Again here it shows that deleting $X_5$ does not decrease the amount of information, and by these four variables we can predict Y as close as with 6 variables in case one. |
| 3 | +3.192230 | 1.135 | 25% | +0.385524 | 2 | | | |
| 4 | -0.014794 | 3.387 | 1% | -0.318462 | 3 | | | |
| 6 | -0.272238 | 3.371 | 1% | -0.188029 | 4 | | | |
| Constant term | -1.07596 | 1.510 | 15% | | | | | |

statement is true we would expect a significant difference between the means of two groups -- apple size in orchards with more than 30 days of deficient soil moisture and apple size in orchards with less than 30 days of deficient soil moisture. We tested this and found that the difference is not significant. The test is as follows:

$\bar{X}_1 = 7.96$ mean for orchards with > 30 dry days

$\bar{X}_2 = 8.40$ mean for orchards with < 30 dry days

$$S_p^2 = (SS_{X_1} + SS_{X_2})/(n_1+n_2-2)$$
$$= 19/14$$

$$S_{\bar{d}} = \sqrt{\frac{19}{14}\left(\frac{16}{63}\right)} = .624$$

$$t = .72/.624 = 1.153 \text{ with } 14 \text{ d.f.}$$

$$t < t_{.05},14 = 2.145$$

Conclusion:

Difference, $\bar{X}_2 - \bar{X}_1$, is not significant. If we, however, consider this factor (dry days before August $15^{th}$) with some other factor, say Nitrogen level, we might reach a different conclusion if our earlier conclusion is wrong and the letter right. To check this we therefore examined the regression of final apple size in each of the two groups on Nitrogen to see if there is any significant difference in the "groups" adjusted means -- 7.5734 and 8.5431. With an $F = 3.0877 < F_{.05}(1,12) = 4.75$ we conclude that our earlier conclusion holds and that the statement in the letter needs some explanation.

Summary of calculations for the above follows.

$$\hat{U}_{y_1 \cdot \bar{\bar{X}}} = \bar{Y}_1 - b_1(\bar{X}-\bar{\bar{X}}) = 7.5734$$

$$\hat{U}_{y_2 \cdot \bar{\bar{X}}} = \bar{Y}_2 - b_2(\bar{X}-\bar{\bar{X}}) = 8.4531$$

$$H_0 : \hat{U}_{y_1 \cdot \bar{\bar{X}}} - \hat{U}_{y_2 \cdot \bar{\bar{X}}} \neq 0$$

Pooled residual SS = 10.23 with 12 d.f.

Sum of squares for $\bar{Y}_{\bar{\bar{X}}}$ = 2.6246 with 1 d.f.

$$F = \frac{2.6246/1}{10.23/12} = \frac{12(2.6246)}{10.23} = 3.0877$$

$$F_{.05}(1,12) = 4.75$$

Conclusion:

$H_0$ is rejected.

Anticipating the question of why we fitted a linear rather than a curlinear regression we constructed diagrams I-V. We can fit a linear regression to each one of the five and this justifies the fitting of multiple linear regression.

In all the above analyses we ignored the year effects. The available data for each year are small and with small samples we have to be cautious about any conclusions, tests of significance and be a little skeptical, too (Snedecor, 1956). Our reluctance to use small samples (each set of annual data) is based on the fact that with five independent variables ($X_1$, $X_2$, $X_3$, $X_5$, $X_6$) and only six observations we have zero degrees of freedom for error (n-k-1 = 6-5-1), whereas when we pooled all annual data (1961 - 1964)we have 16 observations and hence 10 degrees of freedom for error in the multiple regression analysis of variance. Table 4 shows the different F's for each variable after the others have been fitted. This is done by the use of step-wise multiple regression analysis (Searle and Primer, 1964).

Table 4

| Variable being fitted | F, after preceeding variables are fitted | Error degrees of freedom | Standard error of Y | Remarks: (testing at $\alpha$ = .05) |
|---|---|---|---|---|
| $X_1$ | 48.61 | 14 | 0.58 | Significant |
| $X_4$ | 4.45 | 13 | 0.52 | Not significant |
| $X_3$ | 6.36 | 12 | 0.44 | Significant |
| $X_6$ | 3.00 | 11 | 0.404 | Not significant $F(1,11)$ |
| $X_2$ | 1.29 | 10 | 0.40 | Not significant |
| $X_5$ | 0.14 | 9 | 0.42 | Not significant |

A person might expect that the standard error of Y has something questionable, especially the last three values. Recall from Table 1 that after fitting $X_1$, $X_3$, $X_4$, $X_6$ and $X_2$, only 0.12% of variability in Y is explained by fitting $X_5$. This small amount of reduction in the variability of Y does not seem to be worth losing one degree of freedom, and this explains why the last standard error of Y in Table 4 is greater than the preceding one. The degrees of freedom degrees by 10% (i.e. from 10 to 9), while the variability degrees only by 0.12%.

It is easy to see from the above that the most important factors are size as of August 15th (10-15), nitrogen level, while $X_4$ - total dry days is close to being significant because $F_{.05}(1,13) = 4.67$. This is the same type of conclusions we reached via Table 3.

The analysis of variance for all the combinations mentioned so far are available in the computer sheets.

As a result of the scatter diagrams we feel we might do better under a log transformation. Suppose we set up a regression formula of the form

$$C = \alpha^* X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} \cdots X_k^{\beta_k}$$

where $\alpha^*$, $\beta_1$, $\beta_2$, $\cdots$, $\beta_k$ are the constants to be estimated and $X_1, X_2, \cdots, X_k$ are observed variables, then the transformation is: $Y = \log C$ if $\log \alpha^* = \alpha$ then

$$Y = \alpha + \beta_1 \log X_1 + \beta_2 \log X_2 + \cdots + \beta_k \log X_k \quad .$$

We hope to report on this as the analyses come out of the computer.

LITERATURE CITED

1. Steel, R. G. D., and Torrie, J. H. Principles and Procedures of Statistics. McGraw-Hill, New York, 1960.

2. Snedecor, G. W. Statistical Methods. (Fifth edition). Iowa State Univ. Press, Ames, Iowa, 1956.

3. Searle, S. R., and Primer, Mrs. P. L. Multiple Regression Analysis. (Third Issue), 1964.

Diagram I: Final Size × Total Moisture (Dry Days)

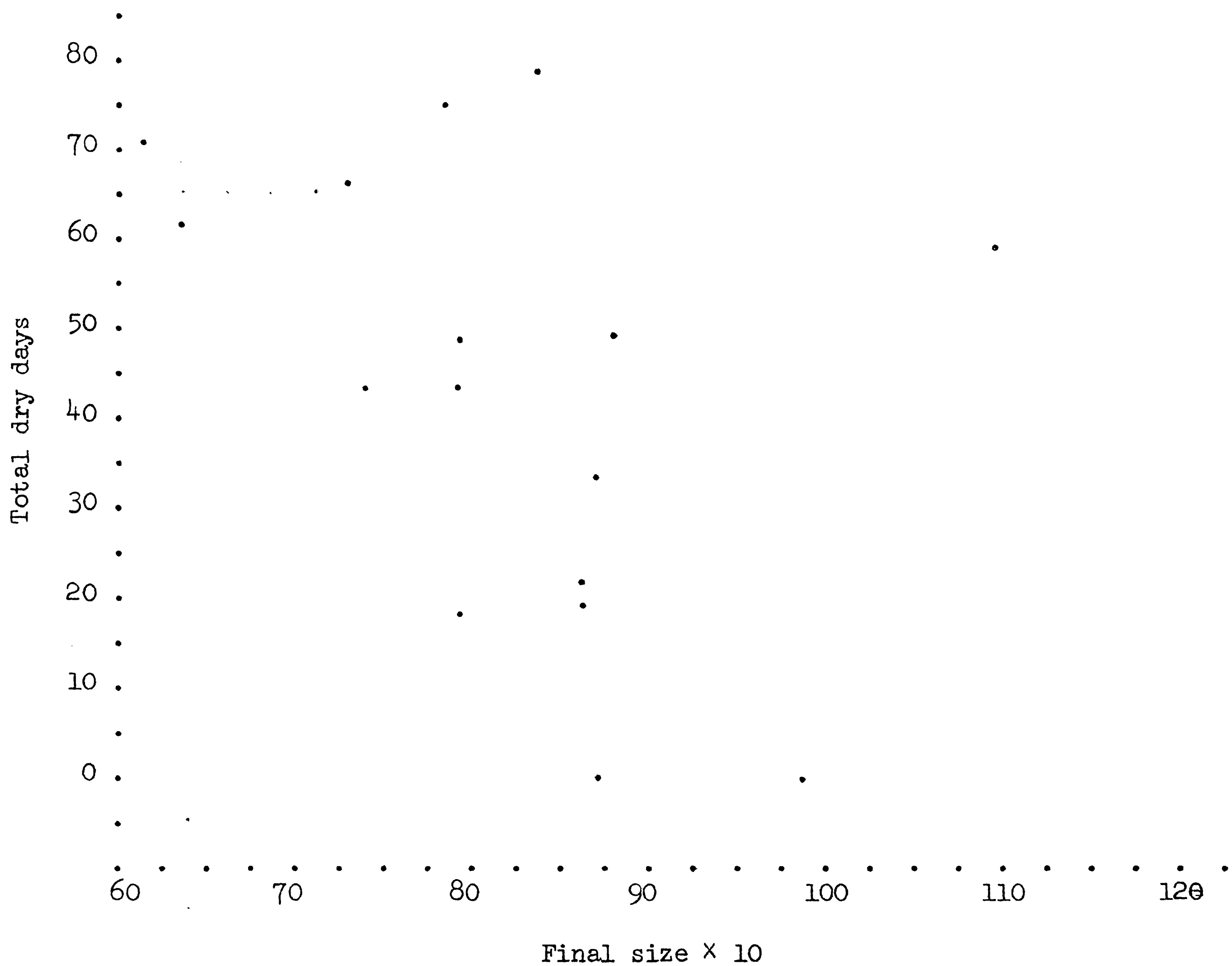

Diagram II: Final Size × Number of Dry Days Before August 15

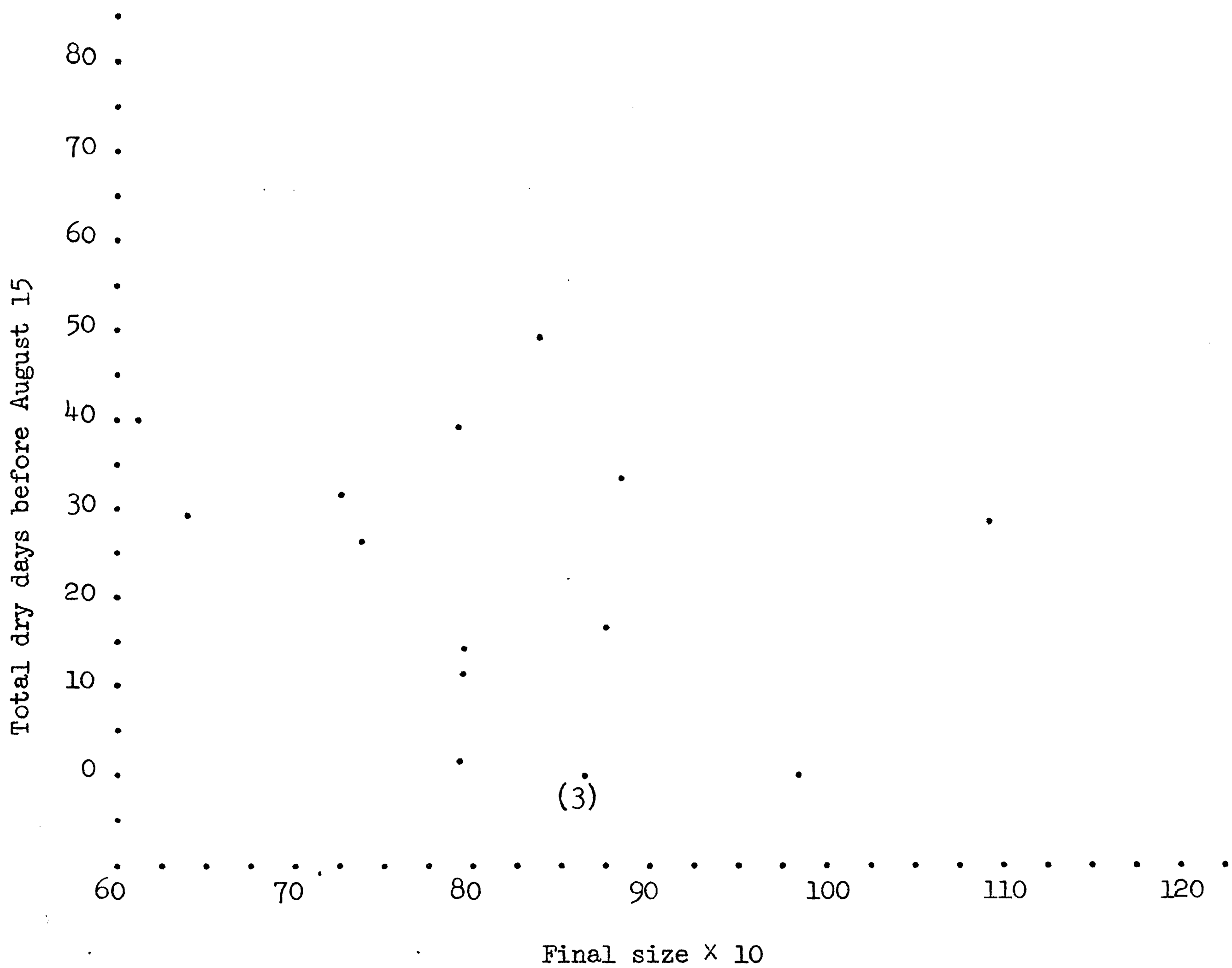

Diagram III: Final Size X Nitrogen Level

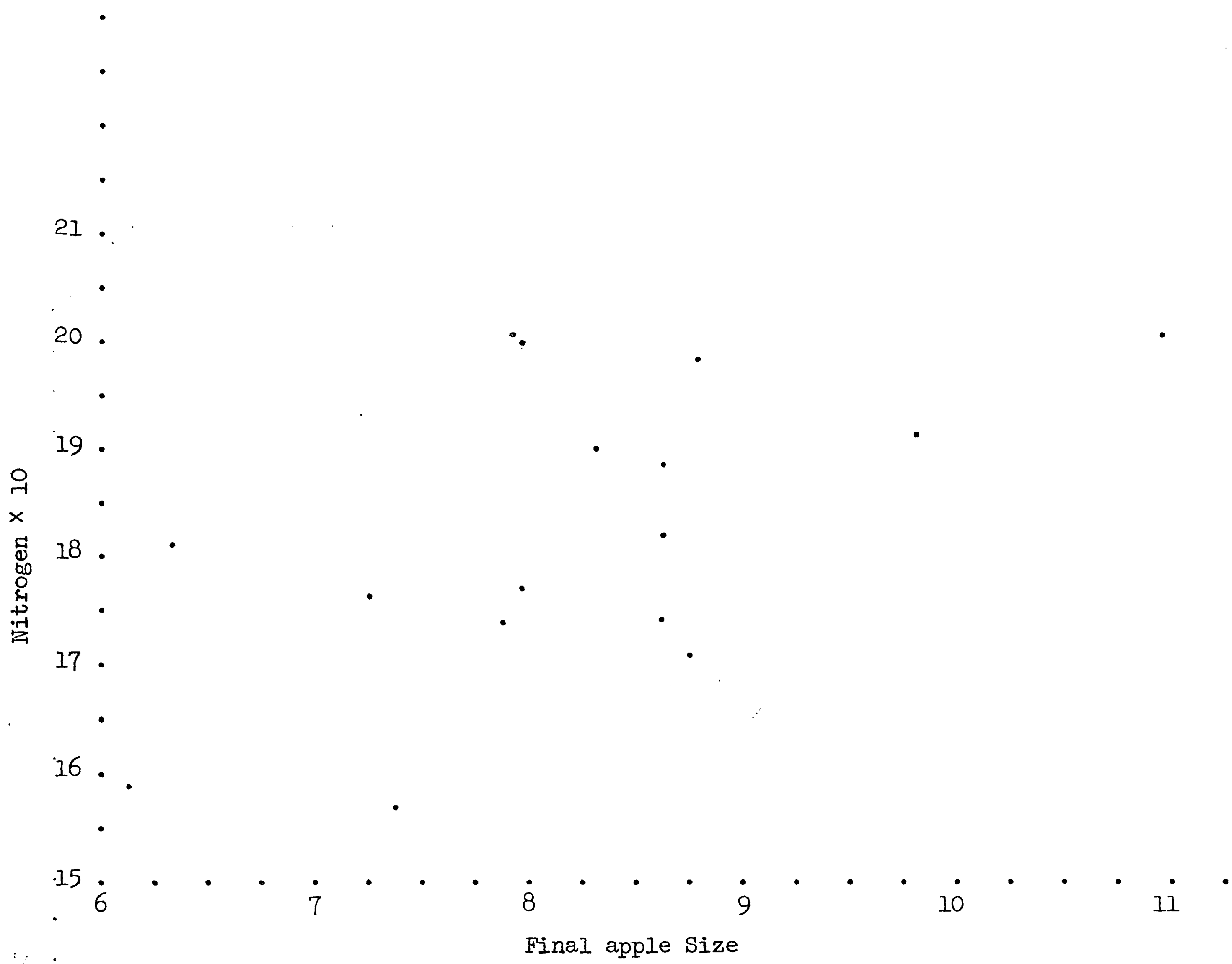

Diagram IV: Final Size X Size at August 10

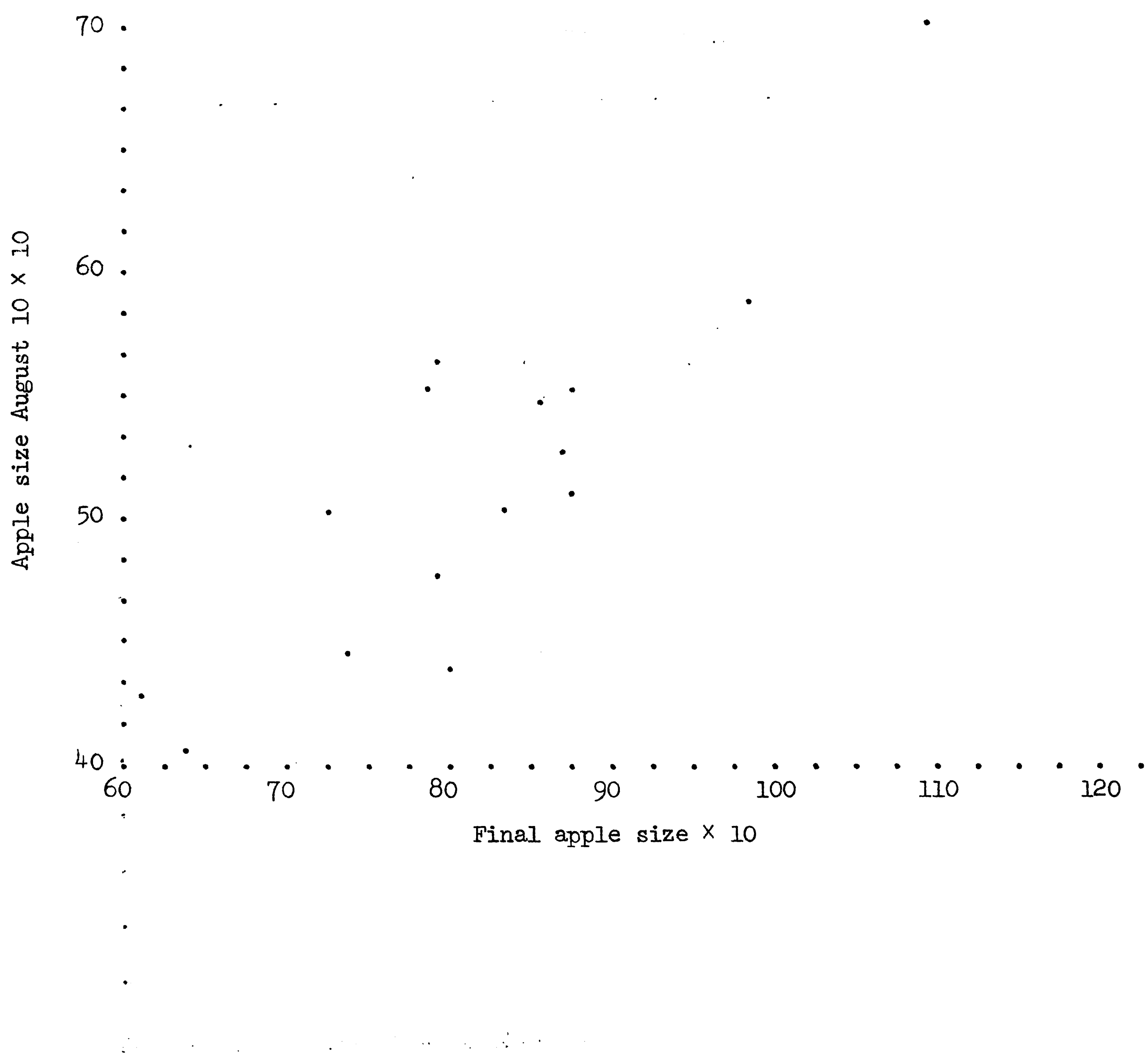

Diagram V: Final Size × Cropload