NEW USEFUL IS THE COEFFICIENT OF
DETERMINATION IN MULTIPLE LINEAR REGRESSION?

by
Gary W. Fowler and Fred N. Bigelow

ABSTRACT: The relationship of the coefficient of determination ($R^2$) to multiple linear regression and some problems associated with its use as a measure of the usefulness, "goodness-of-fit," or predictive precision of a regression equation are discussed. The adjusted coefficient of determination ($R^2_{adj}$) is evaluated and compared to $R^2$. An example using simple linear regression is examined in detail. A set of criteria is presented for evaluating regression equations.

INTRODUCTION

Model I: multiple linear regression is widely used by practitioners in natural resources. The general linear model is

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i \quad (i = 1, \ldots, n) \]

where $Y_i$ = observed value of the dependent or response variable for the $i$th element

$X_{ij}$ = observed value of the $j$th ($j = 1, \ldots, p$) independent or predictor variable for the $i$th element

$p$ = regression constant or $Y$-intercept of the regression equation

$\beta_j$ = regression (not adjusted or partial) coefficient associated with the $j$th ($ j = 1, \ldots, p$) independent variable

$e_i$ = random error term or that part of $Y_i$ not explained by the $X_{ij}$'s

$p$ = number of independent variables

$a$ = number of elements in the sample

The model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$, and is not restricted to a functional linear relationship between $Y$ and the $X_{ij}$'s (i.e., the form

$R^2$ The authors are, respectively, Associate Professor of Biometrics and Graduate Teaching Assistant, School of Natural Resources, The University of Michigan, Ann Arbor, Michigan, 48109.

Published by: USDN, Bureau of Land Management, D 460 Denver Service Center, Denver Federal Center, Bldg. 50 Denver, Colorado, 80225
of the dependence of $Y$ on the $X_i$'s can be curvilinear. The parameters $b_1, b_2, \ldots, b_p$ are unknown and are usually estimated using least squares based on a random sample of a element (comms. or observations) from the population of interest. The least squares sample regression equation is

\[ Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi} + \varepsilon_i \quad (i = 1, \ldots, n) \]

where $Y_i$ is the predicted value of $Y$ for the $i$th element, and $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the Least Squares estimates of the respective population parameters.

The difference between the observed and predicted value of the dependent variable, $e_i = Y_i - \hat{Y}_i$, is the residual or error term associated with the $i$th element.

The assumptions associated with Model 2: multiple linear regression are:

1. Independent observations (the $e_i$'s are independent)
2. Linearity in the parameters
3. Homogeneity of variances (the $e_i$'s are homogeneously distributed around the regression surface)
4. $X$'s are fixed constants and measured without error
5. The $Y_i$'s for $X_i$'s are normally distributed—this assumption is necessary to make statistical inferences.

Multiple linear regression can be applied to help understand causal relationships such as the effect of fertilizers on the growth of tomatoes on a farm. Another application is in modeling situations where a variable that is difficult and expensive to measure is predicted from variables that are easier and less expensive to measure. For example, tree volume can be predicted from tree height and diameter measurements.

Many practitioners in natural resources use the coefficients of determination ($R^2$) as a measure of the usefulness, "goodness-of-fit" (i.e., how close the data points are fit by the regression equation), of the predictive precision (i.e., how well $Y$ predicts $Y$) of the regression equation. $R^2$ is also used as a criterion to help search for the "best" of all possible regression equations.

In the multiple linear regression model, the total sum of squares of the dependent variable is partitioned as follows:

\[ SS_T = SS_R + SS_E \]

where $SS_T$ = total sum of squares

$SS_R$ = sum of squares explained by the regression equation

$SS_E$ = sum of squares not explained by the regression equation (error sum of squares)

\[ R^2 = \frac{SS_R}{SS_T} \]

$R^2$ is the proportion of the total sum of squares explained by the Regression equation.
The multiple correlation coefficient \( R \) is also the simple linear correlation between \( Y_i \) and \( Y \) if \( i = 1, \ldots, n \). \( R^2 \) is more commonly used in multiple linear regression.

The objectives of this paper are to (1) examine the relationship of \( R^2 \) to the multiple linear regression model, (2) present some problems associated with using \( R^2 \) to evaluate a regression equation, (3) evaluate the adjusted coefficient of determination and compare it with \( R^2 \), and (4) present a set of criteria for the natural resource practitioner to use in evaluating regression equations.

**EXAMPLE**

An example using simple linear regression will be used throughout this paper. The model is

\[
Y_i = \beta_0 + \beta_1 X_i + e_i \quad (i = 1, \ldots, n)
\]

where \( \beta_0 \) is the intercept and \( \beta_1 \) is the slope of the population regression line.

The sample least squares regression line is

\[
Y = \hat{\beta}_0 + \hat{\beta}_1 X
\]

The data in Table 1 represent \( n = 20 \) pairs of observations taken on a dependent variable \( Y \) and an independent variable \( X \). We suppose there is a linear relationship between \( Y \) and \( X \) and that all of the assumptions of the regression model are adequately met.

<table>
<thead>
<tr>
<th>Observation</th>
<th>( X )</th>
<th>( Y )</th>
<th>Observation</th>
<th>( X )</th>
<th>( Y )</th>
<th>Observation</th>
<th>( X )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>6</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>11</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>7</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>13</td>
<td>7</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

| Relationship of \( R^2 \) to Multiple Linear Regression

In multiple linear regression, a

\[
R^2 = \frac{(\bar{Y} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}
\]

\( \bar{Y} \)
$$\text{SSR} = \sum \left[ y_i - \bar{y} \right] \left[ y_i - \hat{y}_i \right]$$

SSR = \( \sum_{i=1}^{n} (y_i - \bar{y}) (y_i - \hat{y}_i) \)

$$\text{SSE} = (1 - \hat{r}^2) \sum y_i^2$$

SSE = \( \sum (y_i - \hat{y}_i)^2 \)

where \( \bar{y} = \frac{1}{n} \sum y_i \) and \( \bar{y} = \frac{n}{n} \sum y_i/n \). The calculated F-statistic

(7) \( F_{p,n-p-1} = \frac{\text{SSE}/p}{\text{SSE}/(n-p-1)} \)

where SSR = SSE/p and MSE = SSE/(n-p-1), is used to test the null hypothesis

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad (\text{no significant regression equation}) \]

against the alternative hypothesis

\[ H_1 : \text{not all } \beta_i \text{ values (significant regression equation)} \]

The decision rule to test \( H_0 \) is:

Reject \( H_0 \) if \( F_{p,n-p-1} > F_{p,n-p-1} \); otherwise accept \( H_0 \).

where \( F_{p,n-p-1} \) is the upper critical value of the F-distribution with \( p \) and \( n-p-1 \) degrees of freedom, and \( n \) is the level of significance.

The test statistic

(8) \( F_{p,n-p-1} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \)

is used to test the significance of the improvement of a regression model by adding \( g \) new independent variables to a model already containing \( h \) independent variables. The full model is based on \( p = g + h \) independent variables. \( R^2_p \) is the coefficient of determination for the original model with \( h \) independent variables, and \( R^2_{p+g} \) is the coefficient of determination for the full model with \( p \) independent variables.

Thus, the F-statistic for testing the significance of a regression model and the improvement of a model by the addition of new independent variables are both directly related to \( R^2 \).

In simple linear regression, \( p = 1 \), \( R^2 = r^2 \), and \( R = |r| \), where \( r \) is the simple linear correlation coefficient. For our example (Table 1),

<table>
<thead>
<tr>
<th>( i )</th>
<th>( y_i )</th>
<th>( \hat{y}_i )</th>
<th>( y_i - \hat{y}_i )</th>
<th>( y_i - \bar{y} )</th>
<th>( y_i - \bar{y} ) ( y_i - \hat{y}_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>21</td>
<td>0</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>

\( \hat{y} = \bar{y} - \bar{y} = 0.95 \)
\[ Y \hat{X} = 0.95 = 0.42X \]

\[ r = \frac{20}{20} - \frac{20}{20} \cdot \frac{20}{20} \cdot 0.85 \]

\[ s^2 \cdot r^2 = 0.721 \]

\[ SST = \frac{20}{1} - \frac{20}{20} = 370.53 \]

\[ SSS = \frac{20}{1} - \frac{20}{1} = 411.53 \]

\[ SSE = 20 - \frac{20}{1} = 159.02 \]

\[ \bar{r} = 46.2 \]

where \( n = x \) and \( \bar{x} = \frac{\sum x}{n} \).

The 20 data pairs and the sample regression line (lower line) for our example (Table 1) are plotted in Figure 1.

**Figure 1.** Simple linear regression line having slope angles of 45.6° and 58.8° with the new "goodness-of-fit" as the example based on 20 observations.
The analysis of variance (ANOVA) table for testing the significance of the regression line is given in Table 2.

Table 2. Analysis of variance (ANOVA) table for testing the significance of the simple linear regression equation for our example (Table 1).

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>Calculated F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>431.53</td>
<td>1</td>
<td>431.53</td>
<td>46.58</td>
</tr>
<tr>
<td>Error</td>
<td>136.53</td>
<td>18</td>
<td>8.63</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>568.06</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F(0.05;1,18) = 4.41 for a = 0.05. Thus, there is a significant linear relationship between Y and X. The significance probability associated with the calculated F value is F = 0.005. Because of the relatively high value of $R^2$ and the high significance of the regression equation, many users would say that they have a good regression equation.

**Problems Associated with $R^2$**

$R^2$ not only measures the goodness-of-fit but also the steepness of the regression surface (Barrett 1975). If the goodness-of-fit (RSE) of the regression surface remains constant, $R^2$ increases as the slope of the regression surface increases (9). When the slope of the regression surface increases,

$$n \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \text{ increases, which causes } R^2 = 1 - \frac{SSR}{SST} \text{ to increase. In other words, the simple linear correlation (R) between } Y_i \text{ and } \bar{Y} \text{ increases as the slope of the regression surface increases.}

The slope angle of the simple linear regression line for our example with $R^2 = 0.723$ is 42.6° (Figure 1). A series of different regression lines were constructed with the same "goodness-of-fit" (the set of vertical distances from the data points to the regression line is the same for all lines) as our example but with different slopes and a constant intercept. 87°, 45°, 0°, and the calculated F value (F(0.01;1,18)) for these different regression lines are shown in Table 3 where

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad \text{variance of } Y \text{ about } \bar{Y}$$

$$s_y = \sqrt{\frac{\text{SST}}{n-1}} \quad \text{standard deviation of } Y$$

$R^2$ increases as the slope becomes steeper in either a positive or negative sense. In other words, $R^2$ increases as the absolute value of the slope increases.
Table 3. $\hat{\beta}_1$, $\hat{\eta}_1$, $\hat{\psi}_1$, $\hat{\varphi}_1$, and $\hat{F}_{11,18}$ of the regression lines with equal intercepts, different slopes, and the same "goodness-of-fit" as found in our example (Table 1).

<table>
<thead>
<tr>
<th>Slope angle in degrees ($\theta$)</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\eta}_1$</th>
<th>$\hat{\psi}_1$</th>
<th>$\hat{\varphi}_1$</th>
<th>$\hat{F}_{11,18}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.0005</td>
<td>0.0055</td>
<td>0.17</td>
<td>0.06</td>
<td>0.004</td>
</tr>
<tr>
<td>10.8</td>
<td>0.0405</td>
<td>0.0400</td>
<td>0.29</td>
<td>0.02</td>
<td>1.87</td>
</tr>
<tr>
<td>20.6</td>
<td>0.0330</td>
<td>0.0327</td>
<td>0.22</td>
<td>0.02</td>
<td>1.65</td>
</tr>
<tr>
<td>30.6</td>
<td>0.0220</td>
<td>0.0206</td>
<td>0.20</td>
<td>0.02</td>
<td>1.45</td>
</tr>
<tr>
<td>40.6</td>
<td>0.0130</td>
<td>0.0106</td>
<td>0.20</td>
<td>0.02</td>
<td>1.45</td>
</tr>
<tr>
<td>50.6</td>
<td>0.0090</td>
<td>0.0067</td>
<td>0.20</td>
<td>0.02</td>
<td>1.45</td>
</tr>
<tr>
<td>60.7</td>
<td>0.0066</td>
<td>0.0044</td>
<td>0.20</td>
<td>0.02</td>
<td>1.45</td>
</tr>
<tr>
<td>70.7</td>
<td>0.0050</td>
<td>0.0035</td>
<td>0.20</td>
<td>0.02</td>
<td>1.45</td>
</tr>
</tbody>
</table>

The data points and sample regression lines having the same goodness-of-fit for slopes of 45.6° (our original example) and 30.6° are shown in Figure 1.

$R^2$ is a function of the slope angle ($\theta$ in degrees) and varies from 0.0003 for a slope of 0.6° to 0.994 for a slope of 30.6° (Table 3, Figure 2).

Figure 2. $R^2$ as a function of $\theta$ in degrees for simple linear regression lines having the same "goodness-of-fit" as our example (Table 1).

For simple linear regression,

$$R^2 = \frac{\sum (y_1 - \bar{y})^2}{\sum (y_1 - \bar{y})^2} = \frac{\sum (y_1 - \hat{y})^2}{\sum (y_1 - \bar{y})^2}$$

2
Table 3 also shows that $\alpha_0$ MNE, and $\beta_1$ MRE increase as $\alpha$ increases. The results of Table 3 show consistently that given the same goodness-of-fit, $R^2$ and the significance of the regression equation increase as the slope of the regression line increases.

When evaluating a multiple regression equation, computing different regression equations from different sets of data, or comparing different regression equations from the same set of data, causal relationships or predictive relationships cannot be determined solely by looking at $R^2$ and the significance of the regression equation. A regression equation with a higher $R^2$ could have lower predictive power than a regression equation with a lower $R^2$ if the higher $R^2$ is associated with a smaller $R^2$.

An examination of equation (11) for the calculated $R^2$ value shows that the number of observations must be considered when evaluating the significance of the regression equation. For $E$ constant, the $F$ value and the significance of the regression equation increase as $n$ increases. This creates a problem for the practitioner. The problem of large sample sizes is that a very small $R^2$ will yield an insignificant regression equation if the sample size is large enough. The problem of large sample sizes is that a very large $R^2$ will yield an insignificant regression equation if the sample size is small enough. Another example associated with a small sample size is that a regression equation can be developed with a high $R^2$ that fits the data poorly but has very little in common with the population from which the data points were drawn.

The practitioner should also be aware that $R^2$ is a biased estimate of the population coefficient of determination $r^2$ (Cochran 1933, Kendall and Stuart 1967). When there is no relationship between $Y$ and the $X_j$'s in the population ($\beta_j = 0$), the expectation of $R^2$ is

$$E(R^2) = E((\beta_0 + \beta_1X_1 + \cdots + \beta_pX_p)^2)$$

Therefore, the bias is always positive and equal to $p/(n-1)$. $E(R^2)$ is the mean value of the $R^2$'s calculated from all possible samples of size $n$ from the population assuming the $X_j$'s are normally distributed. $E(R^2)$ for various values of $n$ and $p$ is shown in Table 4. These are exact values.

---
Table 4. $E(k^2)$ for various values of $n$ and $p$ when $n^c = 0$.

<table>
<thead>
<tr>
<th>n</th>
<th>0.100</th>
<th>0.500</th>
<th>1.000</th>
<th>2.000</th>
<th>3.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.100</td>
<td>0.500</td>
<td>1.000</td>
<td>2.000</td>
<td>3.000</td>
</tr>
<tr>
<td>1</td>
<td>0.111</td>
<td>0.222</td>
<td>0.444</td>
<td>0.666</td>
<td>0.888</td>
</tr>
<tr>
<td>2</td>
<td>0.055</td>
<td>0.105</td>
<td>0.211</td>
<td>0.311</td>
<td>0.411</td>
</tr>
<tr>
<td>3</td>
<td>0.030</td>
<td>0.061</td>
<td>0.182</td>
<td>0.282</td>
<td>0.382</td>
</tr>
<tr>
<td>4</td>
<td>0.010</td>
<td>0.020</td>
<td>0.060</td>
<td>0.100</td>
<td>0.140</td>
</tr>
<tr>
<td>5</td>
<td>0.005</td>
<td>0.010</td>
<td>0.030</td>
<td>0.060</td>
<td>0.090</td>
</tr>
</tbody>
</table>

$E(k^2)$ increases as $n^c$ increases as $n$ increases for a given value of $n$ and decreases as $n$ increases for a given value of $n^c$. For $p = 1$ must be greater than 0 to have at least one degree of freedom for the error term in the linear regression.

Even if there is no relationship between $Y$ and the $X_i^c$'s in the population, a very high $k^2$ could be obtained easily by chance (very close to 1), especially when the number of independent variables $p$ approaches the number of observations $n$.

$k^2$ is also a biased estimate of $E(k^2)$ when $n^c = 0$. $E(k^2)$ in this case was shown by Wishart (1932) to be

$$E(k^2) = E(k^2|n^c = 0) = 1 - \frac{p}{n} + \frac{p}{n} \left(1 - n^c\right)$$

where $F(1, 1, \frac{1}{2}, n^c)$ is the hypergeometric function with parameters $1, 1, 1, \frac{1}{2}, n^c$, and $X_i^c$. Equation (11) is applicable to the multivariate situation (Goodall and Stuart 1967) where both $Y$ and the $X_i^c$'s are random variables and is only approximate for Model I multiple linear regression where the $X_i^c$'s are not random variables. The approximation becomes better as $n$ increases. Since equation (11) is relatively difficult to solve, the following approximation based on the first two terms of the hypergeometric series is usually used to calculate $E(k^2)$:

$$E(k^2) = 1 - \frac{p}{n} + \frac{p}{n} \left(1 - n^c\right) - \frac{p}{n} \frac{1}{2} \left(1 - n^c\right)$$

This approximation is accurate to the order $n^{-1}$.

The bias associated with $k^2$, in general, decreases as $n$ increases for a given value of $n^c$ and as $n^c$ increases for a given value of $n$. The bias increases as $p$ increases for $n^c$ and $n$ constant. $E(k^2)$ is given for various values of $n$ and $n^c$ for simple linear regression ($p = 1$) in Table 5.
Table 5. $E(Y)$ for various values of $\gamma_2$ and $n$ for simple linear regression. Values in parentheses for $n = 3$ are exact. All other values are approximations of $E(Y|\theta)$. 

<table>
<thead>
<tr>
<th>$n$</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.513</td>
<td>0.578</td>
<td>0.688</td>
<td>0.828</td>
<td>0.961</td>
</tr>
<tr>
<td>3</td>
<td>0.278</td>
<td>0.391</td>
<td>0.562</td>
<td>0.760</td>
<td>0.951</td>
</tr>
<tr>
<td>10</td>
<td>0.148</td>
<td>0.303</td>
<td>0.515</td>
<td>0.741</td>
<td>0.948</td>
</tr>
<tr>
<td>20</td>
<td>0.096</td>
<td>0.273</td>
<td>0.504</td>
<td>0.746</td>
<td>0.948</td>
</tr>
<tr>
<td>50</td>
<td>0.056</td>
<td>0.258</td>
<td>0.501</td>
<td>0.744</td>
<td>0.949</td>
</tr>
<tr>
<td>100</td>
<td>0.059</td>
<td>0.254</td>
<td>0.500</td>
<td>0.749</td>
<td>0.950</td>
</tr>
<tr>
<td>200</td>
<td>0.054</td>
<td>0.252</td>
<td>0.500</td>
<td>0.749</td>
<td>0.950</td>
</tr>
<tr>
<td>1000</td>
<td>0.051</td>
<td>0.250</td>
<td>0.500</td>
<td>0.750</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Exact values for $E(Y|\theta)$ based on equation (11) are given for $n = 1$ while all other values of $E(Y|\theta)$ are approximations based on equation (12). The bias is positive for $\gamma_2 = 0.50$, but becomes negative for large values of $\gamma_2$.

In using $E(Y|\theta)$ to evaluate a regression equation, the practitioner should remember that $E(Y|\theta)$ is biased for both the cases $\gamma_2 = 0$ and $\gamma_2 = 0$, and that the bias can be quite large for smaller sample sizes.

**THE ADJUSTED COEFFICIENT OF DETERMINATION**

$R^2 = 1 - \frac{SSE/df}{SS/df}$ is the unadjusted coefficient of determination in that it does not account for the difference in degrees of freedom associated with SSE and SS. The adjusted coefficient of determination ($R^2_{adj}$) does account for the difference in degrees of freedom where

$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SS/(p+1)}$ and $\hat{\sigma}^2 = SSE/(n-p-1)$ is the variance of $T$ about the regression equation.

$R^2_{adj}$ is the proportion of $\hat{\sigma}^2$ (or variability around $\bar{T}$) removed by the regression equation. $R^2_{adj}$ is always smaller than $R^2$ with this difference decreasing as $n$ and $p$ increase and $p$ decreases. One nice characteristic is that $E[R^2_{adj}] = 0$ when $\gamma_2 = 0$, $R^2_{adj}$ is more useful to the practitioner than $R^2$ in describing the size of the prediction error associated with a regression equation.

For examining the size of the prediction error, it is probably more meaningful to compare the standard deviation of $\overline{T(x_0)}$ with the standard error of the estimate $(\hat{\gamma}_2, \hat{\gamma}_1)$ as follows:

- 40-
The proportion 1 - \( \frac{\hat{y}}{\hat{y}} + \hat{y} \) is in terms of the units of the variable to be predicted.

The practitioner should be cautioned that \( \hat{y} \) can be negative when \( \beta^2 \) is small and/or \( \hat{y} \) is large compared to \( \hat{y} \). This means that \( \hat{y}^2 \) is greater than \( \hat{y}^2 \), which is caused by the fact that more has been lost in reduced degrees of freedom going from \( \hat{y}^2 \) to \( \hat{y}^2 \) than has been gained by reducing the sum of squares from \( \hat{y}^2 \) to \( \hat{y}^2 \). In other words, \( \hat{y} \) is a better predictor than the regression equation.

Setting equation (13) equal to zero and solving for \( \hat{y} \) (this value of \( \hat{y} \) will be called \( \hat{y} \))

\[
\hat{y} = \bar{y} \pm \frac{\hat{y} \cdot \hat{y}}{\hat{y}} \cdot \hat{y}
\]

For a given value of \( \hat{y} \), a regression equation based on a sample size less than \( \hat{y} \) will have a negative \( \hat{y} \). For example, if \( \hat{y} = 5 \) and \( \hat{y} = 0.3 \), a sample size of 10 or less will yield a negative value of \( \hat{y} \). The importance of simple size in evaluating a regression equation is once again emphasized.

\[
\hat{y} = \bar{y} \pm \frac{\hat{y} \cdot \hat{y}}{\hat{y}} \cdot \hat{y}
\]

are shown (Table 3) for the series of simple linear regression equations with the same "goodness-of-fit" for equal intercepts and different slopes. Both \( \hat{y} \) and \( \hat{y} \) decrease as slope increases, and the difference between \( \hat{y} \) and \( \hat{y} \) decreases as slope increases with \( \hat{y} \), always being larger than \( \hat{y} \). Notice that \( \hat{y} \) is negative for a slope of 0.64. For simple linear regression (p=1), the largest sample size for which \( \hat{y} \) is negative for various values of \( \hat{y} \) is given in Table 6.

Table 6. Maximum sample size (n) for which % of size is negative for various values of \( \hat{y} \) of simple linear regression.

<table>
<thead>
<tr>
<th>n</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>0.02</td>
<td>50</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>0.03</td>
<td>40</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>0.25</td>
<td>14</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

\( \hat{y} \) is negative for any sample size when \( \hat{y} = 0.711 \) from our example, \( \hat{y} \), and \( \hat{y} \) is negative for the practitioner must worry about the predictive precision of a regression equation than does \( \hat{y} \). However, both of these terms can be -11.
negative for small sample sizes, especially for small values of $b^2$.

The contribution to predictive precision of adding a new independent variable to a regression equation already having $b$ independent variables can be partially determined by comparing $R_{adj}^2$, for the $b + 1$ independent variables with the $R_{adj}^2$ for the $b$ independent variables, if the $R_{adj}^2$, for the new model (a less than $R_{adj}^2$ for the old model ($R_{adj}^2$ for new model is larger than $R_{adj}^2$ for old model), the new model with $b$ independent variables has less predictive precision than the old model with only $b$ independent $X$ variables. The $F$ value calculated using equation (8) would indicate this lack of improvement in predictive precision by being less than 1.

Conclusions

$R^2$ and the significance of the regression equation are useful criteria for evaluating regression equations in causal studies and predictive modeling. However, they should not be used alone as they do not tell the entire story. $R^2$ is a measure of both the goodness-of-fit and the steepness of a regression surface. For the same "goodness-of-fit", $R^2$ increases as the slope of the regression equation increases. The significance of the regression equation increases as sample size and $R^2$ increase. $R^2$ is a biased estimate of $\beta_0$ when $\beta_0 = 0$ and $\gamma > 0$ with the bias increasing as $\gamma$ increases and $n$ decreases. $R_{adj}^2$, and $1 - \frac{SSE_{adj}}{SSE}$ are more practical than $R^2$ for measuring the predictive precision of a regression equation and can have negative values for small values of $b^2$, especially for small sample sizes. $R_{adj}^2$, is always smaller than $R^2$ with the difference decreasing as $n$ and $R^2$ increase and $\gamma$ decreases.

We suggest that the natural resource practitioner use the following set of criteria to evaluate a single regression equation: compare regression equations from different sets of data, or compare different regression equations based on the same set of data.

(1) $\gamma$, $\sigma$, and $\beta_0$ and $\beta_1$

(2) $F$, $\theta$, and $\gamma$ (slope angle in degrees)

(3) $R^2$, $ECF^2$, $\theta = 0$, $ECF^2|\beta_0 = 0$, and $R^2$

(4) $R_{adj}^2$, $1 - \frac{SSE_{adj}}{SSE}$, and $R^2$

This set of criteria is given below for our single linear regression example.
The results indicate that the regression equation is very significant. $E^{2}$ is relatively high, considerably larger than $E(N)$ when $\gamma_1 = 0$, and approximately equal to $E(N)$ when $\gamma_1 = 0.7213$, indicating that the probability of $E^{2}$ being this large by chance is negligible. The moderate sample size of 10 and not being considerably less than 200 also indicates this. Both $\gamma_1$ and $\gamma_2$ indicate a relatively strong relationship between $y$ and $z$ for relatively strong positive relationship between $y$ and $z$. The slope of the regression line is 42.69, indicating that $E^{2}$ should not be used solely to evaluate the usefulness of the regression equation. $R^{2}$ and $1 - R_{y-z}^2$ indicate that a major proportion of $R_{y-z}^2$ or $\gamma_2$ are explained by the regression equation. The predictive precision of the regression equation measured by $\gamma_2$ is 2.97 as a 45.82 reduction (improvement over the predictive precision of the regression equation measured by $\gamma_1$) is 5.48.

The shows one of criteria plus the data, testing to see if the assumptions of the model are adequately met, computing confidence and precision intervals, and testing the reliability and validity of the regression model represent a relatively complete evaluation of the usefulness of a regression equation.

**LITERATURE CITED**


A NOTE ON DOUBLE SAMPLING - POINT SAMPLING

by

Barry V. Viess, Jr.

and

Noris Exide 2/3

Double sampling is sometimes used with point sample counting, especially when the measurement of qualifying trees is rather detailed and therefore expensive (Heers and Miller 1964). In-trees are counted on all point samples (X-variable) and volume (Y-variable) is determined on a randomly selected subsample of points. The relationship of Y to X on the subsample is evaluated using a regression estimator and is used to adjust the X of the large sample to estimate Y for the population. The appropriate form of the regression depends on such factors as the Intercept value and the variance of Y at given values (Frisseau 1962). Since information necessary to select the appropriate regression estimator may not be readily available, a JP selection of the subsample as described by Viess (1963) may be preferred as no conditions must be met for valid estimates (Groves and Field 1963).

When using double sampling or the MP method, we recommend the X-variable be estimated number of logs on in-trees (as pioneered beliefs) rather than in-tree counts, since the former is usually, if not always, more strongly related to point-sample determined per-acre volume. Actually, logs on in-trees times a constant value estimates per-acre volume at a given point; for 60° E, 35° L, the constant is about 600 for southern pines, 670 for Appalachian hardwoods (Viess and Morsey 1965). Such consistency is not found for predicting per-acre volume from in-tree counts (i.e., per-acre basal area). In a New Jersey study involving ten separate point sample cruises in a forest, the coefficient of variations for ratios of volume/logs on in-trees were lower in every case than those for volume/in-tree count ratios, averaging 29% and 37%, respectively. It is no obvious advantage to use the X-variable most closely related to the Y-variable when double sampling.

Estimating the number of logs on in-trees in eastern forests takes little more time than in-tree counts. In the New Jersey study, counting logs took only 115 more time than counting in-trees, averaging 1.08 and 1.09 minutes per point, respectively.

2/ The authors -- Barry V. Viess, Jr. is professor, Division of Forestry, West Virginia University, Morgantown and Noris Exide is assistant professor, Department of Horticulture and Forestry, Rutgers University, New Brunswick, New Jersey.
Literature Cited


Current Literature

Please print directly from sources given.

General


---

Elmcrest Digest - A newsletter is available from Technical Insights Inc. 2337 Lemon Ave., P.O. Box 1864, Fort Lee, NJ 07024. Subscription rates are $97.00.

---


---


---

Plant Disease is a new international journal emphasizing the practical aspects of plant pathology and improving plant health. For details and subscription, write The America Phytopathological Society, 1340 Pilot Knob Road, St. Paul, MN 55121.

---
IMHE No. 78/79 "A Handbook for Assessment of Environmental Benefits and Pollution Control Costs" is available from the National Technical Information Service, 5285 Port Royal Avenue, Springfield, VA 22151. The stock No. is PB 291 780/75. The price is $6.75 for paper copy and $3.00 for microfiche.

PM-3-117. "Georgia Key Islands National Park Integrated Resource Survey."
PM-3-118. "St. Lawrence Islands National Park and Surrounding Areas Integrated Resource Survey."
PM-5-121. "Wolf Study for Canadian Forest Resource Data System."
From Peterman National Forestry Inst., Canadian Forestry Service, Chalk River, Ontario K0J 1C0 Canada.


Data Processing
78-57 Linear Regression Analysis Using a Programmable Pocket Calculator.
78-58 Calculation of the Two-way Analysis of Variance (ANOVA) using a programmable pocket calculator.
78-69 Calculation of the Two-way Analysis of Variance (ANOVA) with subsampling using a programmable pocket calculator.
78-90 Calculation of Multiple Regression with Three Independent Variables using a programmable pocket calculator.


"Random Numbers, Means, Regression and The Programmable Calculator" a 143 page monograph by Thomas W. Were is available for $15.00 from T & C Enterprises, P.O. Box 1190, West Lafayette, IN 47906.
Range and Wildlife

"Aerial Census of Wild Horses in Western Dui" by Piel, Peterson and Hall.


"Density Estimation by Variable Area Transects" by Parker in Journal of Wildlife Management 43(1), Apr. 1979 at your local conservation library.

Remote Sensing


FMR-5-118, "Recognition of Tree Species on Aerial Photographs."
FMR-5-1027, "Application De Photographies A Grande Echelle A Un Inventaire Forestier En Algerie".
FMR-5-121, "A Forest Inventory in the Faken Using Large Scale Photo-Interpretation Techniques" all available from Petawawa National Forestry Institute, Canadian Forestry Service, Chalk River, Ontario K0J 1C0, Canada.

Soils


Special Report 522 “Estimated Crop Production Costs on Sandy Soils with Center-Point Irrigation Systems, Oregon’s Columbia Basin, 1978.”

From Cooperative Extension Service, USDA, Oregon State Univ. Extension Hall, Corvallis, OR 97331.

———

Rall, A.R. Rating South Dakota Soils According to Productivity. From Agric. Experiment Sta., South Dakota State Univ., University Station, Brookings, SD 57007.

———

Kansas Improvement Study No. 73 “The 3-F Erosion Bridge - A New Tool for Measuring Soil Erosion” from Calif. Dept. of Forestry, 1416 Ninth Street, Sacramento, CA 95814.

———

Soil Science Fact Sheet. 85-10 “Soil Profile and Horizon Designations” contact Corp. Extension Service, Univ. of Florida, Inst. of Food and Agricultural Sciences, Gainesville, FL 32611 for availability.

———

“Soil Erosion: Prediction and Control” A 793 paged review. Copies are available for 17.00 from Soil Conservation Society of America, 7315 Northeast Ambory Road, Ames, IA 50010.

———

“Klopper Mapping of Soils in San Juan County, New Mexico” by Carey and Keetch. In Journal of Soil and Water Conservation 34(23) Mar-Apr. 79 at your local conservation library.

———

Manual on Soil Sampling and Methods of Analysis. Price $10.00 (Canadian) from Canadian Society of Soil Science, Ottawa, Ontario, K1P 5H8.

———


**

19
A Symposium on "Remote Sensing for Natural Resources: An International View of Problems, Prospects and Accomplishments" Sponsored by IFSS, Society of American Foresters (Remote Sensing SG), American Society of Photogrammetry, and Geographic Society of America. September 10-14, 1976, at the University of Idaho, Moscow, Idaho. Papers will be presented by 50 international leaders in remote sensing of natural resources. One day is reserved for tours, one of which is a 1½-hour trip up the Snake River. For more information, write or call: Continuing Education, University of Idaho, Moscow, Idaho 83843. Ph: (208) 885-1000.

The Mosher Recreation Meeting will be held at Arrow Lake Lodge Resort, Kelowna, British Columbia, October 2-3, 1976. Space will be limited. Those wishing to attend should contact Martin Bauer or Don Mill at NSF Northenders Forest Group, P. O. Box 393, Bellevue, Ohio 43215 or call (614) 389-4471.


Convening in 1980 — The 10th International Congress of the International Society of Photogrammetry to be held in Hamburg, Germany. For additional information, write: The Secretaries, ISPRS Congress 1980, c/o Vereinigung, Hamburg, Federal Republic of Germany.

Call for Papers! Arid Land Resources Inventory, an International Workshop sponsored by IFSS subject group 56.02, ISPRS Section Working Group, the Mexican Association of Professional Foresters, the Mexico Forest Service, the U.S. Forest Service and the ISPRS Section on Land Management. The dates are November 30-December 6, 1980 and the place — La Paz, Mexico.

Theme: The arid lands of the world have recently been considered wastelands. However, these lands contain many values including unique scenery, wildlife populations, herds, shrubs, and trees.

In recent years, demands on arid lands have been increasingly attractive for recreation, urban development, underdeveloped arid lands, domestic livestock grazing and wood products. To assure protection and proper management of aridity and nonrenewable values, identification of the basic attributes to arid lands are required. The relative use economic values of the lands and their readiness sales inventory design a challenge.

The purpose of this workshop will be to discuss cost efficient methods for inventoring the Arid Lands.

20


Invention Analysis and Schedule

November 10
November 1

November 2
November 3
November 4
November 5

November 6

Registration
Welcome and Introductions
Arid Land Characteristics
Resources and Data

Meeting the Challenges

Aridity Planning

Classification Scheme
Economical Mapping System

Field Trip


Economical

Measuring the Challenges (continued)

Sampling Scheme

Efficient Measuring Techniques

Measuring the Challenges (continued)

Resource Data Analysis System

What Do We Go From Here?

Proposals for Exchanging Research

Business Meetings

SAP and ZEBI Groups

Contributed papers are now being solicited for the workshop. All papers submitted must fall within the theme of meeting; address the challenges, and must provide "how to" approaches.

All those interested in submitting a paper should send 3 copies of the title, a short abstract, your name, mailing address and phone number to:

Dr. Richard Brink
USDA Forest Service, RMF & RES
251 West Prosper Street
Fort Collins, Colorado 80521-1150

All contributions must be received by November 10, 1983.

The number of papers that can be accepted will be selected based on their appropriateness of the theme. Acceptance letters will be sent out by February 1, 1984.

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

July, 1983 marked the 4th anniversary of RESOURCE INVENTORY NOTES. The "NOTES" were started by the U.S. Forest Service in 1979 and were continued by the Bureau of Land Management starting in September, 1979. The "NOTES" started with an initial mailing of 50%, however over 2,850 copies are received in some 126 countries. The following is a complete listing of the major articles covered to date in the "NOTES". A limited number of past issues are available from our office.

21
Multi-Stage and Multi-Phase Sampling by Diggle
Preliminary Final - Jasper Valley Tables by Nielson

Converting Existing Bank to Inside Bank
Nonmechanized by Corpsmen

REL 20. April 1979 - Nielson, Director on Ecological land Classification by Majch and Wilson
The Precision of Out Grid Estimates - A Theoretical Approach by Gaver

REL 21. May 1979 - Specification of Unsaturated's Form Class 78 and 80 Cubic-Foot Volume Tables by Boyle and Tomich
UNA's Standard, Non-Standard, Road Inventory System by Nelson and Land

REL 22. June 1979 - Preparation of Type for Manual Digging by Nielson
owsal Probability Stating With Replacement and Without Replacement by Edmondson, Hollon and Babcock
University Distributing Supplies Within a Taper Island by Land

University Volume First Time for Number Inventory by Saddle
The Sampling Program in Statistical Geometry by En

--- End ---

Note: Local **_** and current literature and meeting announcements for publishing in the "NOTES". If announcing a meeting, please allow at least a four-month time.

--- End ---

Change of Address: Be sure to send us your old label. If you want to get **_** or off our mailing list, drop us a line.

--- End ---