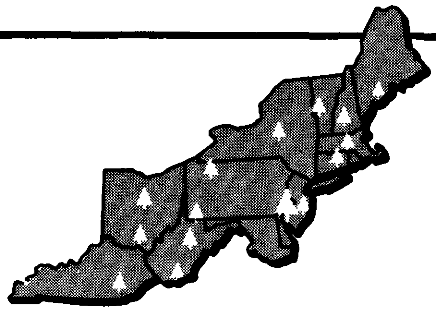
USDA FOREST SERVICE RESEARCH NOTE NE-195

1974

Northeastern Forest Experiment Station



ODC 5--015.5

FOREST SERVICE, U.S. DEPT. OF AGRICULTURE, 6816 MARKET STREET, UPPER DARBY, PA. 19082

WEIGHTED REGRESSION ANALYSIS AND INTERVAL ESTIMATORS

Abstract.—A method for deriving the weighted least squares estimators for the parameters of a multiple regression model. Confidence intervals for expected values, and prediction intervals for the means of future samples are given.

Weighted regression analysis is applicable to many forestry problems. Cunia (*1964*) discusses in detail weighted regression analysis and analyzes the relationship between volume and diameter in black spruce. Many elementary statistics books cover weighted regression analysis, but generally there is little or no discussion of interval estimators. The purpose of this paper is to discuss: (1) confidence intervals for expected values; (2) prediction intervals for means of future samples when the parameters of a multiple regression model are estimated by weighted least squares.

THE REGRESSION MODEL

Suppose we have a sample of n individuals. The model for i^{th} observation is

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{pi}\beta_p + e_i$$

or

$$y_i = x_i\beta + e_i. \qquad (1)$$

The set of n observations can be succinctly written

$$y = X\beta + e \qquad (2)$$

where y is the nx1 vector of observations, X is the nxp design matrix, β is a px1 vector of unknown constants, and e is the nx1 vector of errors. It is also assumed that the errors are normally distributed with a mean and variance-covariance matrix

$$E(e) = 0$$

$$\Sigma_e = E(e\,e') = V\sigma^2$$

where V is a known positive definite matrix. σ^2 is a positive scalar and is assumed to be unknown. In many cases V is a function of X.

The form of V depends on the variances and covariances of the observations. Suppose that the errors have unequal variances and are mutually independent. In this case, the variance-covariance matrix of the errors is

$$V\sigma^2 = \begin{bmatrix} v_1 & & & 0 \\ & v_2 & & \\ & & \cdots & \\ 0 & & & v_n \end{bmatrix}\sigma^2.$$

V can be written in the form $V = P'P = PP = P^2$.

Since V is diagonal we have $P = V^{1/2}$ and $P^{-1} = V^{-1/2}$.

WEIGHTED LEAST SQUARES ESTIMATORS

Weighted least squares estimators can be obtained by transforming the original observations to variables that satisfy the assumptions for ordinary least squares. Pre-multiplying equation (2) by P^{-1} gives

$$P^{-1} y = P^{-1} X\beta + P^{-1} e$$

or

$$Z = Q\beta + f. \qquad (3)$$

Equation (3) is an ordinary multiple regression model; that is,

$$E(f) = 0 \text{ and } E(f f') = I\sigma^2.$$

Unweighted least squares theory can be applied directly to the transformed model. The sum of squares of the transformed errors is

$$\begin{aligned} S &= f'f \\ &= e'V^{-1}e. \end{aligned}$$

Since V is diagonal, the sums of squares can be written as

$$S = \Sigma_i v_i^{-1} e_i^2.$$

Hence S is the weighted sums of squares of the errors.

Weighted least squares estimates are obtained by minimizing S for variation in b where b is the solution vector corresponding to the parameter vector β. The weighted normal equations are

$$Q'Qb = Q'Z$$

or

$$X'V^{-1}Xb = X'V^{-1}y.$$

Solving the weighted normal equations for b gives the weighted least squares estimator. The solution is

$$\begin{aligned} b &= (Q'Q)^{-1}Q'Z \\ &= (X'V^{-1}X)^{-1}X'V^{-1}y. \end{aligned}$$

The sums of squares of transformed residuals is

$$\begin{aligned} SSR &= Z'Z - b'Q'Z \\ &= y'V^{-1}y - y'V^{-1}X \\ &\quad (X'V^{-1}X)^{-1}X'V^{-1}y. \end{aligned}$$

The sample variance is

$$s^2 = SSR / (n-p)$$

which is an unbiased estimator of σ^2. The variance-covariance matrix of b is

$$\begin{aligned} \Sigma_b &= (Q'Q)^{-1}\sigma^2 \\ &= (X'V^{-1}X)^{-1}\sigma^2. \end{aligned}$$

The sample variance-covariance matrix of b is

$$\begin{aligned} \hat{\Sigma}_b &= (Q'Q)^{-1}s^2 \\ &= (X'V^{-1}X)^{-1}s^2. \end{aligned}$$

Given V, values of b, s^2, and $\hat{\Sigma}_b$ are easily obtained by ordinary least squares analysis on the transformed variables z and q.

PREDICTED VALUES AND INTERVAL ESTIMATORS

Consider a future sample of k independent observations (y_1^o, x_1^o), (y_2^o, x_2^o), . . . , (y_k^o, x_k^o). The average of the future values of the dependent variable is

$$\bar{y}^o = \Sigma_{j=1}^k y_j^o / k.$$

The statistic $\bar{y}^o$ is normally distributed with a mean and variance

$$E(\bar{y}^o) = \bar{x}^o\beta$$

and

$$\sigma^2(\bar{y}^o) = \Sigma_{j=1}^k \bar{v}_j \sigma^2$$

The sample estimator of $\sigma^2(\bar{y}^o)$ is

$$s^2(\bar{y}^o) = \bar{v}\, s^2 / k.$$

Now consider the predicted values of dependent variable based on the regression estimates. The prediction for the value of the jth future observation is

$$y^*_j = x^o_j b.$$

The average of the regression estimates is

$$\bar{y}^* = (1/k)\, \Sigma_{j=1}^k\, y^*_j = \bar{x}^o b$$

where $\bar{x}^o = (1/k)\, \Sigma_{j=1}^k x_j^o$

= average of the vectors x_j^o.

The statistic $\bar{y}^*$ is normally distributed with a mean and variance

$$E(\bar{y}^*) = \bar{x}^o\beta$$

and

$$\sigma^2(\bar{y}^*) = \bar{x}^o \Sigma_b \bar{x}^{o\prime}.$$

The sample estimator of $\sigma^2(\bar{y}^*)$ is

$$s^2(\bar{y}^*) = \bar{x}^o\, (Q'Q)^{-1} \bar{x}^{o\prime} s^2.$$

Under the assumptions of the model the statistic

$$t = (\bar{x}^o b - \bar{x}^o \beta)/s\ (\bar{y}^*)$$

has a Student's t distribution with n-p degrees of freedom. Consequently, the confidence interval for the expected value of the average of the regression estimates is obtained from the probability statement

$$P(\bar{x}^o b - ts\ (\bar{y}^*) \leq \bar{x}^o\ \beta\ \leq \bar{x}^o b + ts\ (\bar{y}^*) = 1\text{-}\alpha$$

where $\quad t = t_{1-\alpha/2,n\text{-}p}$. (4)

We are also interested in a prediction interval for the future mean $\bar{y}^o$. The prediction interval gives on a probability basis the range of error of the future mean.
Let $d = \bar{y}^o - \bar{y}^*$. The statistic d is normally distributed with a mean and variance of

$$E(d) \ = E(\bar{y}^o) - E(\bar{y}^*) = \bar{x}^o\beta - \bar{x}^o\beta = 0$$
$$\sigma^2(d) = \sigma^2(\bar{y}^o) + \sigma^2(\bar{y}^*)$$
$$= \sigma^2(\bar{v}/k + \bar{x}^o(Q'Q)^{-1}\bar{x}^{o\prime}).$$

The sample estimator of $\sigma^2(d)$ is

$$s^2(d) = s^2(\bar{v}/k\ +\ \bar{x}^o(Q'Q)^{-1}\bar{x}^{o\prime}).$$

Note that the statistics $\bar{y}^o$ and $\bar{x}^o b$ are statistically independent since they are based on independent samples.

It follows from the assumptions that the quantity d/s_d has a Student's t distribution with n-p degrees of freedom. Therefore the prediction interval for the future mean $\bar{y}^o$ given $(x_1^o, x_2^o, \ldots x_k^o)$ can be calculated from the probability statement

$$P(\bar{x}^o b - ts_d \leq \bar{y}^o \leq \bar{x}^o b + ts_d) = 1\text{-}\alpha. \quad (5)$$

Several examples of situations where these types of intervals arise follow.

I. In regression analysis the confidence intervals for the expected value of y given x^o can be calculated for several values of x^o. The upper and lower confidence limits are often plotted about the estimated regression line. Also, it is a common practice to plot the prediction interval for one future value given x^o. In this case, the quantities $\bar{x}^o$ and $\bar{v}$ in the interval estimators are replaced by x^o and v respectively. The vector x^o is single vector of specific values say $(x_1^o, x_2^o, \ldots x_p^o)$ and v is the weight associated with x^o.

II. Consider a population of a large number (N) of trees where the trees are measurable for volume (y) and diameter at breast height (d). A random sample of n trees is measured and a parabolic regression $E\ (y) = \alpha + \beta d + \gamma d^2$ is estimated by weighted least squares. Suppose that sometime in the near future, k trees of preselected diameters are to be sampled from the forest. The sample will consist of k_1 trees of diameter d_1, k_2 trees of diameter $d_2, \ldots, k_s$ trees of diameter d_s. Also assume that the size of the future sample k is small in respect to the number of trees in the population N.

We are interested in

(1). A point estimator of the average volume $\bar{y}^*$. The estimator is

$\bar{x}^o b = (1, \bar{d}, \overline{d^2})\ (1, \hat{\beta}, \hat{\gamma})'$

where $\bar{d} = \Sigma_{i=1}^{s}\ k_i d_i/k$,

$d^2 = \Sigma_{i=1}^{s}\ k_i d_i^2/k$, and $\hat{\beta}$ and $\hat{\gamma}$ are the weighted least squares estimates of β and γ.

(2). The 1-α confidence interval for the expected value of $\bar{y}^*$ which is obtained from equation (4).

(3). The 1-α prediction interval for the future mean ($\bar{y}^o$) which is obtained from equation (5).

III. Consider the case where an entire forest is harvested. All values of the vector x are measured. Let u be the mean of all vectors x_j. A small random sample of observations (y, x) are measured and the weighted parabolic regression is estimated.

(1). The multiple regression estimator of the average volume is u b.

(2). The confidence interval for u b is is given by

$u\,b \pm t_{1-\alpha/2,\ n\text{-}p}\ [u\ (Q'Q)^{-1}u's^2]^{1/2}$

IV. It may be too expensive to measure the diameter of all the trees. Instead, the diameter is measured on a second large independent sample. The double sampling estimator of u β is $\bar{x}_d b$ where $\bar{x}_d$ is the mean vector from the second sample. Equation (5) is not the proper expression for the confidence interval for $\bar{x}_d \beta$ because $\bar{x}_d$ is a random vector. An approximation of $\sigma^2(\bar{x}_d b)$ is given by Sen (1973).

EXAMPLES

Cunia (*1964*) found a curvilinear relationship between the volume and diameter in black spruce. The relationship can be written

$$y_i = \alpha + \beta d_i + \gamma d_i^2 + e_i \qquad (6)$$

where y_i is the volume and d_i is the diameter of the i^{th} tree. He also found that the variance of the volume can reasonably be assumed to be proportional to the fourth power of the diameter. Assuming the errors are independent, the variance-covariance matrix is

$$\Sigma = \begin{bmatrix} d_1^4 & & & 0 \\ & d_2^4 & & \\ & & \cdots & \\ 0 & & & d_n^4 \end{bmatrix} \sigma^2.$$

The matrices P and P^{-1} are

$$P = \begin{bmatrix} d_1^2 & & & 0 \\ & d_2^2 & & \\ & & \cdots & \\ 0 & & & d_n^2 \end{bmatrix} \text{and } P^{-1} = \begin{bmatrix} d_1^{-2} & & & 0 \\ & d_2^{-2} & & \\ & & \cdots & \\ 0 & & & d_n^{-2} \end{bmatrix}$$

Premultiplying the set of observations by P^{-1}, results in a set of equations whose i^{th} row is

$$y_i/d_i^2 = \alpha(1/d_i^2) + \beta\,(1/d_i) + \gamma(1) + e_i/d_i^2$$

$$\text{or} \quad z_i = \alpha q_{1i} + \beta q_{2i} + \gamma q_{3i} + f_i \qquad (7)$$

The first step in the weighted regression analysis is to transform the data. Cunia (*1964: Table 3*) gives diameters and volumes for 25 black spruce. The original measurements have the form

Tree No.	*Diameter (d)* *(inches)*	*Volume (y)* *(cubic ft.)*
1	3.9	1.0
2	4.1	1.6
—	—	—
—	—	—
—	—	—
25	12.7	25.4

The transformed values needed for analysis for the weighted multiple regression are

Tree No.	$q_1 = 1/d^2$	$q_2 = 1/d$	$q_3 = 1.0$	$z = y/d^2$
1	.065740	.2564	1.0	.0657
2	.059487	.2439	1.0	.0952
—	—	—	—	—
—	—	—	—	—
—	—	—	—	—
25	.006193	.0787	1.0	.1575

The transformed data can be analyzed with any multiple regression program. Most regression programs print b, s^2, and $(Q'Q)^{-1}$. Computer programs with the option for testing hypotheses of the form $x\beta = d$ should also print the quantities xb and $x(Q'Q)^{-1}x'$.

An ordinary least squares analysis of the transformed values was done with BIOMEDX63.* The statistics of interest are

$$b' = (\hat{\alpha}\ \hat{\beta}\ \hat{\gamma}) = (1.19040, -0.76579, 0.19638)$$

$$(Q'Q)^{-1} = \begin{bmatrix} 6218.83 & -2024.18 & 145.87 \\ -2024.18 & 672.12 & -49.59 \\ 145.87 & -49.57 & 3.79 \end{bmatrix}$$

and $s^2 = 0.0053/22 = 0.000241$.

Confidence Intervals for $x\beta$ and Prediction Intervals for one future observation y^o

Suppose we want the expected volume and interval estimates for a 10 inch diameter black spruce. The estimate is xb = (1, 10, 100) b = 13.17070 cubic feet. To compute the .95 confidence interval we also need

$$x(Q'Q)^{-1}x' = 912.12102$$

and $t_{.975,22} = 2.074$.
Then from equation (4) we have

$$P(12.20 \leq \alpha + 10\beta + 100\gamma \leq 14.14) = .95$$

To compute the prediction interval for volume of a single future observation for a 10 inch black spruce we need

$$\bar{v}/k = d^4/1$$
$$= 10^4$$

The prediction interval is

$$P(9.81 \leq y^o \leq 16.53) = .95.$$

* See BIOMEDX63-multivariate general linear hypothesis. Univ. Cal. Publi. in Autom. Comput. 3. W. J. Dixon, Editor. Univ. Cal. Press. 1969.

The sample regression line, confidence intervals, and prediction intervals for one future value are shown in Fig. 1. Note the increasing width of confidence and prediction intervals with increasing diameter. The flare in the intervals is the result of assumption that the variance of the volume is proportioned to the fourth power of the diameter.

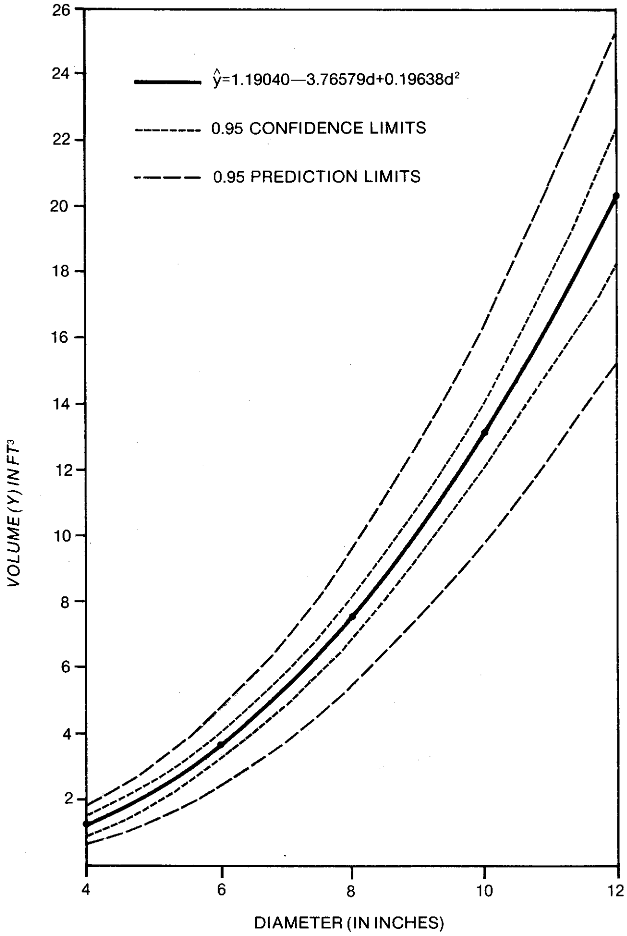


Figure 1.—Weighted regression line, 0.95 confidence intervals, and 0.95 prediction intervals based on the sample of 25 black spruce.

Multiple Regression Estimate

Cunia (*1964: Table 2*) gives the following diameter distribution for 1188 black spruce.

Diameter	*Number*
4	156
5	321
6	265
7	130
8	146
9	84
10	19
11.5	51
13.5	12
15.5	4

The mean vector is $u = (1,\bar{d},\bar{d}^2) = (1, 6.44, 45.77)$. The estimated mean volume for this population of trees is 5.247 cubic feet per tree. The variance of ub is $u(Q'Q)^{-1}u's^2 = 0.023208$. and the 0.95 confidence interval for $u\beta$ is

$$5.247 \pm 2.074\ (.023208)^{1/2}$$
$$= 5.247 \pm 0.315957$$

The examples show that weighted regression is no more difficult than ordinary least square analysis. Interval estimates are easily obtained from ordinary multiple regression analysis of the transformed data. Special computer programs are not needed.

LITERATURE CITED

Cunia, T.
1964. WEIGHTED LEAST SQUARE METHOD AND CONSTRUCTION OF VOLUME TABLES. Forest Sci. 10(2): 180-191.

Sen, A. R.
1973. THEORY AND APPLICATION OF SAMPLING ON REPEATED OCCASIONS WITH SEVERAL AUXILIARY VARIABLES. Biometrics 29(2):381-385.

—DONALD W. SEEGRIST
Biological Statistician
USDA Forest Service
Northeastern Forest Experiment Station
Upper Darby, Pa.

MANUSCRIPT RECEIVED FOR PUBLICATION 13 SEPTEMBER 1973