Resource Inventory Notes

BLM 24 August 1979

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

by

Gary W. Fowler and Fred H. Bigelow[1/]

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

$$(1) \quad Y_i = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i \quad (i = 1, \ldots, n)$$

where Y_i = observed value of the dependent or response variable for the ith element

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

β_o = regression constant or Y-intercept of the regression equation

β_j = regression (net adjusted or partial) coefficient associated with the jth ($j = 1, \ldots, p$) independent variable

e_i = random error term or that part of Y_i not explained by the X_j's ($j = 1, \ldots, p$)

p = number of independent variables

n = number of elements in the sample

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

1/ 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:

USDI, 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_j's can be curvilinear).

The parameters β_o, β_1, ..., β_p are unknown and are usually estimated using Least Squares based on a random sample of n elements (cases or observations) from the population of interest. The Least Squares sample regression equation is

(2) $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip}$ (i = 1, ..., n)

where $\hat{Y}_i$ is the predicted value of Y_i for the i<u>th</u> element, and $\hat{\beta}_o$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ are the Least Squares estimates of the respective population parameters.

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

The assumptions associated with Model I multiple linear regression are:

(1) independent observations (the e_i's are independent)

(2) linearity in the parameters

(3) homogeneous variances (the e_i's are homogeneously distributed around the regression surface)

(4) X_j's are fixed constants and measured without error

(5) the Y_i's (or e_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 fertilizer on tree growth or temperature on brook trout populations. Another application is in modeling situations where a variable that is difficult and expensive to measure is predicted from variables that are easy and inexpensive to measure. For example, tree volume can be predicted from tree height and diameter measurements.

Many practitioners in natural resources use the coefficient 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), or the predictive precision (i.e., how well $\hat{Y}_i$ predicts Y_i) 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:

(3) SST = SSR + SSE

where SST = total sum of squares
SSR = sum of squares explained by the regression equation
SSE = sum of squares not explained by the regression equation (error sum of squares)

R^2 is the proportion of the total sum of squares explained by the regression equation, and

(4) $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

The multiple correlation coefficient (R) is also the simple linear correlation between Y_i and Y_i $(i = 1, \ldots, n)$. 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

(5) $Y_i = \alpha + \beta X_i + e_i \quad (i = 1, \ldots, n)$

where α is the intercept and β is the slope of the population regression line. The sample Least Squares regression line is

(6) $Y_i = \hat{\alpha} + \hat{\beta} X_i$

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

Table 1. Values of the dependent variable (Y) and the independent variable (X) for n = 20 pairs of observations.

Observation No. (i)	Y_i	X_i	Observation No. (i)	Y_i	X_i	Observation No. (i)	Y_i	X_i
1	8	4	8	7	10	15	16	16
2	4	5	9	12	11	16	13	17
3	2	6	10	16	12	17	19	17
4	9	6	11	11	13	18	22	18
5	4	7	12	17	14	19	18	19
6	8	8	13	16	15	20	16	20
7	12	9	14	13	16			

RELATIONSHIP OF R^2 TO MULTIPLE LINEAR REGRESSION

In multiple linear regression,

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} y_i^2$$

$$SSR = R^2 \sum_{i=1}^{n} y_i^2$$

$$SSE = (1-R^2) \sum_{i=1}^{n} y_i^2$$

where $y_i = Y_i - \bar{Y}$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. The calculated F-statistic

$$(7) \quad F_{p,n-p-1} = \frac{MSR}{MSE} = \frac{R^2/p}{(1-R^2)/(n-p-1)},$$

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

H_o: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (no significant regression equation)

against the alternative hypothesis

H_1: H_o false (significant regression equation)

The decision rule to test H_o is:

Reject H_o if $F_{p,n-p-1} > F_{\alpha;p,n-p-1}$; otherwise accept H_o

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

The test statistic

$$(8) \quad F_{g,n-p-1} = \frac{(R_p^2 - R_h^2)/g}{(1 - R_p^2)/(n-p-1)}$$

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_h^2 is the coefficient of determination for the original model with h independent variables, and R_p^2 is the coefficient of determination for the full model with p independent variables.

Thus, the F-statistics 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),

$$\hat{\beta} = \sum_{i=1}^{20} y_i x_i / \sum_{i=1}^{20} x_i^2 = 0.92$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 0.95$$

$$\hat{Y}_i = 0.95 + 0.92X_i$$

$$r = \sum_{i=1}^{20} y_i x_i / (\sum_{i=1}^{20} y_i^2 \cdot \sum_{i=1}^{20} x_i^2)^{1/2} = 0.85$$

$$R^2 = r^2 = 0.7213$$

$$SST = \sum_{i=1}^{20} y_i^2 = 570.55$$

$$SSR = R^2 \sum_{i=1}^{20} y_i^2 = 411.53$$

$$SSE = (1-R^2) \sum_{i=1}^{20} y_i^2 = 159.02$$

$$F_{1,18} = 46.58$$

where $x_i = X_i - \bar{X}$ and $\bar{X} = \sum_{i=1}^{n} X_i/n$.

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

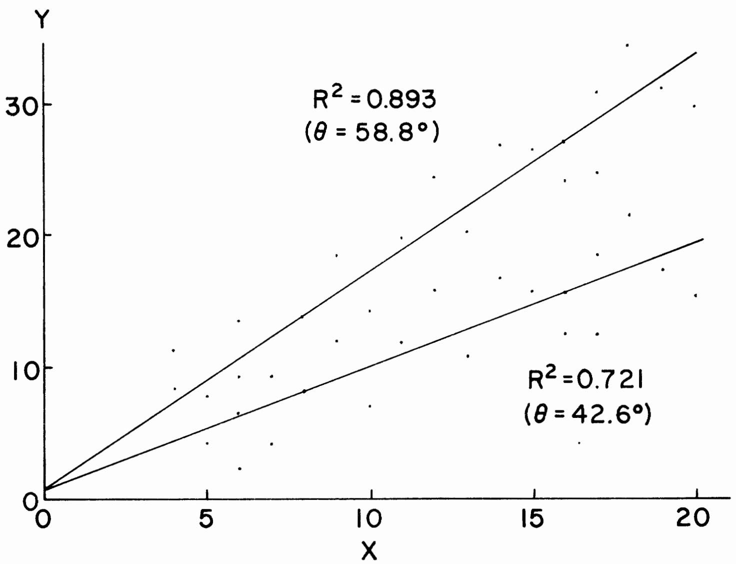

Figure 1. Simple linear regression lines having slope angles of 42.6° and 58.8° with the same "goodness-of-fit" as the example based on 20 observations.

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).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Calculated F Value
Regression (R)	411.53	1	411.53	46.58
Error (E)	159.02	18	8.83	
Total (T)	570.55	19		

$F_{0.05;1,18} = 4.41$ for $\alpha = 0.05$. Thus, there is a significant linear relationship between Y and X. The significance probability associated with the calculated F value is $\hat{P} < 0.005$. 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 1974). If the goodness-of-fit (SSE) of the regression surface remains constant, R^2 increases as the slope of the regression surface increases[2/]. When the slope of the regression surface increases, $SST = \sum_{i=1}^{n} (Y_i-\bar{Y})^2$ increases, which causes $R^2 = 1 - SSE/SST$ to increase. In other words, the simple linear correlation (R) between Y_i and $\hat{Y}_i$ increases as the slope of the regression surface increases.

The slope angle of the simple linear regression line for our example with $R^2 = 0.7213$ 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. R^2, s_Y^2, MSR, and the calculated F value ($F_{1,18}$) for these different regression lines are shown in Table 3 where

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

$$s_Y = \sqrt{s_Y^2} = \text{standard deviation of Y}$$

2/ 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. R^2, $R^2_{Adj.}$, s^2_Y, $1-s_{Y\cdot X}/s_Y$, and $F_{1,18}$ of the regression lines with equal intercepts, different slopes, and the same "goodness-of-fit" as found in our example (Table 1).

Slope angle in Degrees (θ)	R^2	$R^2_{Adj.}$	s^2_Y	MSR	$1-\frac{s_{Y\cdot X}}{s_Y}$	$F_{1,18}$
0.6	0.0003	-0.055	8.37	0.04	-0.03	0.004
10.8	0.099	0.049	9.29	17.40	0.02	1.97
26.6	0.431	0.400	14.72	120.62	0.23	13.65
42.6	0.721	0.706	30.03	411.53	0.46	46.58
58.8	0.893	0.887	78.11	1,324.98	0.66	149.98
74.7	0.976	0.975	348.80	6,468.11	0.84	732.21
87.5	0.999	0.999	13,292.41	252,396.50	0.97	28,495.81

The data points and sample regression lines having the same goodness-of-fit for slopes of 42.6° (our original example) and 58.8° are shown in Figure 1.

R^2 is a function of the slope angle in degrees (θ) and varies from 0.0003 for a slope of 0.6° to 0.999 for a slope of 87.5° (Table 3, Figure 2).

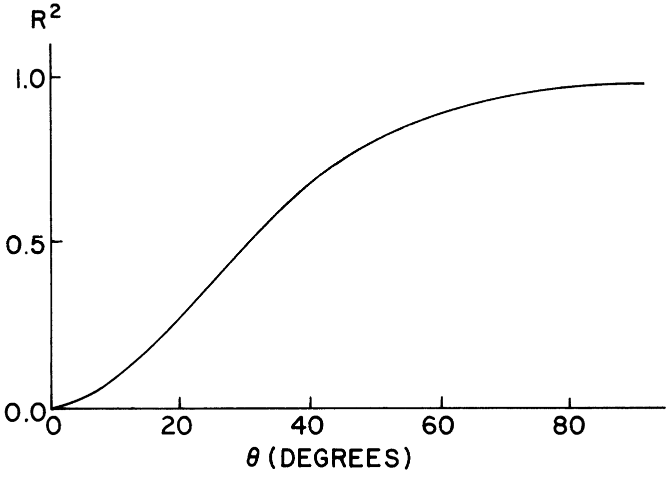

Figure 2. R^2 as a function of θ in degrees for simple linear regression lines having the same "goodness-of-fit" as our example (Table 1).

For simple linear regression,

$$r^2 = (\sum_{i=1}^{n} y_i x_i)^2 / (\sum_{i=1}^{n} y_i^2 \cdot \sum_{i=1}^{n} x_i^2) = \hat{\beta} \sum_{i=1}^{n} y_i x_i / \sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 / \sum_{i=1}^{n} y_i^2$$

$$\beta = \tan \theta = \sum_{i=1}^{n} y_i x_i / \sum_{i=1}^{n} x_i^2$$

$$(9) \quad R^2 = r^2 = (\tan \theta)^2 \sum_{i=1}^{n} x_i^2 / \sum_{i=1}^{n} y_i^2$$

Table 3 also shows that s_Y^2, MSR, and $F_{1,18}$ increase as θ increases. The results of Table 3 show conclusively 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.

Thus, when evaluating a single regression equation, comparing different regression equations from different sets of data, or comparing different regression equations from the same set of data, causal relationships or predictive precision 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 precision than a regression equation with a lower R^2 if the higher R^2 is associated with a steeper slope.

An examination of equation (7) for the calculated F value shows that the number of observations must be considered when evaluating the significance of the regression equation. For R^2 constant, the F value and the significance of the regression equation increase as n increases. This creates 2 problems for the practitioner. The problem of large sample sizes is that a very small R^2 will yield a significant regression equation if the sample size is large enough. The problem of small sample sizes is that a very large R^2 will yield an insignificant regression equation if the sample size is small enough. Another danger associated with a small sample size is that a regression equation can be developed with a high R^2 that fits the data points 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 μ_{R^2} (Wishart 1931, Kendall and Stuart 1967). When there is no relationship between Y_i and the X_j's in the population ($\mu_{R^2} = 0$), the expectation of R^2 is

$$(10) \quad E(R^2) = E(R^2 | \mu_{R^2} = 0) = p/(n-1)$$

Therefore, the bias is always positive and equal to $p/(n-1)$. $E(R^2)$ is the mean value of R^2's calculated from all possible samples of size n from the population assuming the Y_i's are normally distributed. $E(R^2)$ for various values of n and p when $\mu_{R^2} = 0$ is shown in Table 4. These are exact values.

Table 4. $E(R^2)$ for various values of n and p when $\mu_{R^2} = 0$.

n	p			
	1	2	4	8
3	0.500			
5	0.250	0.500		
10	0.111	0.222	0.444	0.889
20	0.053	0.105	0.211	0.421
50	0.020	0.041	0.082	0.163
100	0.010	0.020	0.040	0.081
200	0.005	0.010	0.020	0.040
1000	0.001	0.002	0.004	0.008

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

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

R^2 is also a biased estimate of μ_{R^2} when $\mu_{R^2} > 0$. $E(R^2)$ in this case was shown by Wishart (1931) to be

$$(11) \quad E(R^2) = E(R^2 | \mu_{R^2} > 0) = 1 - \frac{n-p-1}{n-1} (1-\mu_{R^2}) F\left(1,1, \tfrac{1}{2}(n+1), \mu_{R^2}\right)$$

where $F(1,1, \frac{1}{2}(n+1), \mu_{R^2})$ is the hypergeometric function with parameters 1,1, $\frac{1}{2}(n+1)$, and μ_{R^2}. Equation (11) is applicable to the multinormal situation (Kendall and Stuart 1967) where both Y and the X_j's are random variables and is only approximate for Model I multiple linear regression where the X_j'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 $E(R^2)$:

$$(12) \quad E(R^2) \approx \mu_{R^2} + \frac{p}{n-1}(1-\mu_{R^2}) - \frac{2(n-p-1)}{(n^2-1)} \mu_{R^2}(1-\mu_{R^2}).$$

This approximation is accurate to the order $1/n^2$.

The bias associated with R^2, in general, decreases as n increases for a given value of μ_{R^2} and as μ_{R^2} increases for a given value of n. The bias increases as p increases for μ_{R^2} and n constant. $E(R^2)$ is given for various values of n and μ_{R^2} for simple linear regression (p=1) in Table 5.

Table 5. $E(R^2)$ for various values of μ_R and n for simple linear regression. Values in parentheses for n=3 are exact. All other values are approximations of $E(R^2)$.

n	μ_R 0.05	0.25	0.50	0.75	0.95
3	0.513	0.578	0.688	0.828	0.963
	(0.513)	(0.568)	(0.653)	(0.769)	(0.921)
5	0.276	0.391	0.562	0.766	0.951
10	0.148	0.303	0.515	0.747	0.948
20	0.096	0.273	0.504	0.746	0.948
50	0.068	0.258	0.501	0.748	0.949
100	0.059	0.254	0.500	0.749	0.950
200	0.054	0.252	0.500	0.749	0.950
1000	0.051	0.250	0.500	0.750	0.950

Exact values for $E(R^2)$ based on equation (11) are given for n=3 while all other values of $E(R^2)$ are approximations based on equation (12). The bias is positive for $\mu_R < 0.50$, but becomes negative for large values of μ_R.

In using R^2 to evaluate a regression equation, the practitioner should remember that R^2 is biased for both the cases $\mu_R = 0$ and $\mu_R > 0$, and that the bias can be quite large for smaller sample sizes.

THE ADJUSTED COEFFICIENT OF DETERMINATION

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

$$(13) \quad R^2_{Adj.} = 1 - \frac{s^2_{Y\cdot X}}{s^2_Y} = 1 - (1 - R^2)\,\frac{n-1}{n-p-1}$$

and $s^2_{Y\cdot X} = MSE = SSE/(n-p-1)$ is the variance of Y about the regression equation. $R^2_{Adj.}$ is the proportion of s^2_Y (or variability around $\bar{Y}$) removed by the regression equation. $R^2_{Adj.}$ is always smaller than R^2 with this difference decreasing as n and R^2 increase and p decreases. One nice characteristic is that $E(R^2_{Adj.}) = 0$ when $\mu_R = 0$. $R^2_{Adj.}$ is more useful to the practitioner than R^2 in determining 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 $Y(s_Y)$ with the standard error of the estimate $(s_{Y\cdot X} = (s^2_{Y\cdot X})^{1/2})$ as follows:

(14) proportion of s_Y removed by $s_{Y \cdot X}$ (or the regression equation) $= \dfrac{s_Y - s_{Y \cdot X}}{s_Y} = 1 - \dfrac{s_{Y \cdot X}}{s_Y}$ (Crocker 1972)

The proportion $1 - s_{Y \cdot X}/s_Y$ is in terms of the units of the variable to be predicted.

The practitioner should be cautioned that $R^2_{Adj.}$ can be negative when R^2 is small and/or p is large compared to n. This means that $s^2_{Y \cdot X}$ is greater than s^2_Y which is caused by the fact that more has been lost in reduced degrees of freedom going from s^2_Y to $s^2_{Y \cdot X}$ than has been gained by reducing the sum of squares from SST to SSE. In other words, $\bar{Y}$ is a better predictor than the regression equation.

Setting equation (13) equal to zero and solving for n (this value of n will be called n*) yields

(15) $n^* = (p + R^2)/R^2$

For a given value of R^2, a regression equation based on a sample size less than n* will have a negative $R^2_{Adj.}$ For example, if $p = 5$ and $R^2 = 0.5$, a sample size of 10 or less will yield a negative value of $R^2_{Adj.}$ The importance of sample size in evaluating a regression equation is once again emphasized.

$R^2_{Adj.}$ and $1 - \dfrac{s_{Y \cdot X}}{s_Y}$ 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 $R^2_{Adj.}$ and $1 - \dfrac{s_{Y \cdot X}}{s_Y}$ increase as slope increases, and the difference between $R^2_{Adj.}$ and R^2 decreases as slope increases with $R^2_{Adj.}$ always being larger than R^2. Notice that $R^2_{Adj.}$ is negative for a slope of 0.6°. For simple linear regression (p=1), the largest sample size for which $R^2_{Adj.}$ is negative for various values of R^2 is given in Table 6.

Table 6. Maximum sample size (n) for which $R^2_{Adj.}$ is negative for various values of R^2 for simple linear regression.

R^2	n	R^2	n
0.01	100	0.10	10
0.02	50	0.20	5
0.025	40	0.30	4
0.05	20	0.40	3
0.075	14	0.49	3

$R^2_{Adj.}$ is positive for any sample size when $R^2 = 0.7213$ from our example. $R^2_{Adj.}$ and $1 - s_{Y \cdot X}/s_Y$ tell the practitioner much more about the predictive precision of a regression equation than does R^2. However, both of these terms can be negative for small sample sizes, especially for small values of R^2.

The contribution to predictive precision of adding g new independent variables to a regression equation already having h independent variables can be partially determined by comparing $R^2_{Adj.}$ for the p = g+h independent variables with the $R^2_{Adj.}$ for the h independent variables. If the $R^2_{Adj.}$ for the new model is less than $R^2_{Adj.}$ for the old model ($s_{Y \cdot X}$ for new model is larger than $s_{Y \cdot X}$ for old model), the new model with p independent variables has less predictive precision than the old model with only h independent 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 μ_{R^2} when $\mu_{R^2} = 0$ and $\mu_{R^2} > 0$, with the bias increasing as p increases and μ_{R^2} and n decrease. $R^2_{Adj.}$ and $1 - s_{Y \cdot X}/s_Y$ are more practical than R^2 for measuring the predictive precision of a regression equation and can have negative values for small values of R^2, especially for small sample sizes. $R^2_{Adj.}$ is always smaller than R^2 with the difference decreasing as n and R^2 increase and p 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) n, p, s_Y, and $s_{Y \cdot X}$

(2) $F_{p,n-p-1}$, $\hat{P}$, and θ (slope angle in degrees)

(3) R^2, $E(R^2 | \mu_{R^2} = 0)$, $E(R^2 | \mu_{R^2} = R^2)$, and R

(4) $R^2_{Adj.}$, $1 - s_{Y \cdot X}/s_Y$, and n*

This set of criteria is given below for our simple linear regression example.

n	20	R^2	0.7213
p	1	$E(R^2 \mid \mu_{R^2} = 0)$	0.0526
s_Y	5.48	$E(R^2 \mid \mu_{R^2} = R^2)$	0.7178
$s_{Y \cdot X}$	2.97	R	0.85
$F_{p,n-p-1}$	46.58	$R^2_{Adj.}$	0.7058
$\hat{P}$	<0.005	$1 - s_{Y \cdot X}/s_Y$	0.4576
θ	42.6	n^*	2.39

Results indicate that the regression equation is very significant. R^2 is relatively high, considerably larger than $E(R^2)$ when $\mu_{R^2} = 0$, and approximately equal to $E(R^2)$ when $\mu_{R^2} = 0.7213$, indicating that the probability of R^2 being this large by chance is negligible. The moderate sample size of 20 and p=1 being considerably less than n=20 also indicate this. R=0.85 indicates a relatively strong relationship between Y_i and $\hat{Y}_i$ (or relatively strong positive relationship between Y and X). The slope of the regression line is 42.6°, indicating that R^2 should not be used solely to evaluate the usefulness of the regression equation. $R^2_{Adj.}$ and $1 - s_{Y \cdot X}/s_Y$ indicate that a major proportion of s^2_Y or s_Y are explained by the regression equation. The predictive precision of the regression equation measured by $s_{Y \cdot X} = 2.97$ is a 45.8% reduction (improvement over) the predictive precision of $\bar{Y}$ measured by $s_Y = 5.48$.

The above set of criteria plus plotting the data, testing to see if the assumptions of the model are adequately met, computing confidence and prediction 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

Crocker, D.C. 1972. Some interpretations of the multiple correlation coefficient. The American Statistician 26(2):31-33.

Barrett, James P. 1974. The coefficient of determination--some limitations. The American Statistician 28(1):19-20.

Kendall, M.G. and A. Stuart. 1967. The Advanced Theory of Statistics. Vol. II:341-42. Hafner Publishing Co., New York.

Wishart, J. 1931. The mean and second moment coefficient of the multiple correlation coefficient, in samples from a normal population. Biometrika 22:353-361.

* * * * *

A NOTE ON DOUBLE SAMPLING - POINT SAMPLING

by

Harry V. Wiant, Jr.

and

Boris Zeide [2]

Double sampling is sometimes used with point sample cruising, especially when the measurement of qualifying trees is rather detailed and therefore expensive (Beers 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 $\bar{x}$ of the large sample to estimate $\bar{y}$ for the population. The appropriate form of the regression depends on such factors as the intercept value and the variance of Y at given X-values (Freese 1962). Since information necessary to select the appropriate regression estimator may not be readily available, a 3P selection of the subsample as described by Wiant (1976) may be preferred as no conditions must be met for valid estimates (Grosenbaugh 1967).

When using double sampling or the 3P method, we recommend the X-variable be estimated number of logs on in-trees (or pulpwood bolts) 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 BAF = 10, Int. - 1/4", the constant is about 600 for southern pine, 670 for Appalachian hardwoods (Wiant and Maxey 1979). 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 of 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 an 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 11% more time than counting in-trees, averaging 1.88 and 1.69 minutes per point, respectively.

2/ The authors -- Harry V. Wiant, Jr. is professor, Division of Forestry, West Virginia University, Morgantown and Boris Zeide is assistant professor, Department of Horticulture and Forestry, Rutgers University, New Brunswick, New Jersey.

Literature Cited

Beers, T. W., and C. I. Miller. 1964. Point sampling: research results theory and applications. Purdue Univ. Res. Bul. 786.

Freese, F. 1962. Elementary forest sampling. USDA For. Ser. Agri. Handbook 232.

Grosenbaugh, L. R. 1967. The gains from sample-tree selection with unequal probabilities. J. Forest. 65: 203-206.

Wiant, H. V., Jr. 1976. Elementary 3P sampling. WV Agri. & Forest. Exp. Sta. Bul. 650T.

Wiant, H. V., Jr., and W. R. Maxey. 1979. Board-foot factors for point sampling. J. Forest. 77:29.

* * * * *

Current Literature

Please order directly from sources given.

General

"Environmental Analysis for Land Use and Site Planning." W.M. Marsh, from McGraw-Hill, 1221 Ave. of the Americas, New York, NY 10020. Price $21.50.

\- - - - - - - - - - -

Biomass Digest - A new newsletter is available from Technical Insights Inc. 2337 Lemoine Ave., P.O. Box 1304, Fort Lee, NJ 07024. Subscription rates are $87.00.

\- - - - - - - - - - -

Pub. No. 12. "A Handbook of Graphical Solutions to Forest Biometric Problems." From Dept. of Forestry, Southern Illinois Univ., Carbondale, IL 62901.

\- - - - - - - - - - -

AERR-156 "Estimating the Economic Effects of Changes in Land Use: A Guide" From Agricultural Exp. Stn., Univ. of Illinois, Urbana, IL 61801.

\- - - - - - - - - - -

Plant Disease is a new international journal emphasizing the practical aspects of maintaining and improving plant health. For details and subscription rates write The America Phytopathological Society, 3340 Pilot Knob Road, St. Paul, MN 55121.

\- - - - - - - - - - -

IINR Doc. No. 78/29 "A Handbook for Assessment of Environmental Benefits and Pollution Control Costs" is available from the National Technical Information Service, 5825 Port Royale Road, Springfield, VA 22161. The Stock No. is PB 291 786/AS. The price is $6.75 for paper copy and $3.00 for microfiche.

- - - - - - - - - -

FMR-X-117. "Georgian Bay Islands National Park Integrated Resource Survey."

FMR-X-114. "St. Lawrence Islands National Park and Surrounding Areas Integrated Resource Survey."

FMR-X-122. "Pilot Study for Canadian Forest Resource Data System."

From Petawawa National Forestry Inst., Canadian Forestry Service, Chalk River, Ontario KOJ 1JO Canada.

- - - - - - - - - -

"Wild Rivers - Methods for Evaluation" by Sonnen and Davis in Water Resource Bulletin 15(2):404-419 at your local conservation library.

- - - - - - - - - -

Data Processing

TB-47 Linear Regression Analysis Using a Programmable Pocket Calculator.

TB-48 Calculation of the Two-way Analysis of Variance (ANOVA) using a programmable pocket calculator.

TB-49 Calculation of the Two-way Analysis of Variance (ANOVA) with subsampling using a programmable pocket calculator.

TB-50 Calculation of Multiple Regression with Three Independent Variables using a programable pocket calculator.

All from Agri. Exp. Stn., South Dakota State Univ., University Station, Brookings, SD 57007.

- - - - - - - - - -

BULL. 615 "Estimating Water Requirements for Corn with a Pocket Calculator" from: Agricultural Experiment Stn. Kansas State Univ., Manhattan KS 66506.

- - - - - - - - - -

"Random Numbers, Means, Regression and The Programmable Calculator" a 143 paged monogram by Thomas W. Beers is available for $8.00 from T & C Enterprises, P.O. Box 2196, West Lafayette, IN 47906.

- - - - - - - - - -

ERR 159 - "A Computer Program for Factor Analysis Regression." from Agric. Exp. Stn., Univ. of Illinois, Urbana, IL 61801.

- - - - - - - - - - -

Forestry

"Conferencia Internacional sobre utilizacion de los bosques tropicales del mundo." by Miguel Caballero Deloya et al. in Ciencia Forestal, 3(13): 30-47. Contact Revista Ciencia Forestal, Progres No. 5, Coyoacan (21), D.F., Mexico for availability and price.

- - - - - - - - - - -

Res. Report No. 68 "Preliminary Small-Tree Above-ground Biomass Tables for Five Northern Hardwoods", Res. Report No. 69. "Shift-Share Analysis for Measuring Economic Development: A Basic Language Computer Program" from: Agric. Exp. Stn., Taylor Hall, Univ. of New Hampshire, Durham, NH 03824.

- - - - - - - - - - -

AFRI Research Note No. 28. "Comparison of the BAF 20 and BAF 10 Variable - Radius Plot methods for estimating Gypsy Moth Egg Masses".

AFRI Research Report No. 39 "Location Key for Best Management Practices - Silviculture"

AFRI Research Report No. 40. "Whole Tree Weight Tables for New York" Are available from Applied Forestry Research Institute, State University of New York, College of Environmental Science and Forestry, Syracuse, NY 13210.

- - - - - - - - - - -

Gen. Tech, Report NE-44. "Solve II users Manual: A Procedural Guide for a Sawmill Analysis" from Northeastern Forest Exp. Stn., 370 Reed Road, Broomall, PA 19008.

- - - - - - - - - - -

Gen. Tech. Report PNW-71. Plotting Landscape Perspectives of Clearcut units.

Gen. Tech. Report PNW-76. Planning for Prescribed Burning in the Inland Northwest. Both from Pacific Northwest Forest and Range Exp. Stn., P.O. Box 3141, Portland, OR 97208.

- - - - - - - - - - -

Res. Paper. NC-163. "Equations for Estimating Stand Establishment, release, and thinning costs in The Lake States."

Res. Note NC-239. "Distribution of Biomass and Production for Several Northern Woody Species."

"Rare Plants of the Ozark Plateau - a field Identification Guide."

Gen. Tech. Rpt. NC-46. "Proceedings 1977 Midwest Forest Mensurationists Meeting" from Northcentral Forest Exp. Stn., 1992 Folwell Ave., St. Paul, MN 55108

All from Publications, North Central Forest Experiment Station, 1992 Folwell Avenue, St. Paul, MN 55108.

- - - - - - - - - - -

DNR Note No. 27. The Tariff System -- Revisions and Additions. From Dept. of Natural Resources, Olympia, WA 98504.

- - - - - - - - - - -

Range and Wildlife

"Aerial Census of Wild Horses in Western Utah" by Frei, Peterson and Hall.

"A Simple, Lightweight Point Frame" by Sharrow and Tober. Both in Journal of Range Management 32 (1), Jan 1979 at your local conservation library.

- - - - - - - - - - -

"Density Estimation by Variable Area Transect." by Parker in Journal of Wildlife Management 43(2). Apr. 1979 at your local conservation library.

- - - - - - - - - - -

Remote Sensing

"Flood Damage Assessment using computer-assisted analysis of color infrared photography" by Anderson and

"Application of Remote-Sensing technology to soil survey research" by Weismiller and Kaminstey both in Journal of Soil and Water Conservation, Vol. 33/No.6 Nov.-Dec. 1978 at your local conservation library.

- - - - - - - - - - -

FMR-X-118. "Recognition of Tree Species on Aerial Photographs."

FMR-X-107F "Application De Photographies A Grande Echelle A Un Invenaire Forestier En Alberta"

FMR-X-121 "A Forest Inventory in the Yukon Using Large Scale Photo Sampling Techniques" All available from Petawawa National Forestry Institute, Canadian Forestry Service, Chalk River, Ontario KOJ 1J0, Canada.

- - - - - - - - - - -

"Accuracy of Impervious Area Values Estimated Using Remotely Sensed Data" by Jackson and McCuen. In Water Resource Bulletin 15(2):436-446. At your local conservation library.

- - - - - - - - - - -

Soils

Special Report 521 "Estimated Crop Production Costs on Loam Soils with Side-Roll Irrigation Systems, Oregon's Columbia Basin, 1978."

Special Report 522 "Estimated Crop Production Costs on Sandy Soils with Center-Pivot Irrigation Systems, Oregon's Columbia Basin, 1978."
From: Cooperative Extension Service, USDA, Oregon State Univ. Extension Hall, Corvallis, OR 97331.

- - - - - - - - - - -

Bull. 657. Rating South Dakota Soils According to Productivity. from Agric. Experiment Stn., South Dakota State Univ., University Station, Brookings, SD 57007.

- - - - - - - - - - -

Range Improvement Study No. 23 "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. SL-10 "Soil Profile and Horizon Designations" contact Coop. Extension Service, Univ. of Florida, Inst. of Food and Agricultural Sciences, Gainsville, FL 32611 for availablility.

- - - - - - - - - - -

"Soil Erosion: Prediction and Control" A 393 paged review. Copies are available for $7.00 from Soil Conservation Society of America, 7515 Northeast Ankeny Road, Ankeny, IA 50021.

- - - - - - - - - - -

"Helicopter Mapping of Soils in San Juan County, New Mexico" by Carey and Keetch. In Journal of Soil and Water Conservation 34(2) 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 5H4.

- - - - - - - - - - -

"Predicting Site Productivity of Mixed Conifer Stands in Northern Idaho from Soil and Topographic Variables" by Brown and Locwenstein. In Soil Science Society of America Journal. 42(6) 967-971 at your local conservation library.

* * * * * *

Meetings

A Symposium on "Remote Sensing for Natural Resources -- An International View of Problems, Promises and Accomplishments". Sponsored by IUFRO, Society of American Foresters (Remote Sensing WG), American Society of Photogrammetry, and Geological Society of America. September 10-14, 1979, at the University of Idaho, Moscow, Idaho. Papers will be presented by 40 International leaders in remote sensing of natural resources. One day is reserved for tours, one of which is a jet boat trip up the Snake River. For more information, write or call: Continuing Education, University of Idaho, Moscow, Idaho, 83843. Ph (208) 885-6486.

\- - - - - - - - - - - - -

The Midwest Mensuration Meeting will be held at Atwood Lake Lodge Resort, Dellroy, Ohio, October 3-5, 1979. Space will be limited. Those wanting to attend should contact Martin Dale or Don Hilt at USFS Northeastern Forest Exp. Stn., P. O. Box 365, Delaware, Ohio, 43015 or call (614) 369-4471.

\- - - - - - - - - - - - -

1979 Society of American Foresters National Convention. October 14-17, 1979, Boston, Massachusetts. Theme: "Town Meeting Forestry -- Issues for the 80's". Contact E. F. Robie, SAF, 5400 Grosvenor Lane, Washington, D. C., 20014.

Coming in 1980 -- The 14th International Congress of the International Society of Photogrammetry to be held in Hamburg, Germany. For additional information, write The Secretariate, ISP Congress 1980, c/o Hamburg Messe and Congress GmbH, Congress - Organization, P. O. Box 30 23 60, D-2000 Hamburg 36, Federal Republic of Germany.

\- - - - - - - - - - - - -

Call for Papers! Arid Land Resource Inventories. An International Workshop sponsored by IUFRO subject group S4.02, SAF Inventory Working Group, the Mexican Association of Professional Foresters, the Mexican Forest Service, the USDA Forest Service and the USDI Bureau of Land Management. The dates are November 30-December 6, 1980 and the place - La Paz, Mexico.

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

In recent years, demands on arid lands have been increasingly attractive for recreation, urban development, archeological searching, domestic livestock grazing and wood products. To insure protection and proper management of amenity and non-amenity values, inventories of the basic attributes of arid lands are required. The relative low economic values of the lands and their remoteness makes inventory design a challenge.

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

Tentative Outline and Schedule

November 30	Registration
December 1	Welcome and Introductions Arid Land Characteristics, Resources and Uses Inventory Challenges
December 2	Meeting the Challenges Inventory Planning Classification Schemes Economical Mapping Systems
December 3	Field Trip
December 4	Meeting the Challenges (continued) Cost Efficient Sampling Schemes Efficient Measuring Techniques
December 5	Meeting the Challenges (continued) Resource Data Analyses System Where Do We Go From Here? Proposals for Exchanging Research
December 6	Business Meetings SAF and IUFRO 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 Driscoll
USDA Forest Service, RMF & RES
240 West Prospect Street
Fort Collins, Colorado 80526, USA

All contributions must be received by November 30, 1979.

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, 1980.

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

July, 1979 marked the 4th anniversary of RESOURCE INVENTORY NOTES. The "NOTES" were started by the U.S. Forest Service in 1974 and were continued by the Bureau of Land Management starting in September, 1976. The "NOTES" started with an initial mailing of 595. Now over 2,850 copies are received in some 124 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.

USDA Forest Service

No. 1 - July 1975 - Introduction

No. 2 - September 1975 - Definitions

No. 3 - November 1975 - Regional, Management Based and Intensive Inventories

No. 4 - January 1976 - Inventory Planning

No. 5 - March 1976 - A Computer 3P Game by Goldsmith, Russell, and Barrett

No. 6 - May 1976 - Information Needed and the Design of an Inventory by Johnson

No. 7 - July 1976 - 4 PEA Sampling by Lund and LaBau

USDI Bureau of Land Management

BLM 1. September 1976 - 3P or Not 3P by Lund

BLM 2. November 1976 - A Test of the Statistical Validity of a 3P and Point Sampling Design by Wiant and Fountain

BLM 3. January 1977 - Inventorying the Urban Forest by Geiger

BLM 4. March 1977 - Mesavage and Girard's Volume Tables Formulated by Wiant and Castaneda

BLM 5. May 1977 - Sampling Precision and Probability by Kinsinger

BLM 6. July 1977 - An Illustration of List Sampling by Wiant

BLM 7. September 1977 - Forest Site Index Mapping and Yield Model Inputs to Determine Potential Site Productivity by Getter and Creighton

BLM 8. November 1977 - Comparison of Point - 3P Sampling Designs by Wiant

BLM 9. January 1978 - Some Basic Considerations When Sampling Small Woodlands by Ashley

BLM 10. March 1978 - Stratification in Double Sampling -- The Easy Way Out May Sometimes Be the Best Way by Frayer

BLM 11. May 1978 - 3P Random Numbers and a Handheld Programmable Calculator by Estola

BLM 12. July 1978 - On the Precision of Dot Grid Estimates by Zohrer

BLM 13. September 1978 - Inplace, Multiple Resource Inventories at Budget Prices by Lund

BLM 14. October 1978 - Modification of Freese's Chi-Square Test of Accuracy by Rennie and Wiant.

BLM 15. November 1978 - Type Maps, Stratified Sampling and P.P.S. by Lund

BLM 16. December 1978 - A Technique for Combining Related Regressions Into One Equation by Wiant

BLM 17. January 1979 - Variable Radius Plot and 3P Timber Sampling by Estola

Height Accumulation for Programmable Calculators by Hanson

BLM 18. February 19 - Multi-Stage and Multi-Phase Sampling by Nichols

Preliminary Pinon - Juniper Volume Tables by Estola

BLM 19. March 1979 - A Distribution-Free Method for Internal Estimation and Sample Size Determination by Fowler and Hauke

Converting Outside Bark to Inside Bark Diameters by Castaneda

BLM 20. April 1979 - Canada Committee on Ecological Land Classification by Welch and Wiken

The Precision of Dot Grid Estimates: A Theoretical Approach by Chevreu

BLM 21. May 1979 - Metrification of Mesavage's Form Class 78 and 80 Cubic-Foot Volume Tables by Hassler and Fountain

BLM's Standard, Non-Standard, Stand Inventory System by Costello and Lund

BLM 22. June 1979 - Preparation of Maps for Manual Digitizing by Hanson

Unequal Probability Sampling With Replacement and Without Replacement by Oderwald, Wellman and Buhyoff

Uniformly Distributing Samples Within a Type Island by Lund

BLM 23. July 1979 - Sampling Natural Resource Populations: Mutually Exclusive Fixed-Area Sampling Units by Fowler and Davis.

Computing Optimum Plot Size for Wildland Inventories by Taaffe

The Satellite Program in Statistical Ecology by Patil

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

Wanted! Lead articles, current literature and meeting announcements for publishing in the "NOTES". If announcing a meeting, please allow at least a four-month lag time.

- - - - - - - - - - - -

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

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

UNITED STATES
DEPARTMENT OF THE INTERIOR
BUREAU OF LAND MANAGEMENT
DENVER FEDERAL CENTER
BUILDING 50
DENVER, COLORADO 80225

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

POSTAGE AND FEES PAID
U. S. DEPARTMENT OF THE INTERIOR
Int 415

U.S.MAIL

T GREGOIRE 01 002841
I N E R JAMES HALL
UNIVERSITY OF NEW HAMPSHIRE
DURHAM NH 03824