


**RELIABILITY ON PREDICTED VALUES
IN REGRESSION ANALYSIS FOR
FORESTRY RESEARCH**

by
Y. (Jim) Lee

**FOREST RESEARCH LABORATORY
VICTORIA, BRITISH COLUMBIA
INFORMATION REPORT BC-X-36**

**FORESTRY RESEARCH SERVICES
DEPARTMENT OF FISHERIES AND FORESTRY
AUGUST, 1969**



RELIABILITY ON PREDICTED VALUES IN REGRESSION ANALYSIS
FOR FORESTRY RESEARCH

by Y. (Jim) Lee

FOREST RESEARCH LABORATORY
VICTORIA, BRITISH COLUMBIA
INFORMATION REPORT BC-X-36

DEPARTMENT OF FISHERIES AND FORESTRY

AUGUST, 1969

CONTENTS

	<u>Page</u>
Introduction - - - - -	1
Methods of examining residuals - - - - -	1
Examples - - - - -	4
Part I. Hypothetical examples - - - - -	4
Part II. Actual examples - - - - -	8
Discussion - - - - -	8
(1) Correlation coefficient - - - - -	14
(2) Coefficient of determination - - - - -	14
(3) Standard error of estimate - - - - -	15
(4) Analysis of variance - - - - -	15
(5) Confidence limits - - - - -	15
Conclusion - - - - -	17
Acknowledgement - - - - -	17
Literature Cited - - - - -	17

Reliability on Predicted Values in Regression Analysis
for Forestry Research

by

Y. (Jim) Lee^{1/}

Introduction

The major objective of regression analysis in forestry research is to establish relationships which make it possible to predict one or more variables in terms of others. The most obvious and frequently used regression equation is that of volume on dbh and height (Spurr, 1952). Knowing the reliability of predicted values for the regression equation is of primary importance. The correlation coefficient, coefficient of determination and the standard error of estimate are commonly used as criteria in the selection of regressions. At times, analysis of variance is also used or confidence limits expressed (Freese, 1964; Walters, 1967). The critical issue here is that while, by using one or all of these criteria in the selection of a regression, we might know that the selected regression is better than others, but we do not know how much better it is in terms of prediction.

In this paper, which is intended for the forester who is not statistically orientated, the technique of examining residuals between observed and predicted values is described. The technique is illustrated first with hypothetical data and then with actual examples.

Methods of Examining Residuals

Residuals are calculated as the n differences:

$$e_i = Y_i - \hat{Y}_i$$
$$i = 1, 2, 3, \dots, n$$

Where n = the number of observations

e_i = The residual value

Y_i = an observed value

^{1/} Research Scientist, Forest Research Laboratory, 506 West Burnside Road, Victoria, B.C.

\hat{Y}_i = the corresponding predicted values obtained by use of the fitted regression.

The residuals are subsequently plotted against the corresponding \hat{Y}_i .

According to Draper and Smith (1966), should the "plotting" prove to be similar to that shown in Figure 1a, i.e., a "horizontal band" with a constant variance, no abnormality is indicated by the use of the regression. The predicted values would appear not to be invalidated. The narrower the "horizontal band", the greater the reliability of the predicted values.

Should the plotting be similar to that shown in Figure 1b, i.e., "wedge band", the variance is not constant and the predicted values are less reliable. Draper and Smith (1966) state that before determining a regression, the use of weighted least squares or a transformation of the observed values Y_i is needed.

Where the plotting is similar to that shown in Figure 1c, i.e., "descending band", an error in analysis is apparent. Less reliability can be placed upon the predicted values than that shown in 1a. Departure from the fitted regression is such that positive residuals correspond to low predicted values (\hat{Y}_i 's) and negative residuals correspond to high predicted values. The scatter pattern can be caused by incorrectly omitting a β term in the regression analysis.

Should the resultant plot be similar to that shown in Figure 1d, i.e., "arch band", the regression model is inadequate and little reliability can be placed upon the predicted values. In this case, an extra square term, an extra cross-product term, or a transformation on the observed value Y_i is required in the regression analysis (Draper and Smith, 1966).

Other procedures, in addition to the plotting of residuals, may also be useful:

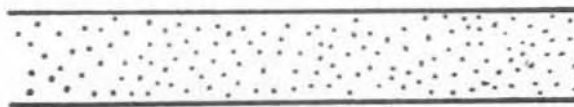
(a) If error is suspected in the sampling process, then it is desirable to increase the sample size or to discard the sample entirely and get one more representative of the population.

(b) Are there systematic patterns in the variation of residuals? Is there a grouping of many positive residuals together, and of many negative residuals together? If so, this would suggest that the regression function fitted is inadequate.

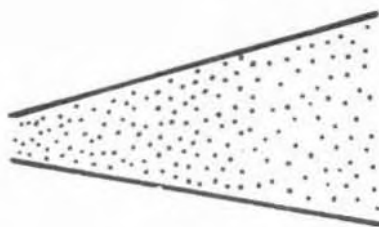
(c) Whether the variation of the residuals varies as the independent variable or variables vary, and if so, how? If the variation of the residuals is not homogeneous, then one may need to decide whether to make a transformation of the variable or variables in order to achieve such homogeneity of residual variation, or whether to use a weighted regression.

FIGURE 1. Scatter patterns of residuals (Draper and Smith, 1966)

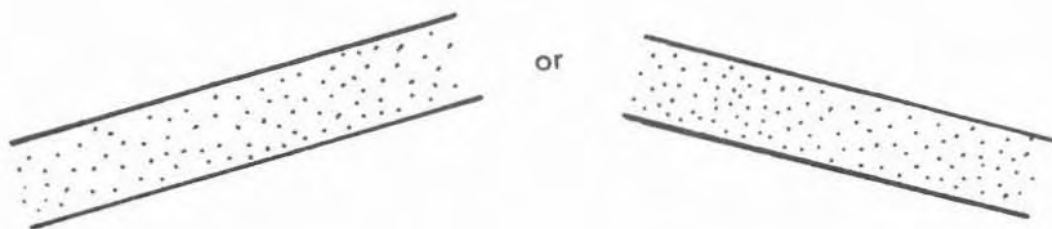
a



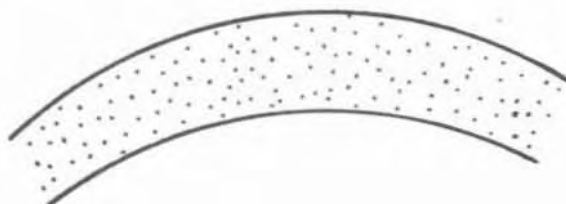
b



c



d



(d) Should the outliers among the residuals be rejected? As a general rule, outliers should be rejected if they can be traced to causes such as error in making or recording the observations, otherwise careful investigation is needed (Draper and Smith, 1966).

It should be noted that there may be complications in the technique of examining residuals when multiple regression, polynomial regression or multi-population cases are involved. In many cases, the plotting of basic data superimposed on the regression line is also useful.

Examples

In the following, the technique of examining residuals is illustrated and demonstrated first with hypothetical examples and then with actual examples.

Part I Hypothetical examples

For each of the 4 sets of hypothetical data, regressions were fitted. Four regressions were obtained and called "original regressions" (Regressions 1, 2, 3 and 4 in Table 1). Residuals for these 4 "original regressions" were calculated and then plotted against the predicted values. Figure 2a, 2b, 2c and 2d show the residual scatter patterns. These 4 patterns were strikingly similar to the scatter patterns provided by Draper and Smith (1966) and shown in Figure 1. According to the methods described previously, improved regressions over the original ones were made, and discussed as follows:

Findings:

(1) Horizontal band

The plotting of residuals against the predicted values from regression 1 shows a "horizontal band" (Figure 2a). Variances of the residuals are homogeneous with errors within ± 3 units. Although no abnormality is indicated by the use of the regression, the reliability of the predicted values depends upon the narrowness of the "horizontal band".

(2) Wedge band

In Figure 2b, an increasing scatter of residuals with increasing predicted Y was observed. This scatter pattern is exactly the same as that of Figure 1b. According to Draper and Smith (1966), an improved regression can be obtained by the use of weighted least squares or a transformation of observed values Y before determining the regression. The following regression models were tested by using standard simple and stepwise forward multiple regression programs and subsequent residuals were plotted.

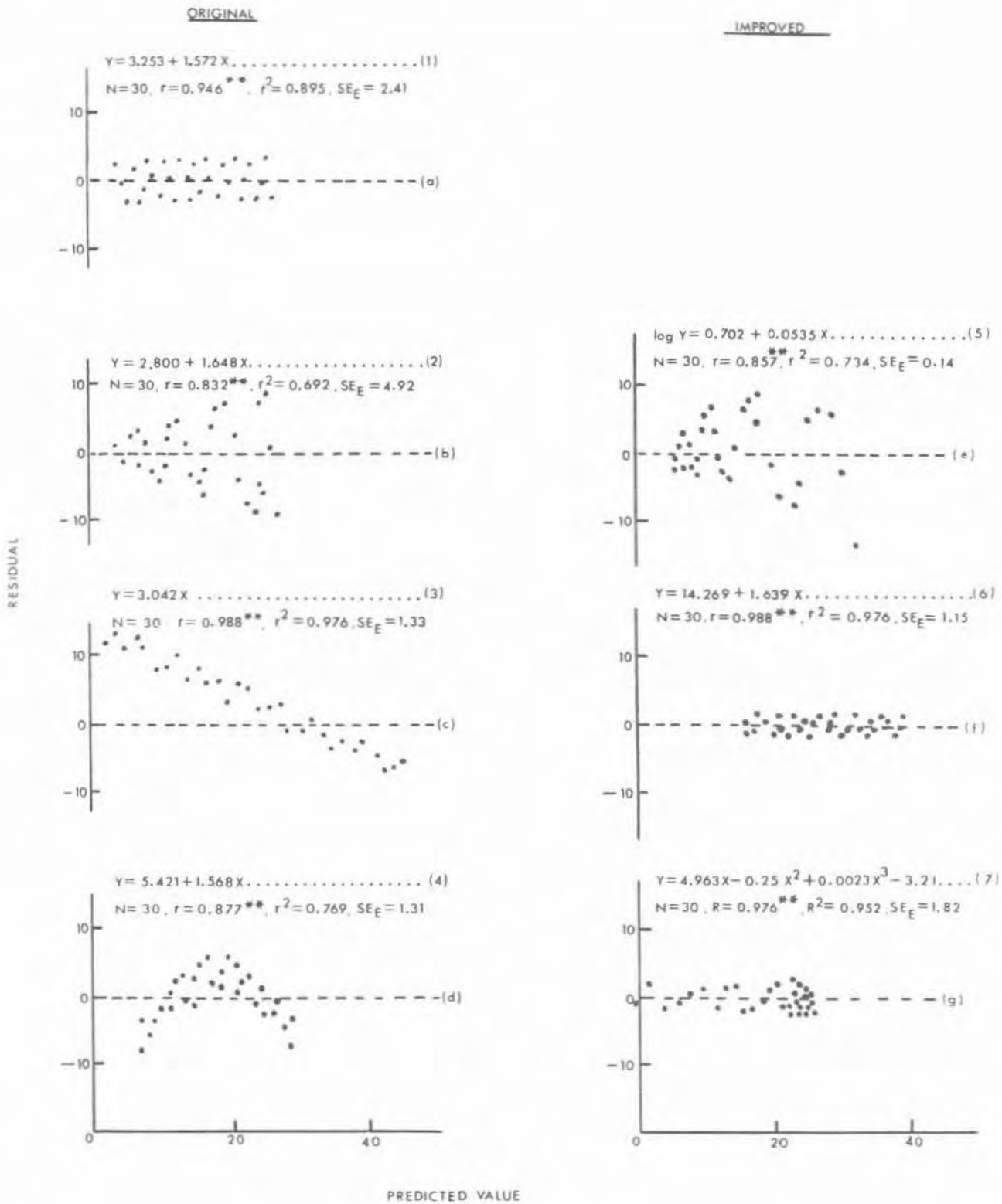
$$\begin{aligned} \log Y &= a + bX \text{ --- (8)} \\ \log Y &= a + bX + cX^2 \text{ --- (9)} \\ \log Y &= a + bX + cX^2 + dX^3 \text{ --- (10)} \end{aligned}$$

Table 1. Hypothetical Examples for Four Scatter Patterns

Original	Improved
$Y = 2.899 + 1.613 X \text{ - - - - - (1)}$ $N = 30, r = 0.803^{**}, r^2 = 0.645, SE_E = 5.35$	
$Y = 2.800 + 1.648 X \text{ - - - - - (2)}$ $N = 30, r = 0.832^{**}, r^2 = 0.692, SE_E = 4.92$	$\log Y = 0.702 + 0.0535 X \text{ - - - - (5)}$ $N = 30, r = 0.857^{**}, r^2 = 0.734, SE_E = 0.14$
$Y = 3.042 X \text{ - - - - - (3)}$ $N = 30, r = 0.988^{**}, r^2 = 0.976, SE_E = 1.31$	$Y = 14.269 + 1.639 X \text{ - - - - - (6)}$ $N = 30, r = 0.988^{**}, r^2 = 0.976, SE_E = 1.15$
$Y = 5.421 + 1.568 X \text{ - - - - - (4)}$ $N = 30, r = 0.877^{**}, r^2 = 0.769, SE_E = 3.84$	$Y = 4.963X - 0.25X^2 + 0.0023X^3 - 3.21 \text{ - (7)}$ $N = 30, r = 0.976^{**}, r^2 = 0.952, SE_E = 1.82$

** = Significant at 1% level

FIGURE 2 Residuals plotted against predicted value for hypothetical examples



$$\begin{aligned}
Y &= a + b (\log X) && \text{--- (11)} \\
Y &= a + b (\log X)^3 && \text{--- (12)} \\
Y &= a + b (\log X) + c (\log X)^2 && \text{--- (13)} \\
Y &= a + b (\log X) + c (\log X)^3 && \text{--- (14)} \\
\log Y &= a + b (\log X) && \text{--- (15)} \\
\log Y &= a + b (\log X) + c (\log X)^2 && \text{--- (16)} \\
\log Y &= a + b (\log X) + c (\log X)^2 + d (\log X)^3 && \text{--- (17)}
\end{aligned}$$

Since the variance of residuals appeared to be proportional to the value of X, the weight of " $W_i = \frac{1}{X_i}$ " was also introduced in the regression analysis. The regression was calculated (Regression 18) and subsequent residuals plotted.

$$Y = 3.201 + 1.596X \text{ --- (18)}$$

The best fitting among the above tested regressions was found to be regression 5:

$$\log Y = 0.702 + 0.0532X \text{ --- (5)}$$

Figure 2e shows the plotting of the residuals. Improvement by the transformation process is obviously negligible. Perhaps, errors were involved in the sampling process in this case. It is best to discard the sample completely and get one more representative of the population or increase the sample size.

(3) Descending band

By forcing a regression through the origin, the plotting of residuals against predicted values often exhibits a descending or ascending band (Figure 2c). By introducing a constant term in regression analysis, the plotting of residuals shows a perfect "horizontal band" with variation of residuals amounting only to ± 2 units. Hence the predicted values become more reliable.

(4) Arch band

The "arch band" scatter pattern is demonstrated by the plotting of residuals for regression 4. Several regression models were fitted. The best improved model was regression 7:

$$Y = 4.963X - 0.250X^2 + 0.0023X^3 - 3.21 \text{ --- (7)}$$

The extra square and cubic terms included in the regression analysis greatly improved the fitting of basic data to a regression. This is demonstrated by the plotting of residuals as shown in Figure 2g. An approximate "horizontal band" is obtained with variances amounting only to ± 3 units.

Part II Actual examples

It is difficult to find actual forestry data which show patterns similar to the 4 scatter patterns mentioned above. Three regressions (19, 20 and 21 in Table 2) selected from a simulated growth model (Lee, 1967) are used to illustrate "horizontal", "ascending or descending" and "wedge" bands. No example is available from the source quoted for illustration of the "arch band" scatter pattern shown in Figure 1d.

From regression 19, the residuals of crown width were calculated and plotted against the predicted crown width (Figure 3). It is noted that the residuals follow the scatter pattern of a very narrow "wedge band", which is almost "horizontal", with variation in residuals amounting to approximately ± 5 feet. No great abnormality is indicated.

From regression 20, the residuals of tree height were determined and plotted in Figure 4. An increasing scatter of residuals with tree height is demonstrated (Figure 4). The scatter pattern appears to follow that of Figure 1b and 1c, i.e., "wedge" and "ascending or descending" bands. The reliability of the predicted tree height is questionable, although the correlation coefficient is highly significant.

In order to improve regression 20, the use of weighted least square, extra square term, extra cross-product term, and/or the transformation on the observed value of Y_1 were introduced. New regressions were then calculated and subsequent residuals plotted. The best among these new regressions was found to be regression 22.

$$\text{Ht (ft.)} = 0.123 + 8.459 (\text{dbh. in.}) - 0.00826 (\text{dbh. in.})^3 \quad \text{--- (22)}$$

$$N = 203, R = 0.872, R^2 = 0.761, SE_E = 13.7 \text{ feet}$$

Figure 5 shows the plotting of residuals for this regression. From Figures 4 and 5, it is apparent that no improvement has been obtained.

From regression 21, the residuals of cubic volume per tree plotted against the predicted volume (Figure 6) show a very narrow "wedge band" which is almost "horizontal". This indicates that the predicted tree volume should be quite reliable. However, Figure 6 indicates a slight underestimate of volume for larger trees.

Discussion

The foregoing indicates how techniques for examining residuals can be used to assess the reliability of the predicted parameter. The following is a discussion of how the techniques can be used to improve the process of selecting best fitted regression for a set of data, in addition to the employment of such criteria as the correlation coefficient, coefficient of determination, standard error of estimate, analysis of variance, and confidence limits.

Table 2. Actual examples

$$\text{CW (ft.)} = 2.860 + 1.629 (\text{dbh, in.}) - - - - - (19)$$

$$N = 169 \quad r = 0.483^{**} \quad r^2 = 0.233 \quad \text{SE}_E = 2.8 \text{ ft.}$$

$$\text{Ht (ft.)} = 10.491 (\text{dbh, in.}) - 0.244 (\text{dbh, in.})^2 + 0.0384 (\text{BA, sq.ft.}) + 10.0 - - - - - (20)$$

$$N = 203 \quad R = 0.877^{**} \quad R^2 = 0.769 \quad \text{SE}_E = 13.5 \text{ ft.}$$

$$V (\text{inside bark, cu.ft.})/\text{BA (outside bark, sq.ft.)} = 0.441 (\text{Ht.ft.}) - - - - - (21)$$

$$N = 456 \quad r = 0.939^{**} \quad r^2 = 0.882 \quad \text{SE}_E = 2.9 \text{ ft.}$$

** = Significant at 1% level

FIGURE 3. Residuals of crown width plotted against predicted crown width.

$$CW \text{ (ft.)} = 2.860 + 1.629(\text{dbh, in}) \dots\dots\dots (19)$$

$N = 169$, $r = 0.483^{**}$, $r^2 = 0.233$, $SE_E = 2.8 \text{ ft.}$

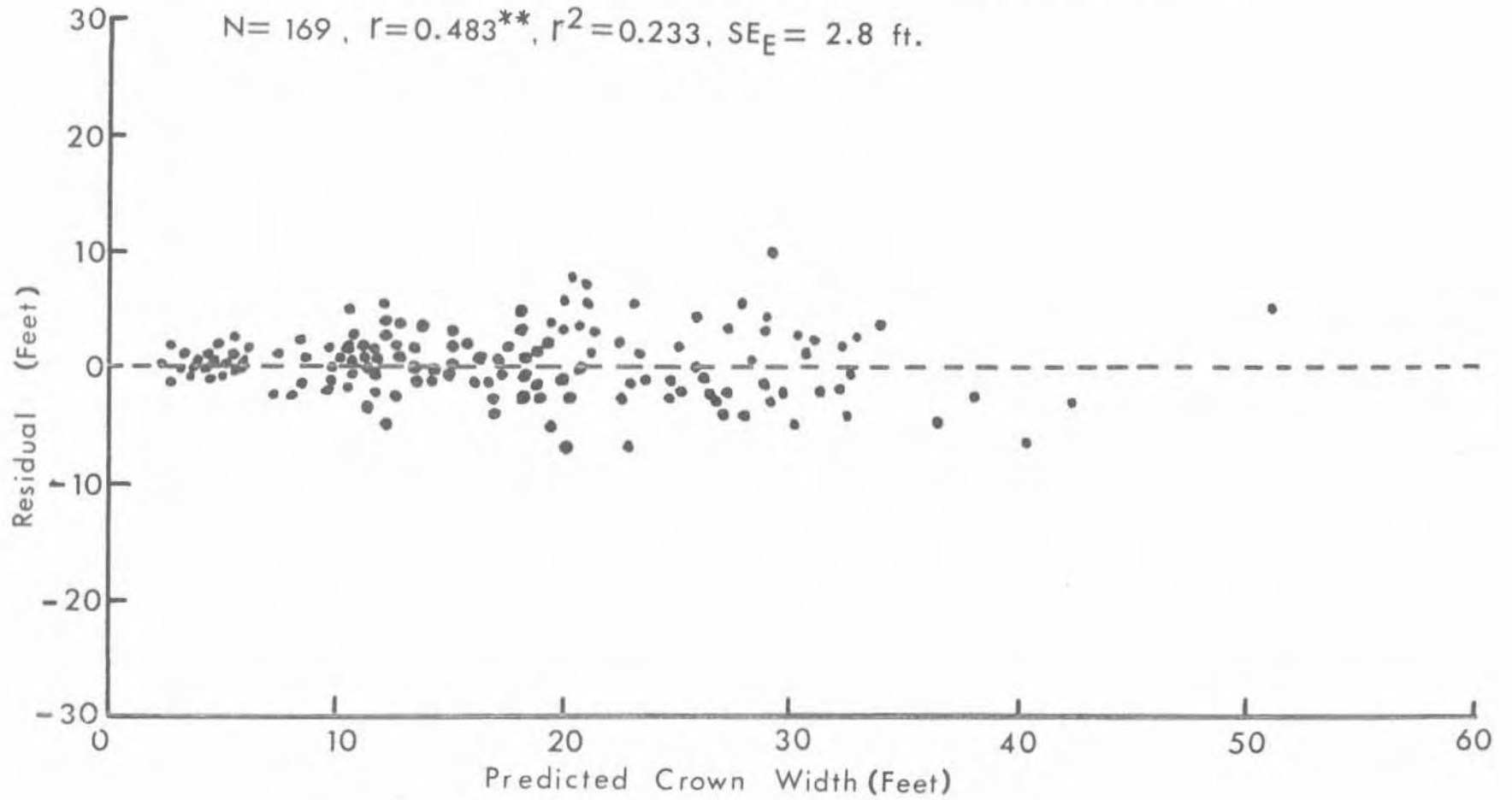


FIGURE 4. Residuals of height plotted against predicted height.

$$\text{Ht. (ft.)} = 10.491(\text{dbh, in.}) - 0.244(\text{dbh, in.})^2 + 0.0384(\text{BA, sq. ft.}) - 10.0 \dots (20)$$

N = 203, R = 0.877, R² = 0.769, SE_E = 13.5 ft.

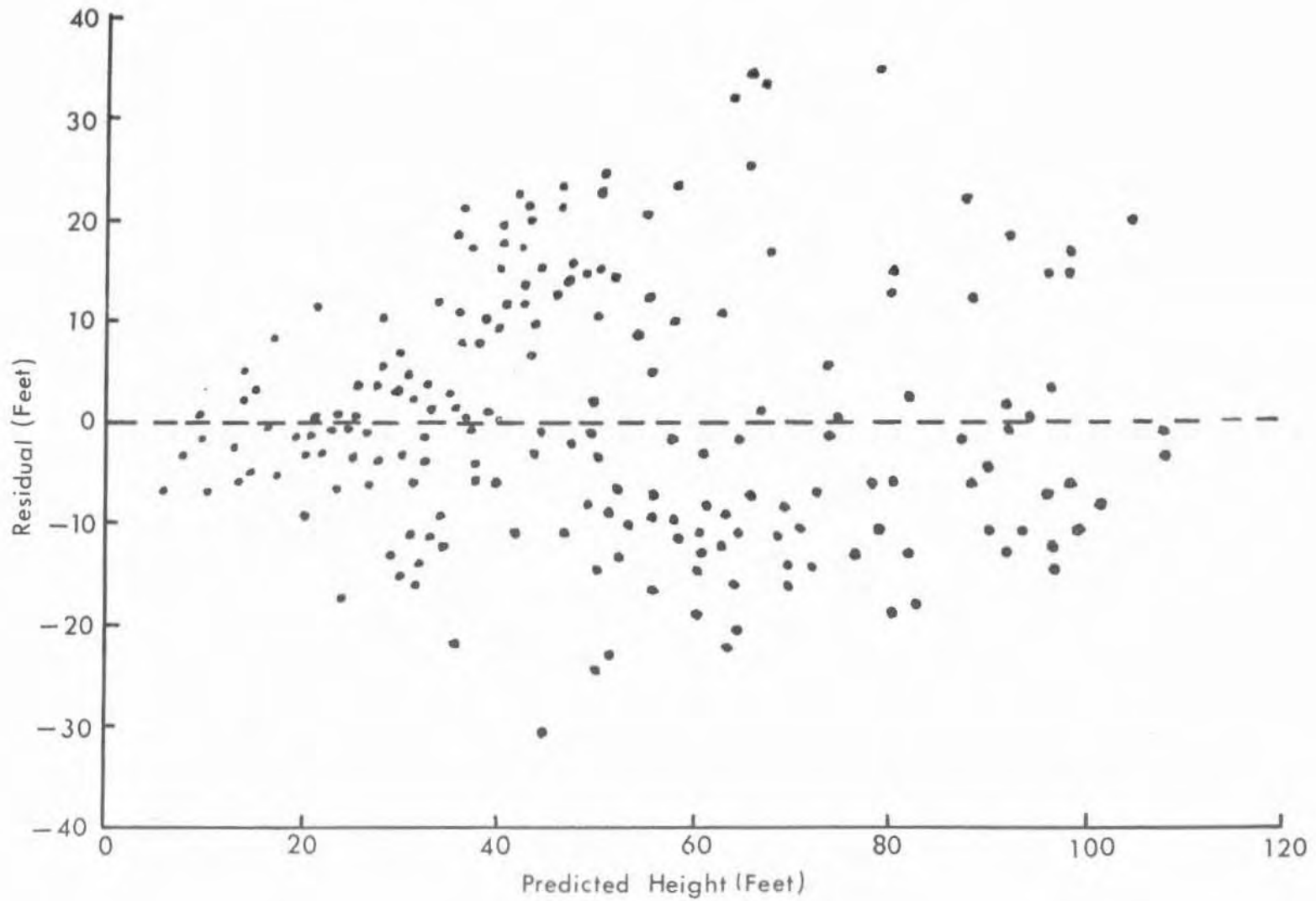


FIGURE 5. Residuals of height plotted against predicted height for regression 22.

$$\text{Ht. (ft.)} = 0.123 + 8.459 (\text{dbh, in.}) - 0.00826 (\text{dbh, in.})^2 \dots (22)$$

$N = 203$, $r = 0.872^{**}$, $r^2 = 0.761$, $SE_E = 13.7$ ft.

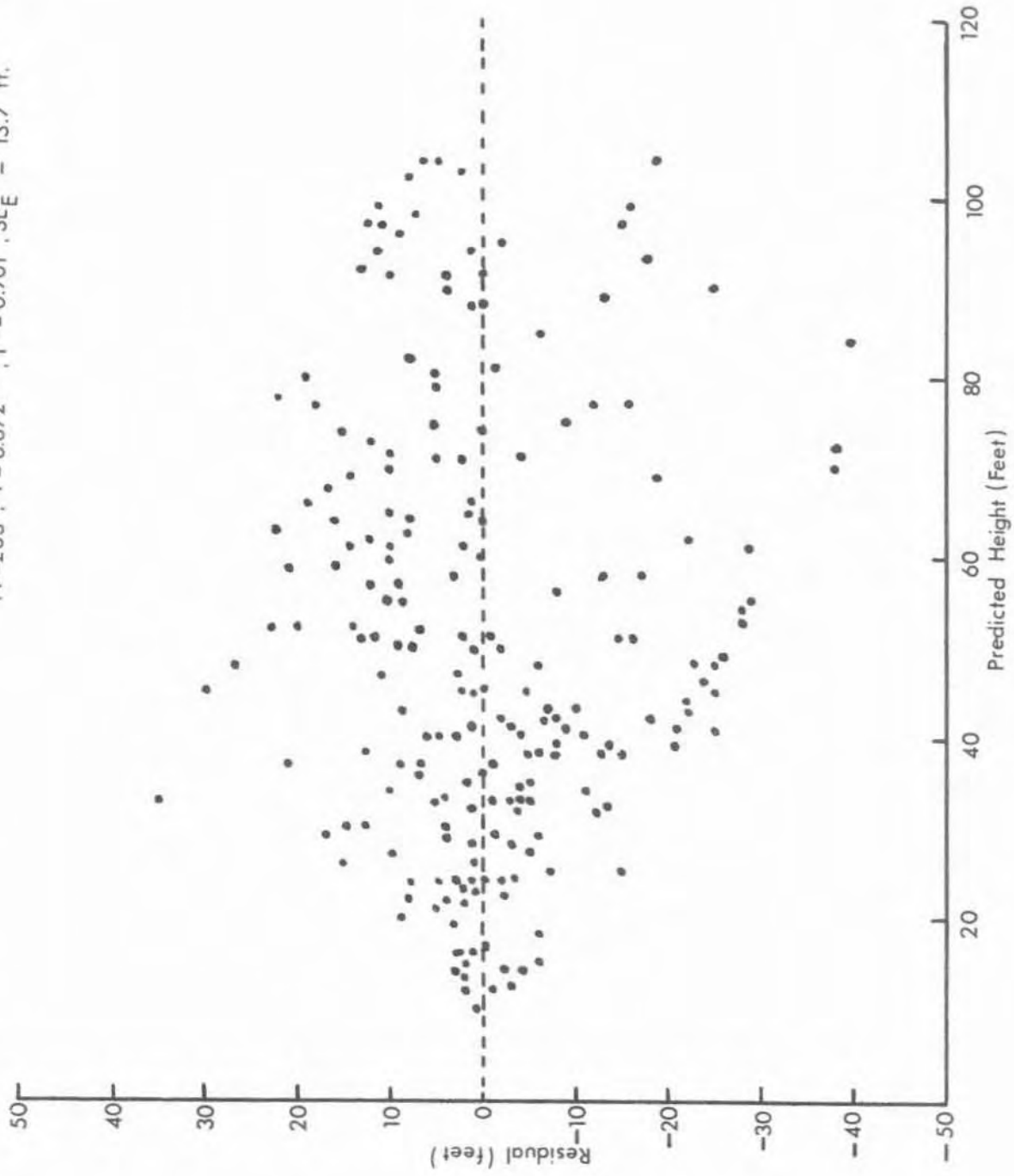
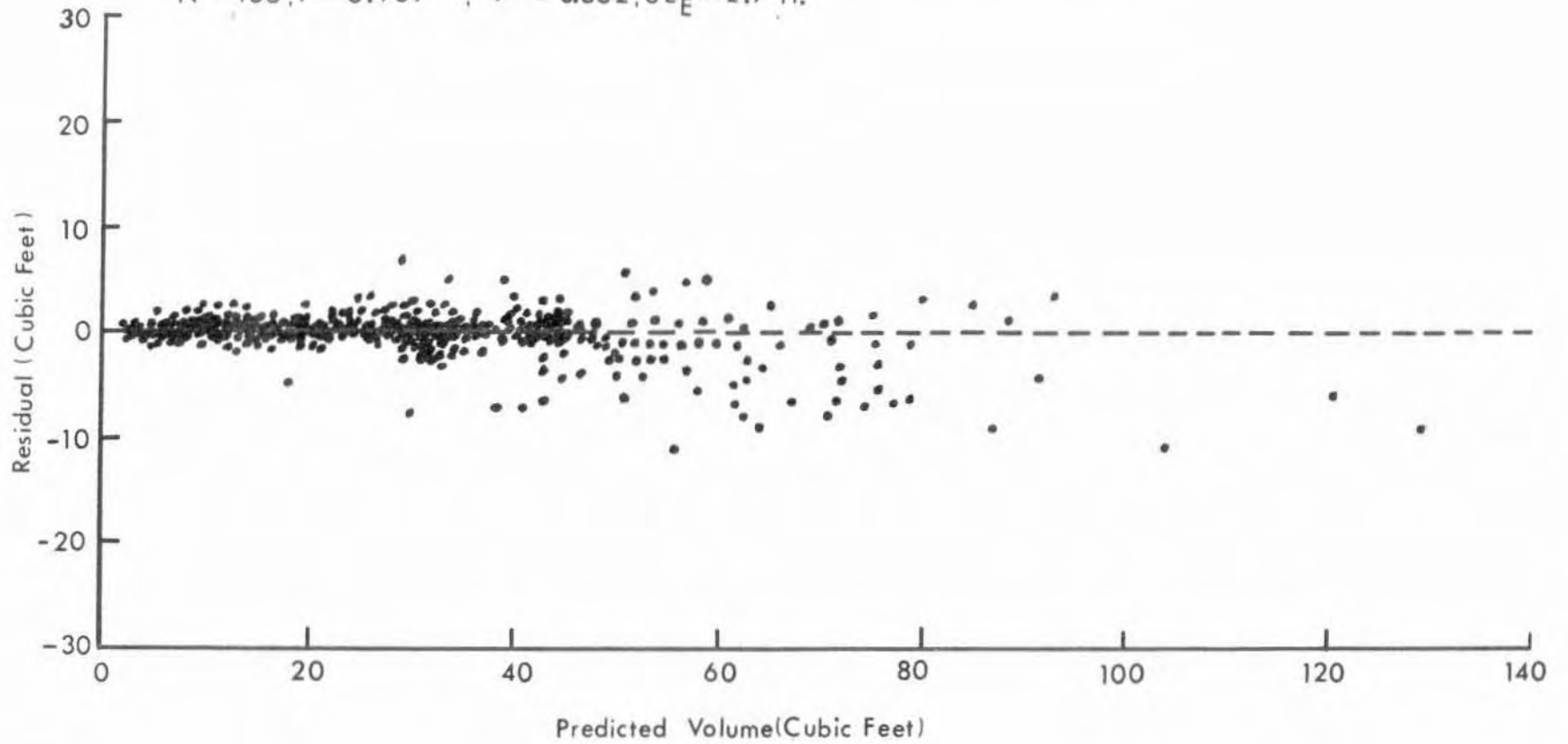


FIGURE 6. Residuals of volume per tree plotted against predicted volume.

$$V (\text{inside bark, cu. ft.}) / BA (\text{outside bark, sq. ft.}) = 0.441 (\text{Ht. ft.}) \dots\dots\dots (21)$$

$N = 456, r = 0.939^{xx}, r^2 = 0.882, SE_E = 2.9 \text{ ft.}$



(1) Correlation coefficient (r or R)

Early literature reported correlation coefficients being used as guides in the selection of independent variables to be fitted in regression analysis. In recent years, however, the role of the correlation coefficient has considerably diminished (Freese, 1964). A regression with a highly significant correlation coefficient may not necessarily be a desirable regression. The following pairs of regressions taken from Table 1 illustrate this point.

(a) Regressions 2 and 5 in Table 1

Both regressions are derived from the same data with different regression models. The correlation coefficients for both regressions are highly significant (at 1% level). However, the plotting of residuals for both regressions (Figure 2b and 2e) indicates that the variance of residuals is not constant. Therefore, the predicted values may be less reliable.

(b) Regressions 4 and 7 in Table 1

Regression 4 has a correlation coefficient of 0.877 which is significant at the 1% level. Regression 7 has a correlation coefficient of 0.971 which is also highly significant. Without the plotting of residuals (Figure 2d and 2g), it is not known how much more accurate the resultant prediction is by choosing regression 7 over regression 4. Obviously, Figure 2d and 2g show that regression 4 is not a desirable regression at all, because its scatter pattern of residuals follows the pattern of "arch band".

(c) Regressions 3 and 6 in Table 1

Both of these regressions are derived from the same data and have the same highly significant correlation coefficient of 0.988. But which one of the two regressions should be chosen? The calculation of \bar{Y} is slightly easier with regression 3 than with regression 6, while the latter has a slightly smaller standard error of estimate (1.15) as compared to that of regression 3 (1.31).

By use of the technique of examining residuals, once the residuals are plotted against the predicted values for both regressions (Figure 2c and 2f), the following results immediately become apparent. Departure from regression 3 is such that positive residuals correspond to low predicted values, and negative residuals correspond to high predicted values (Figure 2c). On the other hand, the plotting of residuals for regression 6 shows a "horizontal band" with variation amounting to less than ± 2 units. Obviously, regression 6 is the desirable regression. The usefulness of the technique of examining residuals is apparent.

(2) Coefficient of determination (r^2 or R^2)

The coefficient of determination is another commonly used measure of how well a regression fits a set of data. It represents, in the dependent

variable Y, the proportion of variation that is associated with the regression in the independent variable or variables. It is merely the squared value of the correlation coefficient discussed in (1) above.

(3) Standard error of estimate (SE_E)

The standard error of estimate is the measure of dispersion of the original observations around the model. The question here is should one regression be chosen over other regressions because of its relatively small standard error of estimate? In most cases, the answer is yes. But how much smaller should the standard error of estimate be? Undoubtedly one can find out how much smaller the standard error of estimate should be by working out a great many sets of standard errors of estimate for each set of data which fits the many sets of regression models, and then compare the standard errors of estimate. But one must ask whether the time spent in working out all these various sets of standard errors of estimate is necessary and worthwhile?

The plotting of residuals would help to make the choice. Refer to regressions 3 and 6 again. The difference of 0.16 in standard error of estimate between these 2 regressions may not be significant, but the plottings of residuals certainly indicate that regression 6 (Figure 2f) is a more reliable predictor than regression 3 (Figure 2c). The variability of residuals for regression 6 is certainly much less than that of regression 3.

(4) Analysis of variance

"Analysis of variance" in regression analysis is a useful tool in testing the significant difference between a maximum model and a hypothetical model (Freese, 1964). However, it does not show pictorially the differences between the observed and predicted values.

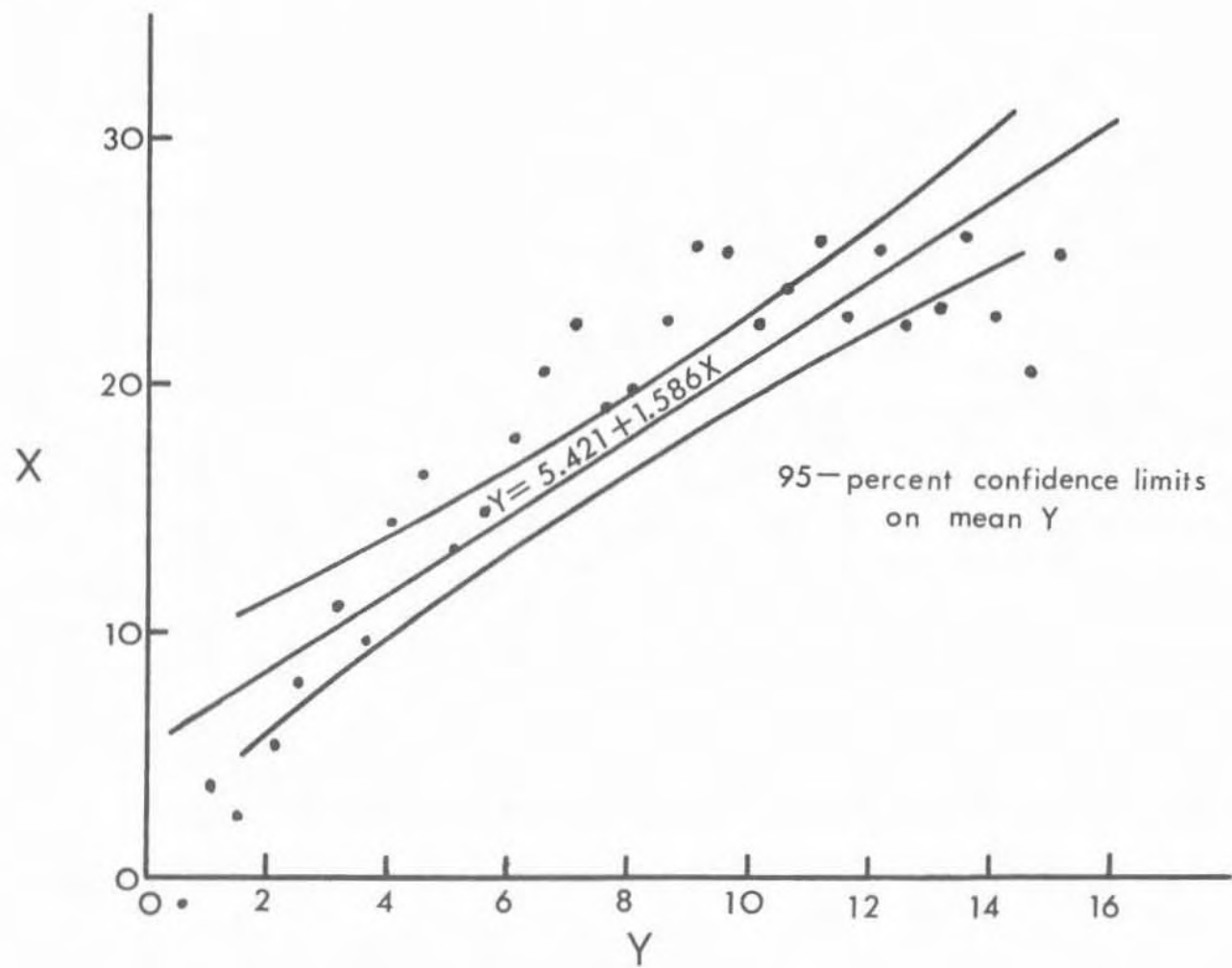
(5) Confidence limits

A point estimate of a parameter is not very meaningful without some measure of the possible error in the estimate. For example, if we compute 95-percent confidence limits for the mean, these limits will include the population mean unless a 1-in-20 chance has occurred in the sampling process.

Confidence limits can be computed for regression. But they do not show the abnormal variability of the residuals. For example, the confidence limits for regression 4 (in Table 1) have been calculated and plotted in Figure 7. It appears that no abnormality (without the plotting of basic data) is indicated in Figure 7. However, the plotting of basic data superimposed on the lines of regression and confidence limits indicates that the regression is inadequate.

By use of the technique of examining residuals, a clearer picture emerges. The plotting of residuals for the same regression 4 indicates an "arch band" which means that the regression model is inadequate, and little reliability can be placed upon the predicted values. In addition, the

FIGURE 7. Regression 4 with 95-percent confidence limits.



plotting of confidence limits for a multiple regression in a graph is not possible (involving multiple regression surfaces), whereas the plotting of residuals for any regression, simple or multiple, is always possible.

Conclusion

It is demonstrated that in addition to the criteria generally utilized in the selection of regressions, the technique of examining residuals between observed and predicted values can be used effectively to indicate the degree of reliability of the predicted values. By providing scatter patterns of residuals, the technique can also assist the investigator to obtain the most significant independent variables in regression analysis in evaluating his data. Furthermore, the degree of reliability of a regression can be more easily understood by the non-statistically-orientated forester if residuals are described pictorially.

Acknowledgement

The author is grateful to Mr. M. Magar, Systems Analyst, for his suggestions and assistance in computing the hypothetical data. The paper was reviewed by Drs. S.W. Nash, Professor of Mathematics, A. Kozak, Professor of Forestry, University of British Columbia; Drs. D.M. Brown, Head, Biometrics Research Services and A.L. Wilson, Statistician, Dept. of Fisheries and Forestry, Ottawa. Their comments are appreciated and were useful to the author.

Literature Cited

- Draper, N.R. and H. Smith. 1966. Applied regression analysis. John Wiley and Sons, Inc. N.Y. 407 pp.
- Freese, F. 1964. Linear regression methods for forest research. U.S.F.S. Res. Paper FPL 17. 136 pp.
- Lee, Y. 1967. Stand models for lodgepole pine and limits to their application. Fac. of For., U.B.C., Ph.D. thesis. 332 pp. Litho.
- Spurr, S.H. 1952. Forest inventory. The Ronald Press Co., N.Y. 476 pp.
- Walters, J. 1967. U.B.C. Research Forest annual report. Fac. of For., U.B.C. 26 pp.