

TABLE OF CONTENTS

PART I: TUTORIAL PAPERS

Combining the Error of Sample Plots and Biomass Regressions

- Error of forest inventory estimates: its main components 1
Tiberius Cunia
- An optimization model to calculate the number of sample trees and plots 15
Tiberius Cunia

Error of Biomass Regressions

- Construction of tree biomass tables by linear regression techniques 27
Tiberius Cunia
- Use of dummy variables techniques in the estimation of biomass regressions 37
Tiberius Cunia
- On the error of tree biomass regressions: trees selected by cluster sampling and double
sampling 49
Tiberius Cunia

Error of Sample Plots

- On the error of forest inventory estimates: stratified sampling and double sampling for
stratification 63
Tiberius Cunia
- On the error of forest inventory estimates: two-stage sampling of plots 71
Tiberius Cunia
- On the error of forest inventory estimates: double sampling with regression 79
Tiberius Cunia
- On the error of forest inventory estimates: Continuous Forest Inventory without SPR 89
Tiberius Cunia
- On the error of forest inventory estimates: Continuous Forest Inventory with SPR 99
Tiberius Cunia

PART II: RESEARCH PAPERS

Biomass Regressions and Measurement Error

- An optimization model for subsampling trees for biomass measurement 109
Tiberius Cunia
- Estimating sample tree biomass by subsampling: some empirical results 119
R. D. Briggs, T. Cunia, E. H. White, and H. W. Yawney
- Unbiased estimation of total tree weight by three-stage sampling with probability
proportional to size 129
Harry T. Valentine, Timothy G. Gregoire, and George M. Furnival
- Measurement errors in forest biomass estimation 133
Daniel Auclair

Biomass of Forest Understory Vegetation

- Biomass-dimension relationships of understory vegetation in relation to site and stand
age 141
Paul B. Alaback

TABLE OF CONTENTS

Biomass estimates for nontimber vegetation in the Tanana River Basin of Interior Alaska	149
Bert Mead, John Yarie, and David Herman	

Biomass Functions in the Eastern United States: Regression Models and Application to Timber Inventories

A summary of equations for predicting biomass of planted southern pines	157
V. C. Baldwin, Jr.	
Summary of biomass equations available for softwood and hardwood species in the southern United States	173
Alexander Clark III	
Methods for estimating the forest biomass in Tennessee Valley Region	189
J. Daniel Thomas and Robert T. Brooks, Jr.	
Areas of biomass research ¹	193
Boris Zeide	

Biomass Studies Outside the United States

Prediction error in tree biomass regression functions for western Canada	199
T. Singh	
Forest biomass studies in France	209
Daniel Auclair	
Biomass studies in Europe - an overview	213
Dieter R. Pelz	
Subsampling trees for biomass	225
C. Kleinn and D. R. Pelz	
Simple biomass regression equations for subtropical dry forest species	229
Joseph D. Kasile	

Use of Simulation Techniques to Evaluate the Validity of Biomass Regression Functions

Evaluating errors of tree biomass regressions by simulation	235
Tiberius Cunia	
Estimation of tree biomass tables by cluster sampling: results of a simulation study	243
Andrew J. Gillespie and Tiberius Cunia	
Error of biomass regressions: sample trees selected by stratified sampling	253
Alexandros Arabatzis and Tiberius Cunia	
Error of biomass regressions: sample trees selected by double sampling	269
John Michelakackis and Tiberius Cunia	
Using simulation to evaluate volume equation error and sampling error in a two-phase design	287
David C. Chojnacky	
High order regression models for regional volume equations	295
Joe P. McClure and Raymond L. Czaplewski	

¹Contributed paper, not presented at the workshop.

TUTORIAL PAPERS

Error of Biomass Regressions

Tiberius Cunia

Professor of Statistics and Operations Research,
State University of New York College of Environ-
mental Science and Forestry, Syracuse, NY, 13210

The paper describes the weighted least squares method of linear regression and its application to the problem of estimating tree biomass regressions and their error. It discusses in detail the four main sources of error (tree selection, biomass measurement of sample trees, statistical model used and application of biomass regressions) and it shows how to calculate the three basic statistics of linear regression functions that are needed in forest inventory, namely [b], the estimate of the vector of regression coefficients, $S_{yy|x}$, the estimate of the conditional variance of y about the regression function and $[S_{bb}]$ the estimate of the covariance matrix of [b]. The procedure is then applied to a sample of trees to construct a biomass table on tree diameter, together with its 95 percent confidence and prediction intervals.

Introduction

In a previous paper, T. Cunia (1986a) proposed a procedure for combining the error of the tree biomass regression function with that from the sample plots, when the error of the forest inventory estimates is being calculated. This procedure requires that (i) the tree biomass regression function be of a linear form, that is

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]' [x]$$

where $x_1 = 1$, x_2, x_3, \dots, x_m are the independent variables and notation [] and []' is used to denote matrices (and vectors) and their transposes respectively and (ii) valid estimates [b] of [β] and $[S_{bb}]$ of the covariance matrix of [b] be available.

It is the objective of the present paper to discuss the problem of error of biomass regressions, identify its main sources and present the procedure for the estimation of this error by the method of least squares linear regression. It will be assumed here that the classical assump-

tions of this method are satisfied. For more details on the procedure of combining the error of the biomass regression with that of the sample plots the reader is referred to the above-mentioned paper.

Main Sources of Error

There are four major sources of error in tree biomass regression functions. The first is that associated with the selection of sample trees. The same selection procedure applied to the same tree population on different occasions results in different sets of sample trees and, thus, in different biomass regressions. The second source of error is the measurement of the sample trees. Different measurements of the same sample trees, made on different occasions by various individuals, do not ordinarily yield the same measurements. While the volume of the merchantable bole of a felled tree, or the green weight of the entire tree can be measured without any appreciable error, the measurement of the oven-dry biomass of a given tree component is usually based on subsampling; small parts of the tree component are selected by some random procedure, their dimensions are measured in the field and the value of their oven-dry biomass determined in the laboratory is used to estimate the biomass of the given tree component. The difference between the true, conceptual value of the tree attribute we want determined and the actual, recorded value obtained by measurement is known as measurement error. The third source of error is that of the statistical model used. Given the same set of sample tree data, different statisticians may use different models and, thus, obtain different biomass regressions. The basic assumptions of the model must be satisfied by both, sample and population, in order for the conclusions to be valid. Finally, the fourth source of error is associated with the application of the biomass regressions to a specific forest inventory. This error component may become extremely important when the regression is applied to a forest population that is very different from that for which it was estimated. Strictly speaking, the biomass regressions are never applied to the original population; the populations of trees are dynamic and change with time.

Let us now consider, in more detail, each source of error separately.

Sample Tree Selection

Because the trees of the forest cannot be all measured for biomass, one must rely on sampling. Any sample of trees, selected by any procedure, can be used to calculate a biomass regression. But only when the selection procedure is statistical and, properly applied, leading to a representative sample of the population of interest, can the resulting regressions be statistically valid.

We say that a sample is representative of a population of interest when the error of the bio-

^{1/}Based on the paper "On volume tables and their contribution to the error of forest inventory estimates" In: Forstliche Nationalinventuren in Europa (National Forest Inventory in Europe) - D. R. Pelz and T. Cunia, (Eds.) Mitteilungen der Abteilung Für Forstliche Biometrie Universität Freiburg, 7800 Freiburg I. Br. - Federal Republic of Germany.

mass regression function calculated from that sample and applicable to that population can be evaluated,, at least conceptually, in quantitative statistical terms. It is not necessary for the estimates of the error to be precise and unbiased; it suffices that these estimates be sufficiently good under the statistical model assumed and that the basic model assumptions be sufficiently well satisfied.

In general, a sample is representative if, for the given selection procedure, each element of the population has a known non-zero probability of selection. It is not necessary for the probability to be known in absolute terms. Knowing, for example, that the elements are selected with equal probability or probability proportional to a measure of tree size is sufficient. However, the probability must be non-zero. For example, if the biomass regression is to be applied to all forest trees in the area of interest, one should not limit the selection of sample trees from those that are dominant and codominant, healthy with no apparent defects, etc.

In addition to being representative, the selection procedure must also be such that the resulting samples could be subjected to valid statistical analysis. Furthermore, it must also be cost-efficient, that is, yield biomass regression functions of acceptable precision at reasonable costs. This implies that one should not use a selection procedure that yields representative samples but for which statistical techniques for a valid analysis have not been devised yet, nor should one use a procedure that is too expensive for all practical purposes.

A search of the forest biomass literature made by Cunia (1979a,b) showed that many authors did not state how the sample trees were selected and from what type of tree population. Or, when they did, the procedure was non-random in the statistical sense; the sample trees were selected from subjectively selected parts of the given forest area, trees of odd shapes or defoliated were discarded from the sample, etc. Finally, when the sampling method was indeed statistical, the biomass regressions were not properly estimated. Sample trees selected by cluster or stratified sampling were analyzed by the ordinary least or weighted least squares method as if they were selected by simple random sampling.

Biomass Measurement of Sample Trees

Once the sample trees are selected, they must be measured for biomass. The biomass is usually required by components; either by major components such as main bole, crown or stump-root system, or minor components such as wood or bark of dead branches, wood of live branches larger than 10 cm. of diameter, bark of merchantable bole, etc. All these components require prior and very precise definition. A search of the literature shows a wide variety of definitions. There is a great need for standardization of terminology, if one is to combine results from the analysis of sample data from various sources

or compare biomass regressions published by various authors.

Assuming that a tree component has been previously defined, it is relatively easy to measure the fresh weight; it can be weighed in one piece or in smaller sections. The only sources of error may be in the physical separation of the component from the tree (with the possible loss of small particles), the actual weighing by mechanical devices and the possible loss of some humidity (between time of harvesting and time of measurement). The measurement of the volume of the components like the main bole that have sections easily approximated by geometrical bodies is also relatively precise. Much more difficult is the problem of measuring the volume of components such as leaves, small roots or branches, etc. or the ovendry weight of any component of large size. In all these cases one must rely on subsampling.

Consequently, the value assigned to a sample tree is most of the time an estimate, not the true value of the tree biomass. As such, it has a subsampling bias and random error. Many of the techniques we know for estimating the tree biomass have an inherent statistical bias, which is ordinarily small if the subsampling is done properly and the sample size is sufficiently large. The random part of the error depends on the method of subsampling, technique of estimation and sample size. When the biomass regressions are calculated by the statistical least squares methods, the random part of the subsampling error is automatically taken into account. This in itself is not a problem. Assuming that the subsampling bias is small, the real problem is that of efficiency. Is it better to have a small sample of trees for which the biomass is precisely estimated, or it is better to have a large sample of trees for which the ovendry biomass estimates have low precision? One way to optimize the combined sizes of the sample of trees and the subsamples within the trees is given in Cunia (1986 b).

Statistical Models

Seldom, if ever the assumptions of a statistical model are satisfied. This in itself is not unusual; it is part of the whole process of solving real world problems by abstracting them first as mathematical models and by solving then these models. The solutions found for the models are assumed to apply equally well to the real world problems they represent.

The errors associated with the mathematical model used are generally small when the model fits well the real world problem. It may become quite large if the basic assumptions of the model are critically violated, either by the population of trees or by the sample drawn from this population. Because we are concerned here with a regression model, let us discuss the basic assumptions of the least squares method and the effect these assumptions have on biomass regressions when they are not satisfied. For a more detailed

discussion of this problem, the reader is referred to Cunia (1979a,b).

There is first the assumption that the true regression function is of the assumed form. Because one deals with finite populations of trees, this assumption is never fulfilled in the strict statistical sense. But this in itself is not a serious drawback. Provided one works only with forms that were shown to be good, and he has no need to extrapolate the application of the estimated regression beyond the sample data, the effect of this assumption is minimal. In particular it does not seem to matter much whether the form of the regression function is linear or non-linear. This is important in view of the controversy going on between biologists who prefer working with non-linear functions of the allometric form and statisticians who most of the time prefer working with linear regressions.

It is true that some biological arguments can be brought in favor of allometric functions; the increase in biomass of a growing tree is proportional to the biomass contained in the tree. But it is also true that the form of the regression function is somewhat affected and greatly obscured by the inherent variation of biomass values from tree to tree. Furthermore, mathematical arguments based on Taylor's Theorem prove that one can always find linear functions that can approximate as closely as desired many non-linear functions as long as this approximation refers to a finite range. Consequently, the decision on whether to use linear or non-linear regression functions must be based entirely on different considerations.

The linear functions have several main advantages. Most of the statistical theory has been developed for linear regressions only. The least squares method is well known, it is simple to apply and has been extended to cover cases such as piecewise linear functions, harmonization and additivity of regression functions of component parts of tree biomass, selection of sample trees by methods other than simple random sampling, etc. Another big advantage is that the error of the biomass regression can be expressed in a convenient form that makes the method to combine it with the error of forest inventory sample plots relatively easy and straightforward to apply. Finally, there is enough empirical evidence to suggest that linear or non-linear functions properly selected are equally good. The only disadvantage one can think of, is that the non-linear regression functions may be better to use when they are applied beyond the range of the sample data; they seem to yield better estimates of the biomass when extrapolated.

The second assumption that the conditional variance of the tree biomass is homogeneous is generally not critical; forest biometricians know how to modify the least squares techniques when this assumption is not satisfied. One way is to transform the variables. Another way is to use the weighted least rather than the least squares method. But the assumption may become critical when (i) one uses \log transformations or

(ii) no correction for bias is made when proper transformations are used or finally, when (iii) wrong weights are used with the weighted least squares method.

The third assumption that the conditional distribution of the tree biomass is normal (in the statistical sense) is not critical unless (i) the sample of trees is relatively small, say 10-15 trees, in which case the central limit theorem does not necessarily apply and the confidence intervals and the results of significance tests may not be valid, or (ii) prediction limits for the biomass of individual trees are desired, in which case the calculated limits will be symmetrical when it is well known that the conditional probability distribution of the tree biomass is highly skewed. Otherwise the regression results are all right; the shape of the conditional distribution is not needed (as an assumption) when the point estimates of the regression coefficients and their covariance matrix are calculated, and with sufficiently large samples, say, at least 20-25 trees, the inferences based on normal distribution (about tests of significance or confidence intervals) are valid, since the central limit theorem applies sufficiently well.

The last assumption is that of statistical independence among the biomass measurements of the sample trees. It is an important assumption because the sample trees are almost never selected completely at random to satisfy this assumption. It is much less expensive to select trees in clusters and much more efficient to use stratification by size (tree diameter) or geographical location (site quality class, forest type and age class, etc.). To see what may happen when the sample trees are selected by methods other than simple random sampling (that insures statistical independence), and the ordinary least or weighted least squares are still applied in their standard form to calculate biomass regressions, one may refer to the empirical results and the results obtained by simulated sampling reported by Cunia (1981, 1986c). For some of the modifications to make to the least squares method when the sample trees are selected by methods other than random sampling, the reader is referred to Cunia (1986d).

Application of Biomass Regressions

The error component associated with the application of the biomass regression functions to a specific forest area is practically zero, when the regressions were estimated from trees properly selected from that specific area. The possible error is negligibly small even though tree populations are dynamic in nature and change with time; and, thus, the population from which the sample trees were selected is no longer the same population when the resulting biomass regressions are applied. The error is also negligibly small when the regression was estimated from a population similar to that being inventoried. However, this error component may be quite important in size for the forest inventory

systems for which old biomass regressions are being used; regressions calculated a long time before, from subjectively selected trees. When this occurs it is impossible to estimate the size of this error component.

Least and Weighted Least Squares Linear Regression Method

Before describing the specific weighted least squares procedure as applied to biomass regression estimation, we shall briefly describe the ordinary least squares method as applied to the estimation of linear regression functions. This will introduce the main concepts and notation used here and the three basic formulae for the three main statistics that summarize all the information from the sample tree data that is usually needed in forest inventory.

Least Squares Method

The basic assumptions of the least squares linear regression method were discussed in the previous section. We shall now be more specific and state them more formally as follows

(1) The expected value of y (the dependent variable) for given values of x_1, x_2, \dots, x_m (the independent variables) is of the linear form

$$E(y|x) = \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]' [x]$$

= regression function of y on $[x]$

where y = biomass of some tree component

x_1, x_2, \dots, x_m = tree characteristics other than biomass; such as, for example, tree diameter, height, species (dummy variable) etc. with $x_1 = 1$ being the variable that introduces a constant term (the intercept) in the regression function

$[x]' = [x_1 \ x_2 \ \dots \ x_m]$ = vector of the fixed variables x

$[\beta]' = [\beta_1 \ \beta_2 \ \dots \ \beta_m]$ = vector of the regression coefficients

and $[]'$ = notation used to denote transposed vectors or matrices $[]$.

(2) The conditional variance of y given $[x]$, denoted here as $\sigma_{yy|x}$ is homogeneous, that is, the variation of y about the regression function remains the same no matter what the given values x_1, x_2, \dots, x_m are. More formally,

$$E(y - E(y|x))^2 = \sigma_{yy|x}, \text{ a constant value}$$

(3) The covariance of any two sample values y_i and y_j is equal to zero, that is, the random variables y_i and y_j are uncorrelated. More formally

$$E(y_i - E(y_i|x_i))(y_j - E(y_j|x_j)) = 0$$

(4) The conditional probability distribution of y given $[x]$ is normal. This assumption is needed only when null hypotheses are tested for significance or prediction and confidence

limits are calculated. It also implies that y_i and y_j are statistically independent since two normally distributed variables that are uncorrelated are also statistically independent.

There are also several implicit assumptions that are not mentioned here. For example, the fixed variables x_1, x_2, \dots, x_m are measured without error, the number n of sample elements is larger than the number m of independent variables x , there are at least m distinct vectors $[x]$ in the sample, or the population from which the sample elements are selected is static, that is, does not change with time.

Let the values y, x_1, x_2, \dots, x_m of the k -th sample element, $k = 1, 2, \dots, n$ be denoted as

$$y_k, x_{k1}, x_{k2}, \dots, x_{km}$$

To apply the least squares method we shall arrange these values in the two matrices $[X]$ and $[Y]$ of sample values,

$$[X] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad \text{and} \quad [Y] = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

The matrices of cross products are now defined as

$$[T] = [X]'[X] \text{ and } [P] = [X]'[Y]$$

Then, the three statistics that summarize the information from the set of n sample elements are

(1) $[b] = [T]^{-1}[P]$ = estimator of the vector of regression coefficients $[\beta]$

(2) $S_{yy|x} = ([Y]'[Y] - [b]'[P]) / (n-m)$ = estimator of the conditional variance $\sigma_{yy|x}$ of y given $[x]$, and

(3) $[S_{bb}] = S_{yy|x} [T]^{-1}$ = estimator of the covariance matrix $[\sigma_{bb}]$ of $[b]$

where $[T]^{-1}$ is the inverse of the matrix $[T]$.

Weighted Least Squares Method as Applied to Biomass Regressions

The basic assumptions of the least squares linear regression are all the same except for the second assumption (about the variance of y) which is now changed to the new assumption that the conditional variance of y given $[x]$ is proportional to a^2 , a value that is known for all sample (and population) elements. Ordinarily, a^2 is a known function of the given independent variables x_1, x_2, \dots, x_m . More formally, this can be expressed as

$$\sigma_{yy|x} = \sigma_{uu|v} a^2$$

= conditional variance of y given [x]

where

a = known value, and

$\sigma_{uu|v}$ = an unknown constant value (factor of proportionality)
 = conditional variance of the new transformed variable $u = (y/a)$ given $[v] = [x]/a$

To calculate the weighted least squares estimates [b] of the vector of regression coefficients $[\beta]$ and $[S_{bb}]$ of the covariance matrix of [b] we apply the least squares method to the new transformed variables

$$u = y/a, v_1 = x_1/a, v_2 = x_2/a, \dots, v_m = x_m/a,$$

that is, we apply the least squares method to the new regression function of u on [v]

$$\hat{u} = \beta_1 v_1 + \beta_2 v_2 + \dots + \beta_m v_m = [\beta]'[v]$$

More specifically, we start with the calculation of the matrices of the transformed values [U] and [V], whose elements are

$$u_k = y_k/a, v_{k1} = x_{k1}/a, v_{k2} = x_{k2}/a, \dots, v_{km} = x_{km}/a$$

We continue with the calculation of the matrices of cross products

$$[T] = [V]'[V] \text{ and } [P] = [V]'[U]$$

and the calculation of the three basic statistics

- (1) $[b] = [T]^{-1}[P]$ = weighted least squares estimator of the vector of regression coefficients $[\beta]$
- (2) $S_{uu|v} = ([U]'[U] - [b]'[P]) / (n-m)$
 = estimator of the conditional variance $\sigma_{uu|v}$ of u given [v], and
- (3) $[S_{bb}] = S_{uu|v} [T]^{-1}$ = estimator of the covariance matrix $[\sigma_{bb}]$ of [b]

Note the following:

- (1) $S_{yy|x} = a^2 S_{uu|v}$ = estimator of the conditional variance of y given [x], a value which varies from tree to tree according to the tree vector [x]

(2) When y is the biomass of the tree bole and [x] is defined as a function of the tree diameter d, empirical evidence shows that "a" is approximately equal to d^2 . As the tree basal area is $\pi d^2/4$, the new variable $u = y/d^2$ represents a value which is proportional to the variable "bole biomass per square inch of tree basal area".

(3) When y is the biomass of the tree bole and [x] is defined as a function of the tree diameter d and height h, empirical evidence shows that "a" is approximately equal to $d^2 h$. As the bole volume is proportional to $d^2 h$ (recall that

the function $y = b d^2 h$ can be used as a regression function of y on d and h), the new variable $u = y/d^2 h$ represents a value which is approximately proportional to the variable "biomass per unit of bole volume".

Testing Null Hypotheses about $[\beta]$

Sometimes it may be of interest to test the null hypothesis that some regression coefficients are equal to zero; that is, the corresponding variables x may be eliminated from the regression function without diminishing the goodness of the regression function to estimate y. This accomplishes two things, it simplifies the regression function and it reduces the error of the regression estimators.

Because the variables x are arranged in an arbitrary way, let us show how to test the null hypothesis that, for some $r < m$,

$$\beta_{r+1} = \beta_{r+2} = \dots = \beta_m = 0$$

The test requires the assumption that the conditional probability distribution of y given [x] is normal and consists of the following steps

Step 1 - Calculate the regression sums of squares

$$CR_1SS = [b_1]'[P_1] \text{ and}$$

$$CR_2SS = [b_2]'[P_2]$$

of the Unrestricted Regression R_1

$$R_1: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \beta_{r+1} x_{r+1} + \dots + \beta_m x_m$$

and the Restricted (under the null hypothesis) Regression R_2

$$R_2: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r$$

Step 2 - Calculate the test statistic

$$F = (n-m)(CR_1SS - CR_2SS) / (m-r) U_1SS$$

where

$$U_1SS = [Y]'[Y] - CR_1SS$$

Step 3 - From a table of F-distribution with (m-r) and (n-m) degrees of freedom, find the critical value F_α , for some probability α of rejecting the null hypothesis when the null hypothesis is true. Then, apply the decision rule to accept the null hypothesis when $F < F_\alpha$ and reject it otherwise.

The above procedure has been written for the least squares method. When the weighted least squares method is being used, the same procedure applies, except that

$$[P_1] = [V_1]'[U_1],$$

$$[P_2] = [V_2]'[U_2], \text{ and}$$

$$U_1SS = [U]'[U] - [b_1]'[P_1]$$

There is an alternate test for the special case where $r = m-1$, that is, for testing the null

hypothesis that $\beta_i = 0$ for some $i = 1, 2, \dots, m$. Then, the test statistic is

$$t = b_i / \sqrt{S_{b_i b_i}}$$

where $S_{b_i b_i}$ is the estimator of the variance of

b_i provided by $[S_{bb}]$. The critical value t_α is taken from a table of t-distribution with $(n-m)$ degrees of freedom for some probability level α of rejecting the null hypothesis, when the null hypothesis is true. The null hypothesis is accepted when $-t_\alpha < t < t_\alpha$ or rejected otherwise.

Constructing $(1-\alpha)$ Confidence Intervals for β_i and $[\beta]$

The $(1-\alpha)$ confidence intervals for some regression coefficient β_i is calculated by the formula

$$b_i \pm t \sqrt{S_{b_i b_i}}$$

where

t = critical value of t read from a table of t-distribution with $(n-m)$ degrees of freedom and for $(1-\alpha)$ confidence level, and

$S_{b_i b_i}$ = estimator of the variance of b_i , the i -th diagonal element of $[S_{bb}]$.

Note that the inferences about the confidence interval of β_i are strictly valid only if the conditional probability distribution of y given $[x]$ is normal; they are acceptably good even when the distribution is not normal but the sample size is sufficiently large.

Sometimes it may be interesting to make inferences about a joint confidence interval for the entire vector $[\beta]$. It is not right to calculate the confidence intervals for each β_i separately and then combine them; they are not statistically independent. To calculate such a joint confidence interval we start by defining

$$\begin{aligned} \text{SSBB} &= [b - \beta]' [T] [b - \beta] \\ &= \text{sum of squares associated with the vector } [\beta] \text{ of regression coefficients.} \end{aligned}$$

Because $[\beta]$ is not known, this sum of squares cannot be calculated. However, it can be used to define the $(1-\alpha)$ joint confidence interval of $\beta_1, \beta_2, \dots, \beta_m$ as the set of values $[\beta]$ that satisfy the inequality

$$[b - \beta]' [T] [b - \beta] < m F S_{yy|x}$$

for the least squares method, or

$$[b - \beta]' [T] [b - \beta] < m F S_{uu|x}$$

for the weighted least squares method, where F is taken from a table of F-distribution with m and $(n-m)$ degrees of freedom at the desired $(1-\alpha)$ confidence level.

Note that the statements about the confidence intervals of the individual regression coefficients β_i or the joint confidence interval of $[\beta]$ are strictly valid only if the assumption about normality of the conditional probability distribution of y given $[x]$ for the least squares (or u for given $[v]$ for the weighted least squares) is satisfied. Of course, because of the central limit theorem, the confidence interval statements are approximately all right when the sample size is sufficiently large.

Constructing Confidence and Prediction Intervals Associated with the Regression Estimates

Assume that an element of the population is selected at random, its values x_1, x_2, \dots, x_m are measured and, thus, they are known and we are interested in calculating point and $(1-\alpha)$ interval estimators for (i) the expected value of y for that given $[x]$ and (ii) the actual value of y for that particular element.

Let the measured values x_1, x_2, \dots, x_m be written as the vector

$$[x_0]' = [x_{01} \ x_{02} \ \dots \ x_{0m}]$$

the expected value (the arithmetic mean) of all population elements that have $[x] = [x_0]$ be written as μ_0 and the actual value (a random variable) taken on by the selected tree be written as y_0 . Then

(1) the point estimator of both μ_0 and y_0 is the regression estimator

$$\hat{y}_0 = [b]' [x_0]$$

(2) the $(1-\alpha)$ confidence interval of the expected value of y given $[x] = [x_0]$ is given by

$$\hat{y}_0 \pm t \sqrt{S_{y_0 y_0}}$$

where

$$\begin{aligned} S_{y_0 y_0} &= [x_0]' [S_{bb}] [x_0] \\ &= \text{estimator of the variance of } \hat{y}_0, \\ &\text{that is, the variance of the estimation error } (\hat{y}_0 - \mu_0), \text{ and} \end{aligned}$$

t = value read from a table of t-distribution with $(n-m)$ degrees of freedom at the $(1-\alpha)$ confidence level.

(3) the $(1-\alpha)$ prediction interval of the actual value of y given $[x] = [x_0]$ is given by

$$\hat{y}_0 \pm t \sqrt{S_{y_0 y_0}}$$

where

$$\begin{aligned} S_{y_0 y_0} &= \hat{S}_{y_0 y_0} + S_{yy|x_0} \\ &= \text{estimator of the variance of the prediction error } (y_0 - \hat{y}_0), \\ &\text{and} \end{aligned}$$

t has been defined above

Note that, for the case of weighted least squares,

$$S_{yy|x_0} = a^2 S_{uu|v}$$

and, thus,

$$S_{y_0 y_0} = S_{uu|v} (a^2 + [x_0]' [T]^{-1} [x_0])$$

It may also be of interest here to state the basic difference between the confidence and prediction intervals; the confidence interval may or may not include the parameter μ_0 , a fixed value, while the prediction interval refers to an interval in which a random variable y_0 may or may not fall in, at some future time.

A Numerical Example

Let us use the data from 353 sample trees reported by Cunia (1985). These trees were selected by simple random sampling and their diameter at breast height (in inches) and total above ground biomass y (in pounds of green weight) are listed by species group in Table 1. To estimate the regression function of y on d , for all species combined, we must select a mathematical model to represent the regression function and the conditional variance of y given d .

From past experience it is known that, an acceptably good mathematical model to express the regression function of the tree biomass y on diameter d is that of the rectangular parabolae

$$\begin{aligned} y &= \beta_1 + \beta_2 d + \beta_3 d^2 \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = [\beta]' [x] \end{aligned}$$

where $x_1=1$, $x_2=d$ and $x_3=d^2$. This can also be seen from the plot of the values y over the values d of Table 1 shown in Figure 1. Also from past experience it is known that the conditional variance of y given d is approximately proportional to d^4 . That this is reasonable can be seen from Figure 2 where the conditional standard deviation of y given d is plotted against the squared diameter d^2 . Note that the statement "conditional variance of y given d is proportional to d^4 " is equivalent to the statement "conditional standard deviation of y given d is proportional to d^2 " and also equivalent to the statement "a straight line passing through the origin of the two axes" is a good expression of the relationship between conditional standard deviation of y given d and the squared diameter d^2 .

To calculate the three main statistics of the regression function by the weighted least squares method we proceed as follows. We start with the calculation of the new transformed variables

$$\begin{aligned} u_k &= y_k / d_k^2 \\ v_{k1} &= x_{k1} / d_k^2 = 1 / d_k^2 \\ v_{k2} &= x_{k2} / d_k^2 = d_k / d_k^2 = 1 / d_k \\ v_{k3} &= x_{k3} / d_k^2 = d_k^2 / d_k^2 = 1 \end{aligned}$$

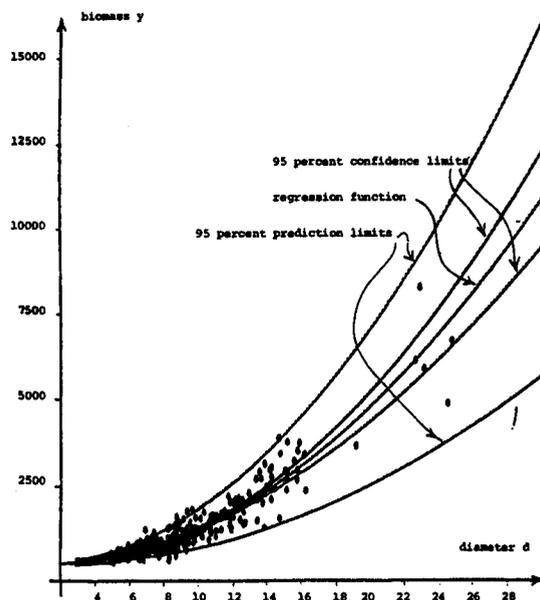


Figure 1 - Total above ground tree biomass y (green weight in pounds) plotted against tree diameter d (inches) together with the parabolic weighted least squares biomass regression and the corresponding 95 percent confidence and prediction limits.

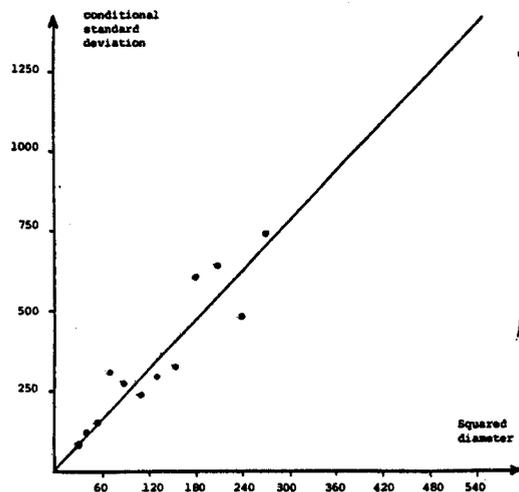


Figure 2 - The sample conditional standard deviation $S_{yy|x}$ of biomass y for given diameter d plotted against the squared diameter, and a straight line passing through the origin, an expression of the relationship between conditional standard deviation and squared diameter.

Table 1 - Diameters d (inches) and total above ground biomass y (pounds of green weight) of 353 trees arranged by species groups.

d	y	d	y	d	y	d	y	d	y	d	y	d	y	d	y
<u>Species group 1</u>															
9.3	557	7.8	621	9.5	813	8.2	546	7.2	322	9.3	647	12.7	1220	14.3	1953
11.4	1529	8.3	790	12.5	1356	8.7	627	12.0	1055	5.6	236	7.5	409	9.3	866
7.9	369	7.2	507	6.2	288	9.3	936	6.5	272	7.6	460	6.1	234	6.6	272
12.4	1993	5.0	206	7.2	356	7.5	436	7.8	533	5.0	123	9.4	419	7.1	503
6.6	336	5.1	240	11.1	851	6.6	400	11.4	1532	7.6	410	13.9	1125	9.2	944
8.4	637	5.7	235	9.8	821	9.5	846	6.5	306	8.5	497	5.5	221	6.3	276
11.9	1277	5.3	251	7.1	394	6.5	296	7.4	475	8.6	444	5.1	244	9.5	647
13.4	2539	7.1	513	5.3	178	9.0	771	14.2	2376	13.5	1315	5.4	275	9.2	1005
7.0	439	7.2	474	8.2	482	9.2	717	6.1	210	7.8	410	6.0	186	9.5	962
11.1	1471	5.1	219	8.6	626	11.7	1534	7.6	414	5.6	234	6.0	141	--	--
6.7	364	22.8	8195	7.7	594	15.8	2518	11.4	1057	6.3	268	5.5	187	--	--
7.0	377	8.0	339	9.5	756	11.9	1864	13.5	1956	9.5	885	5.1	158	--	--
6.9	384	8.1	577	6.8	378	6.9	388	5.2	149	5.8	225	5.2	170	--	--
<u>Species group 2</u>															
7.2	762	5.4	435	6.3	513	6.5	603	5.5	145	14.0	2311	7.0	469	8.8	841
6.8	485	8.7	1091	6.6	355	16.3	2219	8.0	383	8.9	457	6.6	327	7.7	529
6.3	322	6.4	505	5.0	250	5.2	175	12.0	1124	13.8	1831	15.2	2531	11.6	1607
6.1	210	8.7	1186	10.4	1563	11.9	1190	6.2	278	16.2	3266	8.5	779	15.2	3629
7.2	506	8.9	1162	10.7	1138	8.3	472	9.5	803	9.9	701	8.1	450	7.5	569
7.3	473	5.3	325	6.0	246	5.0	142	5.3	184	9.2	657	5.3	146	12.4	1830
11.8	1542	7.5	753	5.2	170	5.2	163	5.1	162	8.9	590	14.8	1395	7.4	816
6.8	465	6.0	443	7.1	546	6.8	331	6.8	391	7.5	434	9.3	881	9.8	1085
13.8	2151	7.1	753	7.8	310	6.9	329	7.1	486	7.1	555	6.8	421	15.8	3385
8.8	950	8.8	612	10.0	940	6.6	265	5.0	101	7.4	614	5.2	175	--	--
6.7	431	6.8	566	8.6	672	5.9	185	15.1	2791	6.8	325	7.4	582	--	--
7.5	956	5.4	411	9.1	1012	5.3	159	6.2	334	8.6	883	6.8	429	--	--
5.7	334	5.2	199	6.8	433	6.4	338	24.5	4797	7.1	526	8.1	477	--	--
7.6	436	6.1	255	11.8	2039	12.1	1531	9.3	795	12.1	1440	6.6	198	--	--
<u>Species group 3</u>															
6.5	535	5.4	373	5.5	281	6.5	284	6.7	540	8.4	167	5.3	320	9.4	1076
11.1	1023	12.6	1931	5.5	218	6.0	235	11.9	1435	6.9	151	8.8	1137	15.9	3606
13.7	2560	11.9	1438	5.0	307	19.2	3544	9.6	910	8.4	134	7.4	686	13.9	2999
9.6	941	6.7	685	6.3	462	7.2	524	10.8	965	8.8	261	15.6	3083	10.8	1445
6.6	527	6.6	548	6.5	415	12.2	1929	10.0	1053	7.0	508	7.8	704	6.7	409
9.0	1242	7.3	599	7.9	516	5.4	374	10.7	740	14.3	2877	10.9	1464	12.5	1361
13.0	2140	6.4	443	5.2	267	15.2	2736	8.5	510	9.2	1340	6.9	143	5.8	379
5.7	238	11.9	1464	7.5	645	12.7	1577	5.5	182	7.1	441	9.7	1063	6.5	405
5.9	327	6.0	367	7.1	581	7.8	676	24.7	6636	14.8	3287	12.8	1255	10.4	1069
5.1	197	13.6	2759	6.1	393	7.7	743	7.4	559	5.0	183	8.3	1053	8.6	825
22.6	6056	6.3	367	9.8	655	12.3	1710	7.4	449	5.8	398	14.3	2611	8.9	748
9.4	1576	7.9	715	5.6	422	7.0	541	8.8	1046	6.3	508	11.5	1531	10.1	1010
14.7	3744	5.3	308	6.8	446	10.7	1277	8.6	790	6.6	584	9.9	1110	14.5	2264
8.4	1106	6.5	511	7.3	488	7.6	546	9.2	678	7.3	521	10.3	841	--	--
9.7	1630	14.2	2819	6.9	404	15.8	2815	10.2	998	7.5	735	15.2	2233	--	--
5.9	423	12.4	1995	7.0	568	5.6	252	6.0	317	5.8	354	12.1	1847	--	--
6.0	321	10.2	1367	8.8	1116	14.3	2221	5.7	278	5.6	328	9.6	1241	--	--
7.0	1035	10.3	1090	5.1	217	5.7	278	8.8	1416	6.4	537	23.1	5804	--	--
6.0	568	6.5	354	8.6	773	10.8	1162	7.8	235	6.0	478	9.9	1398	--	--

for each individual sample tree $k = 1, 2, \dots, 353$ of Table 1. For example, for $k = 1$, the first tree of species 1, these variables are

$$u_1 = 557/(9.3)^2 = 6.4400509$$

$$x_1 = 1/(9.3)^2 = .011562030$$

$$x_2 = 1/9.3 = .10752688$$

$$x_3 = 1$$

The new variables u are arranged in a 353 by 1 vector $[U]$ and the new variables v in a 353 by 3 matrix $[V]$. We continue by calculating the matrices of cross products

$$[T] = [V]'[V] = \begin{bmatrix} .14868724 & .94619303 & 6.3324803 \\ .94619303 & 6.3324803 & 45.275493 \\ 6.3324803 & 45.275493 & 353 \end{bmatrix}$$

$$[P] = [V]'[U] = \begin{bmatrix} 58.746219 \\ 430.50963 \\ 3456.2479 \end{bmatrix}$$

the inverse of the matrix [T]

$$[T]^{-1} = \begin{bmatrix} 1085.6999 & -276.84567 & 16.031589 \\ -276.84567 & 72.496669 & -4.3320188 \\ 16.031589 & -4.3320188 & .27086279 \end{bmatrix}$$

and the three basic statistics

$$[b] = [T]^{-1}[P] = \begin{bmatrix} 5.1818118 \\ -25.653078 \\ 12.988357 \end{bmatrix}$$

= estimate of $[\beta]$
 $S_{uu|v} = ([U]'[U] - [b]'[P]) / (n-3) = 8.0278958$
 = estimate of $\sigma_{uu|v}$, the conditional variance of the transformed variables $u = y/d^2$ given d

and

$$[S_{bb}] = S_{uu|v}[T]^{-1}$$

$$= \begin{bmatrix} 8715.8856 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

= estimate of the covariance matrix $[\sigma_{bb}]$ of $[b]$.

Under the basic assumptions of the weighted least squares linear regression method, the estimators $[b]$, $S_{uu|v}$ and $[S_{bb}]$ are all unbiased.

If the conditional variance of y given d is required, one can estimate it by the formula

$$S_{yy|d} = d^4 S_{uu|v}$$

For example, if $d = 10$ inches, then

$$S_{yy|d} = (10)^4 (8.0278958) = 80278.958 \text{ square pounds}$$

It may be interesting to consider the null hypothesis $\beta_1 = 0$, that is, the regression function passes through the origin of the two axes. Because the test statistic t is equal to

$$t = b_1 / \sqrt{S_{b_1 b_1}} = 5.1818118 / \sqrt{8715.8856} = .06$$

a value which is not significant, the null hypothesis would normally be accepted. However, what happens at $d=0$ is irrelevant; we have no trees of diameter equal to zero and we should not extrapolate the application of the regression function anyway.

We can similarly test the null hypothesis that $\beta_2 = 0$. As the test statistic

$$t = b_2 / \sqrt{S_{b_2 b_2}} = -25.653078 / \sqrt{581.99570} = -1.06$$

is not significant, this null hypothesis would normally be accepted as well. However, the restricted parabolic model

$$\hat{y} = \beta_1 + \beta_3 d^2$$

includes all possible parabolas for which the minimum occurs at $d=0$ and the full parabolic model

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$$

includes the set of all possible parabolas. As the least squares model selects the parabola of least squares, there seems no reason to work with the restricted when we can work with the full parabolic model.

Note that the two separate null hypotheses above are not equivalent to the null hypothesis $\beta_1 = \beta_2 = 0$. This is because, if one of two null hypotheses is accepted, and the corresponding term eliminated from the regression function, the other null hypothesis will not necessarily be accepted also.

Using the formula

$$[b]'[x] \pm t \sqrt{[x]'[S_{bb}][x]}$$

for the confidence interval, the formula

$$[b]'[x] \pm t \sqrt{d^4 S_{uu|v} + [x]'[S_{bb}][x]}$$

where

$$[x]' = [1 \quad d \quad d^2]$$

for $d = 5, 6, \dots, 30$

for the prediction interval of future values and $t=2$ for the 95 percent confidence level, we have calculated the biomass table and its 95 percent confidence and prediction limits shown in Table 2. The corresponding regression function and the 95 percent confidence and prediction limits are shown graphically in Figure 1.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State University of New York, the United States Department of Agriculture Forest Service and the Department of Energy, Grant No. 23-524.

Literature Cited

- Cunia, T. On tree biomass tables and regressions: Some statistical comments. In: 1979 forest resource inventories workshop proceedings, vol. II, W. E. Frayer, (Ed.), Colorado State University, Fort Collins, CO; 1979a.
- Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: 1979 forest resource inventories workshop proceedings, vol. II, W. E. Frayer (Ed.), Colorado State University, Fort Collins, CO; 1979b.
- Cunia, T. Cluster sampling and tree biomass tables construction. In: Interdivisional Proceedings, 17th IUFRO World Congress, September 6-12, 1981, Kyoto, Japan; 1981.

Table 2 - The tree biomass table constructed by the weighted least squares method from sample data of Table 1 and its 95 percent confidence and prediction limits.

Diameter d	Lower 95 Percent Limits		Regression Estimates	Upper 95 Percent Limits	
	Prediction	Confidence		Confidence	Prediction
5	58	178	202	225	345
6	114	304	319	334	523
7	184	443	462	482	740
8	268	605	631	658	995
9	366	794	826	859	1287
10	479	1008	1048	1087	1616
11	607	1244	1295	1346	1982
12	749	1500	1568	1636	2387
13	905	1775	1869	1958	2829
14	1074	2070	2192	2313	3309
15	1258	2385	2543	2701	3828
16	1455	2719	2920	3121	4384
17	1666	3073	3323	3573	4979
18	1890	3446	3752	4057	5613
19	2128	3840	4207	4573	6285
20	2380	4254	4687	5121	6995
21	2644	4687	5194	5701	7744
22	2922	5141	5727	6314	8532
23	3214	5614	6286	6958	9358
24	3519	6108	6871	7633	10223
25	3837	6622	7481	8341	11126
26	4168	7155	8118	9081	12068
27	4513	7709	8781	9853	13049
28	4871	8283	9470	10656	14068
29	5242	8877	10184	11492	15126
30	5627	9491	10925	12359	16223

Cunia, T. On the error of biomass estimates in forest inventories; Part 1: Its major components. Faculty of Forestry Miscellaneous Publication Number 8 CESF 85-004), SUNY College of Environmental Science and Forestry, Syracuse, NY, 1985.

Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the workshop on "Tree Biomass regression functions and their contributions to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.

Cunia, T. An optimization model for subsampling trees for biomass measurement. In: Proceedings of the workshop on "Tree Biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.

Cunia, T. Evaluating errors of tree biomass regressions by simulation. In: Proceedings of the workshop on "Tree Biomass regression functions and their contributions to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY, 1986c.

Cunia, T. On the error of tree biomass regressions: trees selected by cluster sampling and double sampling. In: Proceedings of the workshop on "Tree Biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986d.

Tiberius Cunia

Professor of Statistics and Operations Research, State University of New York, College of Environmental Science and Forestry, Syracuse, NY 13210

If the sample elements can be subdivided into classes, one can use linear regression with dummy variables techniques to estimate individual class regressions and their error when their regression coefficients are interrelated. It is shown how to test null hypotheses about relationships among coefficients and it is illustrated how to apply these techniques to problems of (i) piecewise regressions, (ii) harmonization of biomass regressions and (iii) forcing additivity of biomass tables.

Introduction

In his paper, Cunia (1986a) reviews the method of weighted least squares linear regression as applied to the problem of estimating the regression of tree biomass y on several independent variables x defined in terms of diameter, height, etc. He assumed implicitly that all variables were quantitative, measured by an interval or ratio scale. There are cases, however, where the sample trees can be classified according to some criterion (species, forest type, site quality etc.) into several classes. Biomass regression functions may be required separately by classes or a single biomass regression may be desired for all classes, regression that includes the classification criterion as a set of independent variables x . In both cases it is useful to work with dummy variables techniques of the type described, among others, by Cunia (1973).

The objectives of the present paper are those of describing the general method of linear regressions with dummy variables and showing how this method has been applied to a variety of problems involving calculations of tree biomass regressions. More specifically we shall show how to use the standard, least or weighted least squares linear regression techniques and (i) estimate independent biomass regression functions for individual classes, (ii) test for similarity or identity the corresponding coefficients of various individual class regressions, (iii) calculate common estimators (and their error) of regression coefficients when the corresponding tests show that they are not significantly different and (iv) apply dummy variables to estimate piecewise linear regressions, harmonize biomass regressions and insure additivity of the regressions of several biomass components. In all cases we shall show how to calculate the vectors of regression coefficients $[b^i]$ of class i and the covariance matrices $[S_{bb}^i]$ of $[b^i]$ and $[b^j]$; these

statistics are needed, see Cunia (1986b) to calculate the error of forest inventory estimates.

Estimation of Individual Class Regressions

Assume that the n elements (trees) of the sample are grouped into q mutually exclusive and collectively exhaustive classes and are measured for the dependent variable y and the independent variables x_1, x_2, \dots, x_m which may be referred to as the vector of order m

$$[x]' = [x_1 \quad x_2 \quad \dots \quad x_m]$$

Assume also that the subsample of n_i elements in class i satisfies the basic assumptions of the least squares linear regression of y on $[x]$, in particular that the regression is of the linear form

$$\hat{y}_i = \beta_{i1}x_1 + \beta_{i2}x_2 + \dots + \beta_{im}x_m = [\beta^i]'[x] \text{ for } i = 1, 2, \dots, q$$

and that the conditional variance of y given $[x]$ is homogeneous over all n elements of the sample

Let us define now

- (1) the dummy variables $D_i, i = 1, 2, \dots, q$
 $D_i = 1$ if the tree belongs to class i
 $= 0$ otherwise

- (2) the new variables x_{ij} (with the first subscript denoting the class number $i = 1, 2, \dots, q$ and the second subscript denoting the variable number $j = 1, 2, \dots, m$) as

$$x_{ij} = D_i x_j = x_j \text{ if the tree belongs to class } i \\ = 0 \text{ otherwise,}$$

- and (3) the giant size regression

$$\hat{y} = \sum \beta_{ij} x_{ij} = [\beta]'[x] \\ = \beta_{11}x_{11} + \beta_{12}x_{12} + \dots + \beta_{1m}x_{1m} \\ + \beta_{21}x_{21} + \beta_{22}x_{22} + \dots + \beta_{2m}x_{2m} \\ + \dots \\ + \beta_{q1}x_{q1} + \beta_{q2}x_{q2} + \dots + \beta_{qm}x_{qm}$$

where

$$[\beta]' = [\beta_{11} \quad \beta_{12} \quad \dots \quad \beta_{1m} \quad \beta_{21} \quad \dots \quad \beta_{qm}] \\ = [[\beta^1]' \quad [\beta^2]' \quad \dots \quad [\beta^q]']$$

and a similar expression for $[x]$.

It is not necessary that the same independent variables be included in all regressions; by forcing coefficients β_{ij} to be equal to zero we can eliminate the variables x_j that are not desired in the regression of class i . It is also not necessary for a given variable x_j to have all the regression coefficients β_{ij} distinct; some classes may be defined as having equal regression coefficients for some variable x_j . Because the q individual class regressions are all included in

the giant size regression, one can (i) estimate all regressions in one single step and (ii) test null hypotheses about relationships among the coefficients of the various regressions, in particular hypotheses about equality of two or more coefficients β_{ij} (for given j).

Note that the giant size regression can be written in a long format form, showing in explicit terms the individual class regressions, as

$$\begin{aligned} \hat{y} &= \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1m}x_m && \text{if the tree belongs to class 1} \\ &= \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2m}x_m && \text{if the tree belongs to class 2} \\ &= \dots && \\ &= \beta_{q1}x_1 + \beta_{q2}x_2 + \dots + \beta_{qm}x_m && \text{if the tree belongs to class } q \end{aligned}$$

where some β_{ij} may be equal to zero and not all β_{ij} are distinct.

A simple procedure to calculate the least squares estimators $[B]$ of $[\beta]$ and $[S_{BB}]$ of the covariance matrix $[\sigma_{BB}]$ of $[B]$ is that described by Cunia (1986a). We start by defining the matrices $[Y]$ and $[X]$ of sample data and the matrices $[T] = [X]'[X]$ and $[P] = [X]'[Y]$ of the sums of crossproducts. If $[T]^{-1}$ denotes the inverse of the matrix $[T]$, then

$$\begin{aligned} [B] &= [T]^{-1}[P] \\ \text{and } S_{yy|x} &= ([Y]'[Y] - [B]'[P]) / (n - qm) \\ [S_{BB}] &= S_{yy|x}[T]^{-1} \end{aligned}$$

Note that the number of degrees of freedom $(n - qm)$ of $S_{yy|x}$ is correct only when each individual class regression has m independent variables (including the intercept terms $x_{i1} = 1$). The vector $[B]$ can be written in terms of the subvectors of regression coefficients $[b^i]$ as

$$\begin{aligned} [B]' &= \left[[b^1]' \quad [b^2]' \quad \dots \quad [b^q]' \right] \\ &= [b_{11} \quad b_{12} \quad \dots \quad b_{1m} \quad b_{21} \quad b_{22} \quad \dots \quad b_{qm}] \end{aligned}$$

and the matrix $[S_{BB}]$ can be expressed in terms of the covariance matrices $[S_{bb}^{ij}]$ of $[b^i]$ and $[b^j]$ as

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & \dots & [S_{bb}^{1q}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & \dots & [S_{bb}^{2q}] \\ \vdots & \vdots & & \vdots \\ [S_{bb}^{q1}] & [S_{bb}^{q2}] & \dots & [S_{bb}^{qq}] \end{bmatrix}$$

When the subsamples of various classes are statistically independent, and distinct regression functions are fit in each class, the submatrices

$[S_{bb}^{ij}]$ are all equal to zero, as $[b^i]$ and $[b^j]$ are statistically independent.

When the conditional variance of y given $[x]$ is not homogeneous, one should normally apply the weighted least squares method. In the case of biomass regressions, one can apply the procedure described, among others by Cunia (1986a) which assumes that the conditional variance of y given $[x]$ is proportional to a value a^2 known for all sample (and population) elements. Let us assume more specifically that

$$\sigma_{yy|x} = \sigma_{uu|v} a^2$$

where, for the trees with equal values $[x]$, we have the same function a^2 (independent of the class of trees) and $\sigma_{uu|v}$ is the same for all classes. Then, to calculate the weighted least squares estimates, we define first the transformed variables $u = y/a$, $v_{ij} = x_{ij}/a$ arranged in the matrices $[V]$ and $[U]$ of sample data. If $[T] = [V]'[V]$ and $[P] = [V]'[U]$, then

$$[B] = [T]^{-1}[P]$$

$$S_{uu|v} = ([U]'[U] - [B]'[P]) / (n - qm)$$

$$\text{and } [S_{BB}] = S_{uu|v}[T]^{-1}$$

Ordinarily $\sigma_{uu|v}$ would change from class to class. This is not important, however, since the differences are relatively small and the weighted least squares method is not too sensitive to small departures from its assumption of homogeneity of conditional variance of the transformed variable u . But if the departure is felt to be large and one wishes to take it into account, the procedure is more complex and requires the following steps

Step 1 - Calculate each class regression separately. Let the estimate $S_{uu|v}$ of class i be denoted as $S_{uu|v}^i$.

Step 2 - Proceed with the weighted least squares method as usual, but define the new transformed variables as

$$u = y/a\sqrt{S_{uu|v}^i} \quad \text{and} \quad v_{ij} = x_{ij}/a\sqrt{S_{uu|v}^i}$$

It may be of interest to note that for this case,

$$[U]'[U] - [B]'[P] = n - qm$$

and thus,

$$S_{uu|v} = 1 \quad \text{and} \quad [S_{BB}] = [T]^{-1}$$

Let us apply the above procedures to a numerical example.

Example 1 - Consider the data from the 353 sample trees listed in Table 1 of a paper by Cunia (1986a). These trees were measured for total above ground tree biomass y (pounds of green weight) and diameter at breast height d (inches). They were also classified into three species groups (1 for pines, 2 for maples and 3 for all other species). There are 100, 107 and 146 trees in species groups 1, 2 and 3 respec-

tively. It is desired to calculate a different biomass regression function for each individual group. From past empirical experience it is known that within each species group, (i) the biomass regression function is of the parabolic form, $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$ and (ii) the conditional variance of y given d is proportional to d^4 . It has been verified that these assumptions are also satisfied by our sample data.

To estimate the three regression functions we shall use the following two procedures

Procedure 1 - Applying the weighted least squares method as described by Cunia (1986a) to each species group $i = 1, 2, 3$ separately we obtain the following statistics, where the super-script refers to the group number i .

- (1) The matrices $[T^i]$ and $[P^i]$ of the sums of crossproducts of the variables u and v_{ij}

$$[T^1] = \begin{bmatrix} .04333 & .27451 & 1.83052 \\ .27451 & 1.83052 & 13.00988 \\ 1.83052 & 13.00988 & 100 \end{bmatrix}$$

$$[T^2] = \begin{bmatrix} .04917 & .30862 & 2.02798 \\ .30862 & 2.02798 & 14.15100 \\ 2.02798 & 14.15100 & 107 \end{bmatrix}$$

$$[T^3] = \begin{bmatrix} .05618 & .36306 & 2.47398 \\ .36306 & 2.47398 & 18.11461 \\ 2.47398 & 18.11461 & 146 \end{bmatrix}$$

$$[P^1] = \begin{bmatrix} 14.32026 \\ 104.79302 \\ 836.39548 \end{bmatrix}, [P^2] = \begin{bmatrix} 18.33304 \\ 131.75597 \\ 1027.3029 \end{bmatrix}$$

$$[P^3] = \begin{bmatrix} 26.09292 \\ 193.96063 \\ 1592.5495 \end{bmatrix}$$

- (2) The inverses of the matrices $[T^i]$,

$$[T^1]^{-1} = \begin{bmatrix} 4231.9155 & -1115.5300 & 67.66323 \\ -1115.5300 & 301.30214 & -18.77909 \\ 67.66323 & -18.77909 & 1.21455 \end{bmatrix}$$

$$[T^2]^{-1} = \begin{bmatrix} 3343.8895 & -863.61829 & 50.83846 \\ -863.61829 & 229.43526 & -13.97511 \\ 50.83846 & -13.97511 & .89404 \end{bmatrix}$$

$$[T^3]^{-1} = \begin{bmatrix} 2776.4363 & -687.93798 & 38.30736 \\ -687.93798 & 174.87134 & -10.03959 \\ 38.30736 & -10.03959 & .60337 \end{bmatrix}$$

- (3) The estimates of the vectors of regression coefficients and their covariance matrices

$$[b^1]' = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 295.60183 & -107.06967 & 16.882552 \end{bmatrix}$$

$$[b^2]' = \begin{bmatrix} b_{21} & b_{22} & b_{23} \\ -256.70604 & 40.050701 & 9.1695394 \end{bmatrix}$$

$$[b^3]' = \begin{bmatrix} b_{31} & b_{32} & b_{33} \\ 18.800242 & -20.693393 & 13.156786 \end{bmatrix}$$

$$S^1_{uu|v} = 3.1325346$$

$$S^2_{uu|v} = 7.7487808$$

$$S^3_{uu|v} = 9.0755518$$

$$[S^1_{bb}] = \begin{bmatrix} 13256.622 & -3494.4363 & 211.95740 \\ -3494.4363 & 943.83938 & -58.826156 \\ 211.95740 & -58.826156 & 3.8046237 \end{bmatrix}$$

$$[S^2_{bb}] = \begin{bmatrix} 25911.067 & -6691.9889 & 393.93612 \\ -6691.9889 & 1777.8435 & -108.29003 \\ 393.93612 & -108.29003 & 6.9277171 \end{bmatrix}$$

$$[S^3_{bb}] = \begin{bmatrix} 25197.692 & -6243.4168 & 347.66044 \\ -6243.4168 & 1587.0539 & -91.114812 \\ 347.66044 & -91.114812 & 5.4758847 \end{bmatrix}$$

Procedure 2 - Assuming that the same conditional variance function applies to the sample trees of all three species groups, that is, same function a^2 and $\sigma_{uu|v}$, we start by defining the dummy variables

$$\begin{aligned} D_1 &= 1 \text{ if tree of group 1} \\ &= 0 \text{ otherwise} \\ D_2 &= 1 \text{ if tree of group 2} \\ &= 0 \text{ otherwise} \\ D_3 &= 1 \text{ if tree of group 3} \\ &= 0 \text{ otherwise} \end{aligned}$$

If the single subscripted variables x_j are defined as

$$x_1=1, x_2=d, x_3=d^2$$

we defined the double subscripted variables x_{ij} as

$$\begin{aligned} x_{i1} &= D_i x_1 = 1 \text{ if tree of species } i \\ &= 0 \text{ otherwise} \\ x_{i2} &= D_i x_2 = d \text{ if tree of species } i \\ &= 0 \text{ otherwise} \\ x_{i3} &= D_i x_3 = d^2 \text{ if tree of species } i \\ &= 0 \text{ otherwise} \end{aligned}$$

for $i = 1, 2, 3$.

Because of the assumption that the conditional variance of y given d is proportional to d^4 (that is, $a = d^2$), we define the new variables u and v_{ij} as

$$u = y/d^2 \text{ and } v_{ij} = x_{ij}/d^2$$

Note that

$$\begin{aligned} v_{i1} &= 1/d^2 \text{ if tree of species } i \\ &= 0 \text{ otherwise} \\ v_{i2} &= 1/d \text{ if tree of species } i \\ &= 0 \text{ otherwise} \\ v_{i3} &= 1 \text{ if tree of species } i \\ &= 0 \text{ otherwise} \end{aligned}$$

The matrices $[T]$ and $[P]$ of the sums of crossproducts take the form

$$[T] = \begin{bmatrix} [T^1] & [0] & [0] \\ [0] & [T^2] & [0] \\ [0] & [0] & [T^3] \end{bmatrix}$$

and

$$[P] = \begin{bmatrix} [P^1] \\ [P^2] \\ [P^3] \end{bmatrix}$$

where $[T^i]$ and $[P^i]$ are the matrices of the sums of crossproducts of species i obtained by Procedure 1 above. The inverse of $[T]$ takes the form

$$[T]^{-1} = \begin{bmatrix} [T^1]^{-1} & [0] & [0] \\ [0] & [T^2]^{-1} & [0] \\ [0] & [0] & [T^3]^{-1} \end{bmatrix}$$

and, using matrix multiplication by blocks,

$$[B] = [T]^{-1}[P] = \begin{bmatrix} [T^1]^{-1}[P^1] \\ [T^2]^{-1}[P^2] \\ [T^3]^{-1}[P^3] \end{bmatrix} = \begin{bmatrix} [b^1] \\ [b^2] \\ [b^3] \end{bmatrix}$$

This is verified when the calculations are performed with the sample data. Then, it is found that

$$[B]' = \begin{bmatrix} 295.60183 & -107.06967 & 16.882552 \\ -256.70604 & 40.050701 & 9.1695394 \\ 18.800242 & -20.693393 & 13.156786 \end{bmatrix}$$

$$S_{uu|v} = 6.9986424$$

and the covariance matrix of $[B]$ is

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [0] & [0] \\ [0] & [S_{bb}^{22}] & [0] \\ [0] & [0] & [S_{bb}^{33}] \end{bmatrix}$$

where

$$[S_{bb}^{11}] = \begin{bmatrix} 29617.663 & -7807.1955 & 473.55072 \\ -7807.1955 & 2108.7059 & -131.42815 \\ 473.55072 & -131.42815 & 8.50021 \end{bmatrix}$$

$$[S_{bb}^{22}] = \begin{bmatrix} 23402.687 & -6044.1556 & 355.80023 \\ -6044.1556 & 1605.73531 & -97.80676 \\ 355.80023 & -97.80676 & 6.25706 \end{bmatrix}$$

and

$$[S_{bb}^{33}] = \begin{bmatrix} 19431.285 & -4814.6319 & 268.09952 \\ -4814.6319 & 1223.8620 & -70.26349 \\ 268.09952 & -70.26349 & 4.22275 \end{bmatrix}$$

Note that although the vectors $[b^i]$ of regression coefficients are identical by the two procedures, their covariance matrices are not. This is because in Procedure 2 we assume that the conditional variance of u given $[v]$ is the same in all strata. As this is not the case, and $[S_{bb}] = S_{uu|v}[T]^{-1}$, the covariance matrices $[S_{bb}]$ are not the same, even though the inverse matrices $[T]^{-1}$ are identical.

It may be interesting to realize that we have the following relationships

$$S_{uu|v} = \frac{(n_1-3)S_{uu|v}^1 + (n_2-3)S_{uu|v}^2 + (n_3-3)S_{uu|v}^3}{(n_1-3) + (n_2-3) + (n_3-3)}$$

where (n_i-3) is the number of degrees of freedom of $S_{uu|v}^i$ and that

$$[S_{bb}^{ii}] = (S_{uu|v}^i / S_{uu|v}^i) [S_{bb}^i]$$

Because $[S_{bb}^{ii}]$ of Procedure 1 is not equal to $[S_{bb}^{ii}]$ of Procedure 2, a question arises as to which covariance matrix to use, when it is important to know the error of the biomass regressions or the error of forest inventory estimates based on these regressions. At first sight, it seems that the error is estimated more accurately by $[S_{bb}^{ii}]$. A look at the three sample values $S_{uu|v}^i$ shows that $S_{uu|v}^1=3.13$ of species group 1 (pine trees) is the smallest, followed by $S_{uu|v}^2=7.75$ of species group 2 (maple trees) and $S_{uu|v}^3=9.08$ of species group 3 (all remaining trees). This is not surprising since softwood trees (pines) are expected to be less variable than hardwood trees (maples), and a species group consisting of trees of a variety of species is expected to be the most heterogeneous.

Consequently, when it is important to know the error of the individual regression functions, one is advised to use $[S_{bb}^{ii}]$. This may be the case when individual biomass tables are to be (i) constructed with their confidence and prediction limits and (ii) applied to estimate the error of forest inventory estimates by species group. But if the interest lies with the calculation of biomass estimates for all species combined, it does not really matter which matrix is used; the error of these estimates will be approximately the same. However, because working with $[S_{bb}^{ii}]$ presents advantages not shared by $[S_{bb}^i]$ we prefer using the estimates calculated by Procedure 2.

One of the main advantages of dummy variables techniques and giant size regressions (of Procedure 2) is that of using the entire set of all sample tree data to (i) test null hypotheses about the values taken on by the coefficients of some variable x_j in one or more individual class regressions (as, for example, whether these coefficients are equal to zero, have common values or are related in some specific way), (ii) estimate the values of the coefficients of some variable x_j , when based on a significance test they can be assumed to be in a given relationship and (iii) to estimate the covariance matrix $[S_{bb}^{ij}]$ when $[b^i]$ and $[b^j]$ have common coefficients or, in general, are not statistically independent.

Testing Null Hypotheses about Regression Coefficients

Cunia (1986a) describes the way to test the general null hypothesis that several regression coefficients are all equal to zero. The same test applies, albeit in a different form, to null hypotheses about linear combinations of regression coefficients.

To describe this new form of significance test let us assume that we deal with a given, unrestricted regression R_1 defined as

$$R_1: \hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

The null hypothesis is expressed as the following $r < m$ linear combinations among the m coefficients

$$\begin{aligned} a_{11}\beta_1 + a_{12}\beta_2 + \dots + a_{1m}\beta_m &= [a_1]'[\beta] = 0 \\ a_{21}\beta_1 + a_{22}\beta_2 + \dots + a_{2m}\beta_m &= [a_2]'[\beta] = 0 \\ &\vdots \\ a_{r1}\beta_1 + a_{r2}\beta_2 + \dots + a_{rm}\beta_m &= [a_r]'[\beta] = 0 \end{aligned}$$

Assume that we can solve the r equations in m variables β_j for any set of r regression coefficients in terms of the remaining $(m-r)$. Substitute this solution in R_1 , rearrange the terms under the remaining $(m-r)$ regression coefficients and, by using a different notation for the coefficients and the corresponding new independent variables, define the restricted (under the null hypothesis) regression R_2 as

$$R_2: \hat{y} = \beta'_1 x'_1 + \beta'_2 x'_2 + \dots + \beta'_{m-r} x'_{m-r}$$

We can now test the null hypothesis with the test statistic

$$F = \frac{(CR_{1SS} - CR_{2SS})/r}{U_1SS/(n-m)}$$

which has the F -distribution with r and $(n-m)$ degrees of freedom, where

$$\begin{aligned} CR_{1SS} &= [b_1]'[P_1] \\ &= \text{regression sum of squares of } R_1 \end{aligned}$$

$$\begin{aligned} CR_{2SS} &= [b_2]'[P_2] \\ &= \text{regression sum of squares of } R_2 \end{aligned}$$

$$\begin{aligned} U_1SS &= [Y]'[Y] - CR_{1SS} \\ &= \text{unexplained sum of squares of } R_1 \end{aligned}$$

and n = sample size.

For the case of weighted least squares method all of the sums of squares above refer to the transformed variables u and v

Let us now illustrate how to apply this test to the giant size regression function of the previous section.

Example 2 - Consider the sample data and the giant size regression of Example 1. The reader can verify that additional calculations yield

$$\begin{aligned} [U]'[U] &= \sum u^2 = \sum (y/d)^2 = 36961.259 \\ CR_{1SS} &= [B]'[P] = 34553.726 \\ U_1SS &= 36961.259 - 34553.726 = 2407.533 \\ U_1SS/(n-m) &= 2407.533/(353-9) = 6.9986424 \end{aligned}$$

where R_1 is the giant size regression

$$\begin{aligned} R_1: \hat{y} &= \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13} \\ &+ \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23} \\ &+ \beta_{31}x_{31} + \beta_{32}x_{32} + \beta_{33}x_{33} \end{aligned}$$

We would like to test now the null hypothesis

$$\beta_{12} = \beta_{22} = \beta_{32} = \beta_2 \text{ (same common value), and}$$

$$\beta_{13} = \beta_{23} = \beta_{33} = \beta_3 \text{ (same common value)}$$

When this null hypothesis is true, the three biomass regression functions are "parallel"; that is, they differ only by their intercept. The null hypothesis can also be expressed as the set of four linear combinations

$$\beta_{12} - \beta_{22} = [a_1]'[\beta] = 0$$

$$\beta_{22} - \beta_{32} = [a_2]'[\beta] = 0$$

$$\beta_{13} - \beta_{23} = [a_3]'[\beta] = 0, \text{ and}$$

$$\beta_{23} - \beta_{33} = [a_4]'[\beta] = 0$$

where

$$[a_1]' = [0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$[a_2]' = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0]$$

$$[a_3]' = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0]$$

and

$$[a_4]' = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad -1]$$

Solving for β_{22} , β_{32} , β_{23} and β_{33} in terms of β_{12} and β_{13} , which we shall now denote by β_2 and β_3 respectively, we can write

$$\begin{aligned} R_2: \hat{y} &= \beta_{11}x_{11} + \beta_2 x_{12} + \beta_3 x_{13} \\ &+ \beta_{21}x_{21} + \beta_2 x_{22} + \beta_3 x_{23} \\ &+ \beta_{31}x_{31} + \beta_2 x_{32} + \beta_3 x_{33} \\ &= \beta_{11}x_{11} + \beta_{21}x_{21} + \beta_{31}x_{31} + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

where

$$x_{12} + x_{22} + x_{32} = x_2, \text{ and}$$

$$x_{13} + x_{23} + x_{33} = x_3$$

The statistics associated with the restricted regression R_2 , are the following

$$[T_2] = \begin{bmatrix} .04333 & 0 & 0 & .27451 & 1.83052 \\ 0 & .04917 & 0 & .30862 & 2.02798 \\ 0 & 0 & .05618 & .36306 & 2.47398 \\ .27451 & .30862 & .36306 & 6.33248 & 45.27549 \\ 1.83052 & 2.02798 & 2.47398 & 45.27549 & 353 \end{bmatrix}$$

$$[P_2]' = [14.3203 \quad 18.3330 \quad 26.09292 \quad 430.5096 \quad 3456.2479]$$

Note that there is a relationship between the matrices $[T_2]$ and $[P_2]$ of the restricted regression R_2 and the corresponding matrices $[T_1]$ and $[P_1]$ of the unrestricted regression R_1 (Procedure 2, Example 1). Some values are the same, as for example .04333, .04917, etc., others are the sums of three values, as for example, 6.33248 (row 4, column 4 of $[T_2]$) which is equal to the sum of the three values 1.83052, 2.02798 and 2.47398 (row 2, column 2) of the submatrices $[T^1]$, $[T^2]$ and $[T^3]$ respectively of $[T_1]$ of the unrestricted regression R_1 .

After calculating $[T_2]^{-1}$, we find

$$[B_2]' = \begin{bmatrix} -36.543351 & 17.250897 & 77.620963 \\ & -29.919713 & 13.174939 \end{bmatrix}$$

$$CR_2SS = [B_2]'[P_2] = 34473.439$$

and,

$$[S_{bb}] = \begin{bmatrix} 7856.12 & 7688.97 & 7736.77 & -1977.43 & 114.489 \\ 7688.97 & 7832.29 & 7734.47 & -1977.83 & 114.599 \\ 7736.77 & 7734.47 & 7910.14 & -1987.70 & 114.949 \\ -1977.43 & -1977.83 & -1987.70 & 518.766 & -30.9890 \\ 114.489 & 114.599 & 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

The test statistic is

$$F = \frac{(CR_1SS - CR_2SS)/(4)}{U_1SS/(353-9)} = 2.868$$

with 4 and 344 degrees of freedom. As the critical F values for the .05 and .01 probability of rejection are 2.37 and 3.32 respectively, the usual decision would be to reject the null hypothesis. However, for the purpose of illustrating how to express the results as a 9 by 1 giant size vector [B] (that contains the three biomass regressions in an explicit form) we shall assume that the null hypothesis is accepted and we use the statistics of the restricted regression R_2 . We can write the estimate of the restricted regression as

$$\begin{aligned} R_2: \hat{y} &= b_{11} + b_2d + b_3d^2 \text{ for species group 1} \\ &= b_{21} + b_2d + b_3d^2 \text{ for species group 2} \\ &= b_{31} + b_2d + b_3d^2 \text{ for species group 3} \end{aligned}$$

This implies that the biomass regressions of the three species groups and the corresponding covariance matrices are

$$[b^1] = \begin{bmatrix} b_{11} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} -36.543351 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$[b^2] = \begin{bmatrix} b_{21} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 17.250897 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$[b^3] = \begin{bmatrix} b_{31} \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 77.620963 \\ -29.919713 \\ 13.174939 \end{bmatrix}$$

$$[S_{bb}^{11}] = \begin{bmatrix} S_{b_{11} b_{11}} & S_{b_{11} b_2} & S_{b_{11} b_3} \\ S_{b_2 b_{11}} & S_{b_2 b_2} & S_{b_2 b_3} \\ S_{b_3 b_{11}} & S_{b_3 b_2} & S_{b_3 b_3} \end{bmatrix}$$

$$= \begin{bmatrix} 7856.12 & -1977.43 & 114.489 \\ -1977.43 & 518.766 & -30.9890 \\ 114.489 & -30.9890 & 1.93720 \end{bmatrix}$$

and similarly

$$[S_{bb}^{22}] = \begin{bmatrix} 7832.29 & -1977.83 & 114.599 \\ -1977.83 & 518.766 & -30.9890 \\ 114.599 & -30.9890 & 1.93720 \end{bmatrix}$$

and

$$[S_{bb}^{33}] = \begin{bmatrix} 7910.14 & -1987.70 & 114.949 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

Because $[b^i]$ and $[b^j]$ are not statistically independent (they have common values b_2 and b_3) we have $[S_{bb}^{ij}] \neq 0$. Then, the reader can verify that

$$[S_{bb}^{12}] = \begin{bmatrix} S_{b_{11} b_{21}} & S_{b_{11} b_2} & S_{b_{11} b_3} \\ S_{b_2 b_{21}} & S_{b_2 b_2} & S_{b_2 b_3} \\ S_{b_3 b_{21}} & S_{b_3 b_2} & S_{b_3 b_3} \end{bmatrix}$$

$$= \begin{bmatrix} 7688.97 & -1977.43 & 114.489 \\ -1977.83 & 518.766 & -30.9890 \\ 114.599 & -30.9890 & 1.93720 \end{bmatrix}$$

and similarly

$$[S_{bb}^{13}] = \begin{bmatrix} 7736.77 & -1977.43 & 114.489 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

$$[S_{bb}^{23}] = \begin{bmatrix} 7734.47 & -1977.83 & 114.599 \\ -1987.70 & 518.766 & -30.9890 \\ 114.949 & -30.9890 & 1.93720 \end{bmatrix}$$

and

$$[S_{bb}^{21}] = [S_{bb}^{12}]', [S_{bb}^{31}] = [S_{bb}^{13}]' \text{ and } [S_{bb}^{32}] = [S_{bb}^{23}]'.$$

Consequently, the giant size vector [B] containing the subvectors of regression coefficients of the three individual species groups is the 9 by 1 vector

$$[B]' = \begin{bmatrix} [b^1]' & [b^2]' & [b^3]' \end{bmatrix}$$

and its covariance matrix is estimated by the 9 by 9 matrix

$$[S_{BB}] = \begin{bmatrix} [S_{bb}^{11}] & [S_{bb}^{12}] & [S_{bb}^{13}] \\ [S_{bb}^{21}] & [S_{bb}^{22}] & [S_{bb}^{23}] \\ [S_{bb}^{31}] & [S_{bb}^{32}] & [S_{bb}^{33}] \end{bmatrix}$$

The numerical values of these subvectors and submatrices are shown above.

Piecewise Linear Regressions

In the previous sections the sample elements were classified into groups by a criterion other than their values [x]. This is not necessary, as sometimes classes may be defined in terms of the variables [x]. Then, the resulting regressions will be of the piecewise form, that is regressions for which the function changes its form over the range of independent variables [x].

For example, when estimating biomass regressions one may have difficulties finding a single mathematical function that would fit sufficiently well the data over the entire range of tree diameters. The fit may be poor for the small trees, or for the large trees or for both. In this case, one may be able to fit piecewise regression functions consisting of one branch for the small and a second branch for the large trees; or, if the relationship seems to be of an S-shaped form, to fit a parabolic branch for the left-hand side, a linear branch for the middle and a second parabolic branch for the right-hand side of the regression function.

Fitting different mathematical functions to different classes of tree diameter would generally result in regression functions that present points of discontinuity at the class border; and this is not desirable. Then, one can set conditions on the choice of the regression coefficients of the various branches and, thus, obtain piecewise continuous regression functions. However, the transition from one to the next branch may not be smooth, even though the regression function is continuous; and this may also be undesirable. Then, one can force the two branches intersecting at the border of the two classes to have equal derivatives. In other words, one can set additional conditions on the choice of the regression coefficients and force the piecewise regression function to be continuous and have a first derivative everywhere.

Let us now show how to use linear regression with dummy variables to fit piecewise linear functions consisting of two branches. The extension to more than two branches is straightforward. We shall start with the grouping of the sample trees into two classes according to their diameter; class 1 containing the small trees with diameter less than or equal to d_0 (say $d_0 = 5$ inches) and class 2 containing the remaining trees. We shall assume that we want to fit different parabolic regressions to the two classes, that is, we shall make the assumption that

$$\hat{y} = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \quad \text{if } d \leq d_0$$

$$= \beta_{21} + \beta_{22}d + \beta_{23}d^2 \quad \text{if } d > d_0$$

This can be written as the giant size regression

$$R_1: y = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13}$$

$$+ \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23}$$

where

$$x_{11} = 1 \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{12} = d \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{13} = d^2 \quad \text{if } d \leq d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{21} = 1 \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{22} = d \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

$$x_{23} = d^2 \quad \text{if } d > d_0, \text{ or } 0 \text{ otherwise}$$

The two parabolic branches will not general-

ly intersect at the point $d=d_0$. To force them to intersect at that point we shall make the assumption that the estimates of the biomass of the trees of diameter $d=d_0$ by the two regression branches are equal, that is

$$\beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 = \beta_{21} + \beta_{22}d_0 + \beta_{23}d_0^2$$

This can be written as

$$[a]'[\beta] = 0$$

where

$$[a]' = [1 \quad d_0 \quad d_0^2 \quad -1 \quad -d_0 \quad -d_0^2]$$

Solving for β_{21} in terms of the other β coefficients, substituting this solution for β_{21} of regression R_1 , rearranging the terms by regression coefficients and by redefining the variables x we obtain the regression R_2 as follows:

$$\beta_{21} = \beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2$$

$$\hat{y} = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13}$$

$$+ (\beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2)x_{21}$$

$$+ \beta_{22}x_{22} + \beta_{23}x_{23}$$

$$= \beta_{11}(x_{11}+x_{21}) + \beta_{12}(x_{12}+d_0x_{21}) + \beta_{13}(x_{13}+d_0^2x_{21})$$

$$+ \beta_{22}(x_{22}-d_0x_{21}) + \beta_{23}(x_{23}-d_0^2x_{21})$$

and

$$R_2: \hat{y} = \beta_{11}x'_{11} + \beta_{12}x'_{12} + \beta_{13}x'_{13}$$

$$+ \beta_{22}x'_{22} + \beta_{23}x'_{23}$$

where

$$x'_{11} = x_{11} + x_{21} = 1$$

$$x'_{12} = x_{12} + d_0x_{21} = d \quad \text{if tree diameter } \leq d_0$$

$$= d_0 \quad \text{otherwise}$$

$$x'_{13} = x_{13} + d_0^2x_{21} = d^2 \quad \text{if tree diameter } \leq d_0$$

$$= d_0^2 \quad \text{otherwise}$$

$$x'_{22} = x_{22} - d_0x_{21} = 0 \quad \text{if tree diameter } \leq d_0$$

$$= (d-d_0) \quad \text{otherwise}$$

$$x'_{23} = x_{23} - d_0^2x_{21} = 0 \quad \text{if tree diameter } \leq d_0$$

$$= (d^2-d_0^2) \quad \text{otherwise}$$

Of course, R_2 can also be written in terms of the individual class regressions as

$$\hat{y} = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \quad \text{if tree diameter } \leq d_0$$

$$= \beta_{21} + \beta_{22}d + \beta_{23}d^2 \quad \text{otherwise}$$

with $\beta_{21} = \beta_{11} + \beta_{12}d_0 + \beta_{13}d_0^2 - \beta_{22}d_0 - \beta_{23}d_0^2$

Note that R_1 presents a discontinuity (a jump in the estimate of the biomass) at $d=d_0$ and that R_2 is a continuous function everywhere. If one wishes to test the null hypothesis that the size of the jump of regression R_1 at $d=d_0$ is equal to zero, he can use the F-statistic

$$F = (n-6)(CR_1SS - CR_2SS)/U_1SS$$

with 1 and n-6 degrees of freedom, where CRSS and USS are the regression and unexplained sums of squares. Of course, this test does not make sense in our case since we do not expect a jump. Consequently, the form of the regression function of tree biomass on diameter is that of R_2 .

Although continuous, the regression R_2 may present a sharp change as we move from one to the other branch of the regression at $d=d_0$. This may be undesirable. To obtain a smooth transition point we shall assume a second relationship among the regression coefficients; that the derivatives of the two parabolic branches are equal at $d=d_0$. More formally we make the assumption that in regression R_2 we have equality of the two derivatives at $d=d_0$, that is

$$\beta_{12} + 2d_0 \beta_{13} = \beta_{22} + 2d_0 \beta_{23}$$

that is $[c]'[\beta] = 0$

$$\text{where } [c]' = [0 \quad 1 \quad 2d_0 \quad -1 \quad -2d_0]$$

Solving for β_{22} , substituting this solution in R_2 , rearranging the terms and redefining the new variables, we obtain the new regression R_3 as follows

$$\beta_{22} = \beta_{12} + 2d_0 \beta_{13} - 2d_0 \beta_{23}$$

$$\begin{aligned} \hat{y} &= \beta_{11}x'_{11} + \beta_{12}x'_{12} + \beta_{13}x'_{13} \\ &+ (\beta_{12} + 2d_0\beta_{13} - 2d_0\beta_{23})x'_{22} + \beta_{23}x'_{23} \\ &= \beta_{11}x'_{11} + \beta_{12}(x'_{12} + x'_{22}) \\ &+ \beta_{13}(x'_{13} + 2d_0x'_{22}) + \beta_{23}(x'_{23} - 2d_0x'_{22}) \end{aligned}$$

that is,

$$R_3: \hat{y} = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$$

where

$$x_1 = 1 \quad \text{for all trees}$$

$$x_2 = d \quad \text{for all trees}$$

$$x_3 = d^2 \quad \text{if tree diameter} \leq d_0 \\ = d_0(2d - d_0) \quad \text{otherwise}$$

$$x_4 = 0 \quad \text{if tree diameter} \leq d_0 \\ = (d - d_0)^2 \quad \text{otherwise}$$

In terms of the individual class regressions, R_3 can be written as

$$R_3: \hat{y} = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \quad \text{if tree diameter} \leq d_0 \\ = \beta_{21} + \beta_{22}d + \beta_{23}d^2 \quad \text{if tree diameter} > d_0$$

where

$$\beta_{11} = \beta_1, \quad \beta_{12} = \beta_2, \quad \beta_{13} = \beta_3,$$

$$\beta_{21} = \beta_1 + d_0^2(\beta_4 - \beta_3)$$

$$\beta_{22} = \beta_2 + 2d_0(\beta_3 - \beta_4) \quad \text{and} \quad \beta_{23} = \beta_4$$

The individual class regression coefficients can also be written, for $i=1,2$ and $j=1,2,3$ as

$$\beta_{ij} = [a_{ij}]'[\beta]$$

where

$$[\beta] = [\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4]$$

$$[a_{11}]' = [1 \quad 0 \quad 0 \quad 0]$$

$$[a_{12}]' = [0 \quad 1 \quad 0 \quad 0]$$

$$[a_{13}]' = [0 \quad 0 \quad 1 \quad 0]$$

$$[a_{21}]' = [1 \quad 0 \quad -d_0^2 \quad d_0^2]$$

$$[a_{22}]' = [0 \quad 1 \quad 2d_0 \quad -2d_0]$$

$$[a_{23}]' = [0 \quad 0 \quad 0 \quad 1]$$

The procedure to estimate the regression R_3 is to apply the least squares method to the sample data y , x_1 , x_2 , x_3 and x_4 (or the weighted least squares method to the transformed sample data u , v_1 , v_2 , v_3 and v_4). More specifically, we define the matrices of sample data $[Y]$ and $[X]$ (or $[U]$ and $[V]$ in the case of weighted least squares) and we calculate the matrices of the sums of crossproducts $[T]$ and $[P]$, the vector of regression coefficients $[b]$ and the covariance matrix $[S_{bb}]$ of $[b]$ by the usual formulae. This yields the four regression coefficients b_1 , b_2 , b_3 and b_4 and the estimates of all their variances and covariances.

It is necessary sometimes to show the individual class regressions in explicit form, that is, to define the giant size vector $[B]$ of regression coefficients as

$$[B]' = \begin{bmatrix} [b^1]' & [b^2]' \end{bmatrix} \\ = [b_{11} \quad b_{12} \quad b_{13} \quad b_{21} \quad b_{22} \quad b_{23}]$$

where $b_{ij} = [a_{ij}]'[b]$ were defined above. To calculate the variances and covariances of b_{ij} is much more complex; for each b_{ij} and b_{hk} one should use the formula

$$\text{Cov}(b_{ij}, b_{hk}) = [a_{ij}]'[S_{bb}][a_{hk}]$$

For example, the reader can verify that

$$\text{Cov}(b_{11}, b_{11}) = [a_{11}]'[S_{bb}][a_{11}] = S_{b_1 b_1}$$

$$\text{Cov}(b_{11}, b_{12}) = [a_{11}]'[S_{bb}][a_{12}] = S_{b_1 b_2}$$

$$\text{Cov}(b_{11}, b_{21}) = [a_{11}]'[S_{bb}][a_{21}] \\ = S_{b_1 b_1} - d_0^2 S_{b_1 b_3} + d_0^2 S_{b_1 b_4}$$

etc.

It may be a good idea, at this point to verify that our regression R_3 is continuous and that the two branches have equal derivatives at $d=d_0$. Then,

(1) R_3 is continuous at $d=d_0$ since

$$\begin{aligned} &\beta_{21} + \beta_{22}d_0 + \beta_{23}d_0^2 \\ &= \beta_1 - d_0^2\beta_3 + d_0^2\beta_4 + d_0\beta_2 + 2d_0^2\beta_3 - 2d_0^2\beta_4 + d_0^2\beta_4 \end{aligned}$$

$$= \beta_1 + d_0 \beta_2 + d_0^2 \beta_3$$

$$= \beta_{11} + \beta_{12} d_0 + \beta_{13} d_0^2$$

and

(2) R_3 has a derivative at $d=d_0$ since derivative of right-hand side branch =

$$= \beta_{22} + 2d_0 \beta_{23}$$

$$= \beta_2 + 2d_0 \beta_3 - 2d_0 \beta_4 + 2\beta_4 d_0$$

$$= \beta_2 + 2d_0 \beta_3 = \beta_{12} + 2d_0 \beta_{13}$$

= derivative of left-hand side branch

Note that, to apply this procedure one needs to know the value of d_0 . Ordinarily this is not the case. However, if one knows the approximate region (range of diameter values) where the regression function changes its form, one can use trial and error and determine that value of d_0 for which the sum of squared residuals is approximately minimized.

Sometimes one may have doubts about the curvilinearity of one or the other branch. To test the null hypothesis that $\beta_{13} = \beta_3 = 0$ one can use the test statistic $t = b_3 / \sqrt{s_{b_3 b_3}}$ which has the t-distribution with $(n-4)$ degrees of freedom. A similar test can be devised for the null hypothesis $\beta_{23} = \beta_4 = 0$. To test the null hypothesis that simultaneously $\beta_{13} = \beta_{23} = 0$, that is $\beta_3 = \beta_4 = 0$, one can use the F distribution; the unrestricted regression is R_3 and the restricted regression is the straight line

$$R_4: \hat{y} = \beta_1 + \beta_2 d$$

Another null hypothesis of interest may be that of two identical parabolic branches, that is,

$$\beta_{11} = \beta_{21}, \beta_{12} = \beta_{22} \text{ and } \beta_{13} = \beta_{23}$$

A first test starts with the giant size regression R_1 defined as the unrestricted regression and regression R_5

$$R_5: \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$$

as the restricted one. The test statistic is

$$F = (n-6)(CR_1SS - CR_5SS) / (3)(U_1SS)$$

with 3 and $(n-6)$ degrees of freedom. A second test starts with regression R_3 defined as the unrestricted regression and R_5 as the restricted one. Note that in this approach the three equalities $\beta_{11} = \beta_{21}$, $\beta_{12} = \beta_{22}$ and $\beta_{13} = \beta_{23}$ reduce to the same equality $\beta_3 = \beta_4$. This is because $\beta_{11} = \beta_{21}$ can be written as

$$\beta_1 = \beta_1 + d_0(\beta_4 - \beta_3)$$

and this implies that $\beta_4 - \beta_3 = 0$. Similarly, for $\beta_{12} = \beta_{22}$

$$\beta_2 = \beta_2 + 2d_0(\beta_3 - \beta_4)$$

implies that $\beta_3 = \beta_4$. Finally, $\beta_{13} = \beta_{23}$ is a different way of expressing the same equality $\beta_3 = \beta_4$.

The restricted regression is R_3 because

$$\hat{y} = \beta_1 + \beta_2 d + \beta_3 (d^2 + 0) \text{ for tree diameters } \leq d_0$$

$$= \beta_1 + \beta_2 d + \beta_3 (0 + d^2) \text{ for tree diameters } > d_0$$

For more on piecewise regressions and an illustrative example, the reader is referred to Cunia (1973).

Harmonizing Biomass Regressions of Nested Components

A problem that arises quite often when biomass regression functions are estimated for various tree components, is that the estimated regressions may not behave the way they should with respect to each other. For example, if one tree component is a part of another component, their regressions should not intersect; or if they do, they should intersect outside the applicable range of tree diameter.

Jacobs and Cunia (1980) and Cunia and Briggs (1985a) have illustrated the application of linear regression with dummy variables techniques to the problem of harmonization of a set of regression functions estimating the biomass of the tree bole up to various top diameters. The regressions were forced to be similar in shape, the spacing between successive regressions had to behave rationally over the range of tree diameters of interest and the regressions were not allowed to cross over. It is the objective of the present section to illustrate the application of this approach to the harmonization of a specific set of nested components. In the present context, we shall define nested components as a set of successive components such that the biomass of one is smaller than, or at most equal to the biomass of the component preceding it in the set.

To better illustrate the technique let us consider the specific problem of estimating the biomass regression functions for the following six components; component 1 = entire tree bole, component 2 = bole up to 10 cm of top diameter, component 3 = bole up to 15 cm of top diameter, component 4 = bole up to 20 cm of top diameter, component 5 = bole up to 25 cm of top diameter and component 6 = bole up to 30 cm of top diameter. We have a sample of trees where each tree is measured for the biomass of at least one and as many as all six bole components.

An overall set of regression functions are desired, one function for each component so that (i) no part of a component may have more biomass than the component itself, (ii) the spacing between the regressions of the successive components should follow a reasonable pattern, (iii) the biomass estimates of the set should be increasing functions of tree diameter and decreasing functions of top diameter and (iv) no biomass estimate should be allowed to be negative within the applicable range of tree breast and top diameter.

To construct a regression model to fit the set of six regressions, one for each component, we shall start with a decision about the form of the regression function. Past experience has shown that (i) the regression function of the bole biomass y (of some component) on tree diameter at breast height d is of the parabolic form $\hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2$ and (ii) the conditional variance of y given d is proportional to d^4 . Let us assume here that an analysis of the sample data shows that these assumptions are sufficiently well satisfied and, thus, they can be used as the assumptions of our basic regression model.

We shall now define the six dummy variables

$$D_i = 1 \text{ if component } i, \\ = 0 \text{ otherwise}$$

for $i=1, 2, \dots, 6$ and the doubly subscripted variables

$$x_{ij} = D_i x_j \text{ for all } i=1,2,\dots,6 \text{ and } j=1,2,3$$

where $x_1=1$, $x_2=d$ and $x_3=d^2$. The giant size regression can now be written as

$$R_1: \hat{y} = \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13} \\ + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23} \\ + \dots \\ + \beta_{61}x_{61} + \beta_{62}x_{62} + \beta_{63}x_{63}$$

Note that, the giant size regression can also be written as a set of six individual component regressions

$$y = \beta_{11} + \beta_{12}d + \beta_{13}d^2 \text{ for component 1} \\ = \beta_{21} + \beta_{22}d + \beta_{23}d^2 \text{ for component 2} \\ \cdot \\ \cdot \\ = \beta_{61} + \beta_{62}d + \beta_{63}d^2 \text{ for component 6}$$

Because all regression coefficients are now contained in the same giant size regression, we can force them to enter into specific relationships. For our specific problem, there are two approximate relationships that seem reasonable. These relationships can be expressed as null hypotheses and, thus, can be tested for significance.

Relationship 1: The difference between the biomass of two components of the same tree is independent of tree size. Expressed otherwise, the six regression functions are "parallel", that is, the "distance" (on the vertical direction) between two regression functions is the same over the entire range of tree diameters.

This relationship can be justified by the following arguments. The difference between two bole components of the same tree is a log with end diameters equal to the top diameters of the given two components. Of course, we assume here

implicitly that both components can be defined on the tree. It is also reasonable to expect the shape of the upper part of the tree bole to be approximately the same for the trees of all sizes. This implies that the length and the end diameter (and, thus, the biomass) of this log are approximately the same for the trees of all sizes. Of course, there are differences between individual trees of the same or of a different size. But on the average we shall expect to have, for a component $i < j = 1, 2, \dots, 6$,

$$\hat{y}_i - \hat{y}_j = (\beta_{i1} + \beta_{i2}d + \beta_{i3}d^2) - (\beta_{j1} + \beta_{j2}d + \beta_{j3}d^2) \\ = \text{a value which is independent of the tree diameter } d.$$

This implies that $\beta_{i2} = \beta_{j2}$, $\beta_{i3} = \beta_{j3}$ and the difference $(\beta_{i1} - \beta_{j1})$ represents the expected biomass of the average i - j log expressing the difference between the two bole components i and j .

The null hypothesis that expresses this relationship in the form suitable for testing is

$$\beta_{12} = \beta_{22} = \dots = \beta_{62} \text{ and}$$

$$\beta_{13} = \beta_{23} = \dots = \beta_{63}$$

To test this null hypothesis, we apply the procedure outlined in a previous section. The giant size regression is the unrestricted regression R_1 and the restricted regression is

$$R_2: \hat{y} = \beta_{11}x_{11} + \beta_2 d + \beta_3 d^2 \\ + \beta_{21}x_{21} + \beta_{31}x_{31} + \dots + \beta_{61}x_{61}$$

Let us assume that this null hypothesis has been accepted and the new regression function is of the form R_2 . Then, we may continue with the testing of another relationship that seems also reasonable, namely

Relationship 2: The spacing between the regressions of successive bole components (which, under relationship 1 are "parallel") varies in a "quadratic" fashion. The spacing may be defined as the difference between the heights (the intercepts) of the regressions. If we assume that the intercepts β_{i1} are of the parabolic form

$$\beta_{i1} = \beta_1 + \beta_4 z_i + \beta_5 z_i^2$$

where $z_i =$ top diameter = 0,10,15,...,30, the height of successive regressions would change in a quadratic fashion. This relationship can be justified by the following arguments.

Consider a large tree that has all six components and section the tree at the points of top diameter equal to 30, 25, 20, 15 and 10 cm. This yields six sections or logs, which starting from the top of the tree will be denoted here as log 1, 2, ..., 6. Note that bole component 1 contains all six logs, bole component 2 contains only the logs 2, 3, ..., 6, etc. and that each log (except log 6) represents the difference between two successive bole components. The volume (and, thus, the biomass) of each consecu-

tive log varies with the length (which may be a linear function of the top diameter) and the square of the log diameter (which goes for consecutive logs from 0 to 10, 15, ..., 30 cm).

The null hypothesis associated with this relationship is that, for some fixed parameters β_1 , β_4 , and β_5 , we have the relationships

$$\begin{aligned}\beta_{11} &= \beta_1 \\ \beta_{21} &= \beta_1 + 10\beta_4 + 100\beta_5 \\ \beta_{31} &= \beta_1 + 15\beta_4 + 225\beta_5 \\ &\vdots \\ \beta_{61} &= \beta_1 + 30\beta_4 + 900\beta_5\end{aligned}$$

or, in terms of the new variable $z_i = 0, 10, 15, \dots, 30$,

$$\beta_{i1} = \beta_1 + \beta_4 z_i + \beta_5 z_i^2$$

The restricted regression for this null hypothesis is

$$R_3: \hat{y} = \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 z + \beta_5 z^2$$

and the unrestricted regression is R_2 above.

For a numerical application of this technique, the reader is referred to Jacobs and Cunia (1980) who use the weighted least squares method to estimate the regressions and test the null hypotheses. Because the biomass of the various components of the same tree are not statistically independent, Cunia and Briggs (1985) use the generalized least squares method to estimate the same regressions and test the same null hypotheses.

Forcing Additivity of Biomass Regressions

If a tree component is subdivided into several mutually exclusive and collectively exhaustive subcomponents, it is usually desirable for the regression estimate of the biomass of the component to be equal to the sum of the regression estimates of the biomass of the subcomponents. For example, if we assume that proper definitions of the components are given, the estimate of the bole biomass should be equal to the sum of the estimates of the unmerchantable top biomass and the biomass of the merchantable bole.

To define the problem in more formal terms, let us assume that (i) the tree is subdivided into (s-1) mutually exclusive and collectively exhaustive components, (ii) the biomass of the component $i = 1, 2, \dots, (s-1)$ is denoted by y_i and the total biomass of the tree by y_s and (iii) for some independent variables x_1, x_2, \dots, x_m the regression function of the component $i = 1, 2, \dots, s$ is denoted by

$$\hat{y}_i = r_i(x_1, x_2, \dots, x_m)$$

Then, it can be shown that (i) in the absence of

measurement error, the value y_s of a particular tree is equal to the sum of the values y_1, y_2, \dots, y_{s-1} and (ii) the true regression value of the total y_s is equal to the sum of the true regression values $\hat{y}_i, i = 1, 2, \dots, (s-1)$.

Consider now a representative sample of n trees and its n sets of sample values y_1, y_2, \dots, y_s . It is then known that, if the regression functions of the component parts and that of the total are estimated separately, the resulting biomass regressions are not necessarily additive. And this is true in spite of the fact that, for every sample tree, we have $y_1 + y_2 + \dots + y_{s-1} = y_s$. To insure additivity is a problem considered, among others, by Kozak (1970), Cunia and Briggs (1984, 1985b), Chiyenda and Kozak (1984) and Reed and Green (1985).

One can recognize three specific methods to insure additivity.

Method 1 - The "best" (from a statistical point of view) regression function is calculated for each component $y_i, i = 1, 2, \dots, s-1$ separately, by whatever procedure is thought best. The regression function of the total y_s is then defined as the sum of the regression functions of the components, that is

$$\hat{y}_s = \hat{y}_1 + \hat{y}_2 + \dots + \hat{y}_{s-1}$$

The method can be used with linear, non linear or combinations of linear or non linear functions; different independent variables x can be used in different regressions; it is not necessary that all components be measured on all sample trees (that is, missing data for some components of some trees presents no problem); and it calculates the "best" regression for each component. On the other hand, the estimate of regression of the total is not necessarily the "best" and what is an important drawback sometimes, the method does not provide means for the estimation of the error of the total regression; the regressions of various components are not statistically independent and the error of their sum is practically impossible to estimate.

Method 2 - The additivity is obtained by using the same independent variables x_1, x_2, \dots, x_m in the linear regression function of all components. If the weighted least squares method is used, the same weights must also be used in all regressions. The method has the main advantage that it is simple and easy to understand. On the other hand the regression functions of some components may not be the "best" in the statistical sense; regression coefficients may be estimated even though they may not be significantly different from zero; the same weights must be used with the weighted least squares method even when different weights may be better for different regressions, etc. Furthermore, if some data for some components are missing for some trees, the method fails to insure additivity; to obtain this additivity, one must ignore all information from the trees that have some data missing.

Method 3 - The additivity is obtained by

using linear regression techniques with dummy variables. More specifically we assume that the regression function of the biomass y on x_1, x_2, \dots, x_m within a component i is of the linear form

$$\hat{y}_i = \beta_{i1}x_1 + \beta_{i2}x_2 + \dots + \beta_{im}x_m,$$

for $i = 1, 2, \dots, s$

These s regression functions are expressed together as the giant size regression

$$\hat{y} = \beta_{11}x_{11} + \beta_{12}x_{12} + \dots + \beta_{1m}x_{1m} \\ + \beta_{21}x_{21} + \beta_{22}x_{22} + \dots + \beta_{2m}x_{2m} \\ + \dots \\ + \beta_{s1}x_{s1} + \beta_{s2}x_{s2} + \dots + \beta_{sm}x_{sm}$$

subject to the additivity constraints

$$\beta_{11} + \beta_{21} + \dots + \beta_{s-1,1} = \beta_{s1} \\ \beta_{12} + \beta_{22} + \dots + \beta_{s-1,2} = \beta_{s2} \\ \vdots \\ \beta_{1m} + \beta_{2m} + \dots + \beta_{s-1,m} = \beta_{sm}$$

In this regression,

$$x_{ij} = x_j \text{ if component } i \\ = 0 \text{ otherwise}$$

for all $i = 1, 2, \dots, s$ and $j = 1, 2, \dots, m$

It is not necessary for the regression function of each component to contain the same independent variables as all others; regression coefficients β_{ij} may be made equal to zero as desired; different sets of weights (in the weighted least squares method) may also be used for different regressions; missing biomass measurements for some components of some trees present no problem; the set of all regressions is "best" even though each individual regression is not necessarily the best; it is possible to measure the error of each regression separately using the information from data of all components. The fact that the biomass measurements of the various components of a given sample tree are not independent presents no problem as the generalized least squares method will take this dependence into account. The main drawback of the method is the fact that it is not well known and is not as easy to understand and apply as the other two methods.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State of New York, the United States Department of Agriculture Forest Service and the Department of Energy, Grant No. 23-524.

Literature Cited

- Chiyenda, S. S.; Kozak, A. Additivity of component biomass regression equations when the underlying model is linear. *Canadian Journal of Forest Research*, Vol. 14:441-446; 1984.
- Cunia, T. Dummy variables and some of their uses in regression analysis. In: *Proceedings of the June 1973 meeting of IUFRO Subject Group S4.02, Nancy-France, Vol. 1*, T. Cunia, K. Kuusela and A. J. Nash, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1973.
- Cunia, T. Construction of tree biomass tables by linear regression techniques. In: *Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates"*, May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.
- Cunia, T. Error of forest inventory estimates: its main components. In: *Proceedings of the workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates"*, May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.
- Cunia, T.; Briggs, R. D. Forcing additivity of biomass tables: some empirical results. *Canadian Journal of Forest Research*, Vol. 14:376-384; 1984.
- Cunia, T.; Briggs, R. D. Harmonizing biomass tables by generalized least squares. *Canadian Journal of Forest Research*, Vol. 15:331-340; 1985a.
- Cunia, T.; Briggs, R. D. Forcing additivity of biomass tables: use of the generalized least squares method. *Canadian Journal of Forest Research*, Vol. 15:23-28; 1985b.
- Jacobs, M. W.; Cunia, T. Use of dummy variables to harmonize tree biomass tables. *Canadian Journal of Forest Research*, Vol. 10:483-490; 1980.
- Kozak, A. Methods for ensuring additivity of biomass components by regression analysis. *Forestry Chronicle*, Vol. 46:402-404; 1970.
- Reed, D. D.; Green, E. J. A method of forcing additivity of biomass tables when using a non-linear model. *Canadian Journal of Forest Research*, Vol. 15:1184-1187; 1985.

Professor of Statistics and Operations Research,
 SUNY College of Environmental Science and Forestry,
 Syracuse, NY, 13210

The least squares method used in the estimation of biomass regression functions requires that the sample tree values be statistically uncorrelated. This requirement is normally satisfied when the trees are selected by simple random sampling. However, this is seldom, if ever the case in real life; the trees are more efficiently selected in clusters and by methods other than simple random sampling. As the computer packages used in the least squares estimation of the biomass regression functions assume that the tree biomass values are uncorrelated, valid questions are raised about the validity of the estimators $[b]$ of the vector of regression coefficients and $[S_{bb}]$ of the covariance matrix of $[b]$. To insure the validity of these estimators one would normally have to modify the least squares techniques so as to take into account the effect of the method by which the sample trees were selected. Two such modifications are presented; one that can be applied when the trees are selected by cluster random sampling, the other, when the trees are selected by a two-phase or double sampling procedure.

Introduction

In his paper, Cunia (1986a) describes an approach to combine the error of biomass regression functions with the error from sample plots when forest biomass estimates are calculated and their error estimated. This approach requires that (i) the true biomass regression function be of the linear form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

where y is the biomass (of some tree component) and $[x]$ is a vector of known attributes other than biomass and (ii) valid estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are given. Note that we have used $[]$ and $[]'$ to denote vectors or matrices and their transposes respectively.

The standard least squares method of linear regression as needed here is described by Cunia (1986b). He (i) defines the basic matrices $[X]$ and $[Y]$ of the values $[x]$ and y of the n sample elements and (ii) calculates the estimators $[b]$ of the vector $[\beta]$ of regression estimators and $[S_{bb}]$ of the covariance matrix of $[b]$ by the formulae

$$[b] = [T]^{-1}[P] \text{ and } [S_{bb}] = S_{yy|x}[T]^{-1}$$

where

$$[T] = [X]'[X] \\ [P] = [X]'[Y]$$

$[T]^{-1}$ is the inverse of $[T]$, and

$$S_{yy|x} = ([Y]'[Y] - [b]'[P]) / (n-m) \\ = \text{estimator of the conditional variance of } y \text{ given } [x]$$

In the same paper, Cunia (1986b) describes the changes to make in the procedure when the least is replaced by the weighted least squares method. Assuming that the conditional variance of y given $[x]$ is proportional to the known value a^2 , we start by defining the new, transformed variables $u = y/a$ and $[v] = [x]/a$. These new variables are arranged in the new matrices $[V]$ and $[U]$ of sample data and the estimators $[b]$ and $[S_{bb}]$ are defined by the formulae

$$[b] = [T]^{-1}[P] \text{ and}$$

$$[S_{bb}] = S_{uu|v}[T]^{-1}$$

where

$$[T] = [V]'[V] \\ [P] = [V]'[U]$$

and

$$S_{uu|v} = ([U]'[U] - [b]'[P]) / (n-m) \\ = \text{estimator of the conditional variance of } u \text{ given } [v]$$

Note that

$$S_{yy|x} = a^2 S_{uu|v} \\ = \text{estimator of the conditional variance of } y \text{ given } [x]$$

For the special case of the regression function of biomass y on tree diameter d , the conditional variance of y given d can be assumed to be approximately proportional to d^4 , that is, $a = d^2$. When the regression of biomass y on tree diameter d and height h is desired, the conditional variance of y given d can be assumed to be approximately proportional to $d^4 h^2$, that is, $a = d^2 h$.

The method of least or weighted least squares in its standard form above assumes implicitly that the variables y and $[x]$ are measured in the interval or ratio scale. When some of these variables are qualitative and their measurement scale is nominal or ordinal, one may still use the least or weighted least squares; but then, these variables must be transformed first to vectors of dummy variables. These procedures are not specifically needed here, even though they can be applied with the methodology described in this paper. The interested people are referred to Cunia (1986c), among others.

To obtain valid estimators $[b]$ and $[S_{bb}]$, certain basic assumptions of the least or weighted least squares must be satisfied first, at least approximately. For a more detailed discussion of the relationship between estimators and basic assumptions of the least squares regression method, the reader is referred to papers by Cunia (1979a, b, 1986b). This is not repeated here. Of interest in the present study is only the

assumption that the sample values y are statistically uncorrelated. This assumption is automatically satisfied when the trees are selected by simple random sampling. However, this is seldom, if ever the case; it is much more efficient to select the sample trees in clusters or by methods other than simple random sampling. As the computer packages used in the least squares estimation of the biomass regression functions assume that the tree biomass values are uncorrelated, valid questions may be raised about the validity of the estimators $[b]$ and $[S_{bb}]$. To insure their validity, appropriate modifications must be made to the standard least squares method; modifications that take into account the effect of the method by which the sample trees were selected.

The objectives of the present paper are those of presenting two modifications of the least or weighted least squares; one that can be applied when the sample trees are selected by cluster sampling, the other when the data of the sample trees are selected by a two-phase or double sampling design. The basic discussions on these two modifications can be found in Cunia (1981, 1982), Briggs and Cunia (1982) and Cunia and Michelakackis (1983). Using simulation techniques, these modified least squares procedures were later tested by Gillespie and Cunia (1986) and Michelakackis and Cunia (1986), among others.

Sample Tree Selection by Cluster Sampling Method

The method of cluster sampling as considered here can be simply described as the selection, by some random procedure of groups or clusters of sample trees (rather than individual trees) from a given forest area of interest. Usually, these clusters consist of trees growing within sample plots of fixed area or trees counted by relascope from fixed points in the forest. They may also consist of subsamples (fixed number or percentages) of trees selected from randomly selected plots or points.

Ideally, this requires (i) a prior subdivision of the tree population into non-overlapping plots of fixed area, and if some conditions are satisfied, overlapping plots or relascope points, (ii) a random sample of plots selected by some statistical sampling procedure and possibly (iii) a random subsample of trees from the plots so selected. Strictly speaking, the cluster sampling (or one-stage cluster sampling) is defined when all the trees of the sample plots are included in the sample and two-stage sampling (or two-stage cluster sampling) when each sample plot is subsampled. For convenience, both methods will be known here simply as cluster sampling.

The way the cluster sampling is actually applied to selection of sample trees for construction of biomass tables is, however, somewhat different. Points are selected in the forest, more or less at random and an arbitrary number of sample trees are selected, by some random, arbitrary or subjective procedure from the forest area around these points. These trees are measured for the characteristics of interest (diameter, height, species, site quality, etc. and

biomass of various tree components) and they constitute the sample trees from which the biomass regression function is being estimated.

The main advantage of cluster sampling is the large decrease in the average sampling costs per tree; as more and more trees are measured from the same location, there is less and less ground movement of the logging equipment and crew. It is also more convenient to work in a few areas than go from tree to tree all over the forest. But there are also disadvantages. One major disadvantage is that the average amount of information per tree decreases with the increase of the average number of sample trees per cluster; trees growing close to each other differ less among themselves than trees growing farther apart. A second major disadvantage is that, unless the least squares techniques are appropriately modified, the validity of the statistical inferences made under the standard assumptions of the least squares method are highly questionable.

Several such modifications have been proposed and applied to samples of trees. A first modification makes use of the theory of ratio estimators as described in standard textbooks on sampling techniques. It is described, and the results of its application discussed by Kotimaki and Cunia (1981), Cunia and Gillespie (1985) and Gillespie and Cunia (1986). This modification is not considered here. A second modification makes use of linear regression techniques with dummy variables. It is described by Cunia (1986b) but, to my knowledge, it has never been applied. It is also not considered here. A third, probably the best modification is to make use of the generalized least squares method. It is a complex procedure, to my knowledge has never been applied to sample trees selected by cluster sampling and is of no further interest here. The last modification, and this is the modification we shall discuss in this paper, uses the ordinary weighted least squares applied to a set of appropriately defined cluster variables; the usual techniques are applied to individual tree variables.

Modification of the Least Squares Regression Method: Trees Selected by Cluster Sampling

Let us make the following, more formal assumptions.

(1) The sample trees are selected in clusters (plots) and the trees are measured for the dependent variable y (biomass of some tree component) and the independent variables x_1, x_2, \dots, x_m (tree measurements other than biomass).

(2) The true regression function of y on $[x]$ is of the form

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]' [x]$$

and the conditional variance of y given $[x]$ is proportional to some known tree value a^2 , that is

$$\sigma_{yy|x} = a^2 \sigma^2$$

where a^2 but not σ^2 is known.

(3) The sample values y within a cluster may be correlated, but the values y belonging to different clusters are uncorrelated.

There are also additional implicit assumptions not mentioned here as, for example, that conditional probabilities distribution of y given $[x]$ is normal, that the variables x are fixed variables measured without error, that the number of clusters is greater than the number m of variables, that we deal with static populations that do not change with time, etc.

Let us now sum up, for each cluster, the tree variables y, x_1, x_2, \dots, x_m , and write the new cluster variables

$$t = \Sigma y, s_1 = \Sigma x_1, s_2 = \Sigma x_2, \dots, s_m = \Sigma x_m$$

where Σ means summation over the trees of a given cluster. For example, if the cluster is a plot and $y =$ tree biomass, $x_1 = 1$, $x_2 =$ tree diameter d and $x_3 = d^2$, the new cluster variables are

$$t = \Sigma y = \text{plot biomass}$$

$$s_1 = \Sigma x_1 = \text{number of trees per plot}$$

$$s_2 = \Sigma x_2 = \text{sum of tree diameters } d \text{ per plot}$$

$$s_3 = \Sigma x_3 = \text{sum of squared diameters } d^2 \text{ per plot}$$

As the regression values y of the biomass of the trees of a given plot can be written as

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

we can also write the regression value of the plot biomass as

$$\begin{aligned} \hat{t} &= \Sigma \hat{y} = \Sigma (\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) \\ &= \beta_1 s_1 + \beta_2 s_2 + \dots + \beta_m s_m = [\beta]' [s] \end{aligned}$$

Although this looks very much like the true regression of t on $[s]$, it is not. The true regression of the plot biomass t on plot variables $[s]$ is defined as the conditional expected value of t given $[s]$. Its form is not necessarily the same as that of the true regression of tree biomass y on tree variables $[x]$. Various subdivisions of the population of trees into sets of plots yields various regressions of t . The fact that the regression of y on $[x]$ is linear does not mean that the regression of t on $[s]$ is also linear.

Let us consider now the total biomass of the given forest population. Obviously, the total plot biomass is equal to the total tree biomass and also equal to the total regression values of the tree biomass. This means that we can write

$$\begin{aligned} \Sigma y &= \Sigma \hat{y} = \Sigma \hat{t} = \Sigma t \\ &= \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \dots + \beta_m \Sigma x_m \\ &= \beta_1 \Sigma s_1 + \beta_2 \Sigma s_2 + \dots + \beta_m \Sigma s_m \end{aligned}$$

and we can view the function $t = [\beta]' [s]$ as a model for the regression function of plot biomass on plot variables $[s]$. This in turn implies that we can use the sample plot data to estimate $[\beta]$. Furthermore, because t is the sum of several variables y with variances $a^2 \sigma^2$, it is reasonable to assume that the conditional variance of t is at least proportional, if not equal to the sum $\sigma^2 \Sigma a^2$ of variances, where again Σ means summation over the trees of a given plot.

Let us reverse somehow the process and make the following assumptions.

(1) The true regression function of the plot biomass t on plot variables s_1, s_2, \dots, s_m is of the linear form

$$\hat{t} = \beta_1 s_1 + \beta_2 s_2 + \dots + \beta_m s_m = [\beta]' [s]$$

and the conditional variance of t given $[s]$ is proportional to the known sum of tree values Σa^2 .

(2) The sample plot values t are uncorrelated random variables.

(3) The probability distribution of t given $[s]$ is normal; this assumption is needed only for testing null hypotheses or calculating interval estimates.

Then, the weighted least squares estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ are the least squares estimates of the regression of the transformed variables $u = t / \sqrt{\Sigma a^2}$ on $v_1 = s_1 / \sqrt{\Sigma a^2}, v_2 = s_2 / \sqrt{\Sigma a^2}, \dots, v_m = s_m / \sqrt{\Sigma a^2}$. That is, if we define the matrices of sample plot values u and v_1, v_2, \dots, v_m as $[U]$ and $[V]$ and the matrices of the sums of crossproducts as $[T] = [V]' [V]$ and $[P] = [V]' [U]$, then

$$[b] = [T]^{-1} [P]$$

$$[S_{bb}] = S_{uu|v} [T]^{-1}$$

$$\text{where } S_{uu|v} = ([U]' [U] - [b]' [P]) / (q - m)$$

where q is the number of sample clusters.

Note what we are doing. We assume a linear regression function $t = [\beta]' [s]$ of the plot biomass t on plot variables s_1, s_2, \dots, s_m and, using the sample plot data, we calculate the estimates $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$. The total biomass in the forest area can now be written successively as the following expressions

$$\Sigma t = \Sigma \hat{t} = \Sigma [\beta]' [s]$$

This is due to a property of the linear least squares regressions; the sum of regression values is equal to the sum of actual values.

$$\begin{aligned} \Sigma \hat{t} &= \beta_1 \Sigma s_1 + \beta_2 \Sigma s_2 + \dots + \beta_m \Sigma s_m \\ &= \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \dots + \beta_m \Sigma x_m, \text{ because} \\ &\quad \text{the variables } s_i \text{ are defined as the sums} \\ &\quad \text{of the variables } x_i \text{ in a given plot} \\ &= \Sigma (\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) = \Sigma [\beta]' [x] \end{aligned}$$

$= \Sigma \hat{y} = \Sigma y$, because sum of the tree biomass (Σy) is identically equal to the sum of the plot biomass ($\Sigma \hat{x}$).

The vector $[\beta]$ of the function $\hat{y} = [\beta]'[x]$ is not the vector of coefficients of the regression function of tree biomass on $[x]$; it is the vector of the regression function of plot biomass on $[s]$. Nevertheless, the function $\hat{y} = [\beta]'[x]$ can serve as an approximation or a substitute for the true regression function of the tree biomass, and $\hat{y} = [b]'[x]$ may serve as an estimate of this regression function. Whether $\hat{y} = [b]'[x]$ is a good estimate of the true regression function of y on $[x]$ depends, among other things on how close the vector $[b]$ of the plot regression is to the vector $[\beta]$ of the tree regression. Only experience with actual forest biomass data will show whether the proposed modified procedure above yields acceptably good results.

It may be of interest to show that this modified least squares method makes sense when applied within the specific forest inventory design having plots, not trees as the sampling units. Consider, for example, the two-phase or double sampling design consisting of (i) a first phase, relatively large, simple random sample of plots whose trees are measured for the variables x_1, x_2, \dots, x_m and (ii) a second phase, relatively small, simple random sample of plots whose trees are measured for the biomass y in addition to the variables x_1, x_2, \dots, x_m of the first phase. The plots of the first phase provide plot measurements s_1, s_2, \dots, s_m and estimates $[\bar{s}] = [\bar{s}_1 \bar{s}_2 \dots \bar{s}_m]$ of the corresponding population means and $[S_{\bar{s}\bar{s}}]$ of the covariance matrix of $[\bar{s}]$. The plots of the second phase provide estimators $[b]$ of $[\beta]$ and $[S_{bb}]$ of the covariance matrix of $[b]$ associated with the biomass regression function

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = [\beta]'[x]$$

by the modified procedure presented above.

Then, the estimate of the mean biomass per plot is

$$\bar{t} = [b]'[\bar{s}]$$

and an estimate of the variance of \bar{t} is

$$S_{\bar{t}\bar{t}} = [b]'[S_{\bar{s}\bar{s}}][b] + [\bar{s}]'[S_{bb}][\bar{s}]$$

The estimator \bar{t} is known as the double sampling with (multiple linear) regression estimator. It can be shown, see Cunia (1986a) that the same estimator \bar{t} can be obtained by the usual procedure where (i) the biomass of each sample tree of the first phase is calculated by the regression $\hat{y} = [b]'[x]$, (ii) the biomass of each plot of the first phase is calculated as the sum of the biomass \hat{y} of all its trees (this is the same as the value given by the formula $\hat{t} = [b]'[s]$ applied to the variables s_1, s_2, \dots, s_m of the given plot) and (iii) the average \bar{t} is calculated as the average of the plot biomass values t .

An Illustrative Example

To show how this procedure can be applied we shall consider the following numerical example.

Example 1. Consider the data from the 353 sample trees used by Cunia (1986b) to estimate the following biomass regression,

$$\begin{aligned} \hat{y} &= b_1 + b_2 d + b_3 d^2 \\ &= b_1 x_1 + b_2 x_2 + b_3 x_3 = [b]'[x] \end{aligned}$$

where

y = total above ground tree biomass (pounds of green weight)

d = tree diameter at breast height (inches)

$$[b]' = [5.1818118 \quad -25.653078 \quad 12.988357]$$

= estimate of the vector $[\beta]$ of regression coefficients

and

$$[S_{bb}] = \begin{bmatrix} 8715.8855 & -2222.4882 & 128.69992 \\ -2222.4882 & 581.99570 & -34.776995 \\ 128.69992 & -34.776995 & 2.1744582 \end{bmatrix}$$

= estimate of the covariance matrix of $[b]$.

These trees have been assumed to have been selected by simple random sampling in order for Cunia (1986b) to illustrate the application of the ordinary weighted least squares method. In reality they were selected by simulated cluster sampling; more specifically by a two-stage simulated sampling procedure. In the first stage, 30 one-fifth acre sample plots were selected at random and without replacement, in the second stage 30 percent of their trees were selected, again by simple random sampling without replacement. As there is a total of 353 trees in the sample, the average number of sample trees per cluster is about 12; the number vary from a low of 1 to a high of 22 trees.

To apply the modified least squares procedure shown here we start with the grouping of the trees in clusters by plot number and the calculations of the following plot variables

t = plot biomass (Σy)

s_1 = number of trees ($\Sigma 1$) per plot

s_2 = sum of tree diameters (Σd) per plot

s_3 = sum of squared diameters (Σd^2) per plot

and

$\Sigma d^4 = a^2$ = the value, such that the conditional variance of t given $[s]$ is assumed proportional to.

Note that the ordinary weighted least squares method used by Cunia (1986b) assumes that the conditional variance of y given d is proportional to d^4 .

We continue with the calculation of the new transformed variables $u = t/\sqrt{\Sigma d^4}$ and $v_i = s_i/\sqrt{\Sigma d^4}$, $i=1,2,3$. The plot variables t, s_1, s_2, s_3 ,

$a = \sqrt{\sum d^4}$, u , v_1 , v_2 and v_3 are listed in Table 1. For convenience, some of the values were reported with a limited number of significant digits, even though, to calculate them the computer used double precision.

The matrices of the sums of cross products are

$$[T] = [V]'[V] = \begin{bmatrix} .04887035 & .38092564 & 3.2785336 \\ .38092564 & 3.0330882 & 26.811322 \\ 3.2785336 & 26.811322 & 245.29889 \end{bmatrix}$$

$$[P] = [V]'[U] = \begin{bmatrix} 33.258523 \\ 273.51112 \\ 2520.8582 \end{bmatrix}$$

and the inverse of $[T]$ is

$$[T]^{-1} = \begin{bmatrix} 12250.429 & -2696.1925 & 130.96286 \\ -2696.1925 & 603.15120 & -29.888939 \\ 130.96286 & -29.888939 & 1.5205770 \end{bmatrix}$$

As $[U]'[U] = 26860.109$, we finally determine the statistics of interest

$$[b]' = [131.33607 \quad -48.59752 \quad 13.833052]$$

$$S_{uu|v} = 33.809815$$

$$[S_{bb}] = \begin{bmatrix} 414184.73 & -91157.769 & 4427.8299 \\ -91157.769 & 20392.431 & -1010.5395 \\ 4427.8299 & -1010.5395 & 51.410428 \end{bmatrix}$$

At first sight these values look quite different from the values $[b]$ and $[S_{bb}]$ calculated by the ordinary weighted least squares method. However, we never work with individual values of the regression coefficients or their error; we always work with linear combinations. For example, if we wish to calculate the regression estimate of the tree biomass of a 10 inch diameter tree (that is, to calculate the estimate of the average biomass of all 10 inch trees in the forest) we use the formula

$$y = b_1 + 10b_2 + 100b_3$$

Using the ordinary weighted least squares as applied by Cunia (1986b) we obtain the value

$$y = 5.18 - (10)(25.653) + (100)(12.9884) = 1047 \text{ pounds}$$

The modified weighted least squares method presented here yields the value

$$y = 131.34 - (10)(48.600) + (100)(13.8331) = 1029$$

The difference is about 2 percent, well within the inherent sampling error.

The regression estimates and their 95 percent confidence limits, as calculated by the ordinary and the modified least squares procedures are shown in Table 2. Note that the 95 percent confidence limits are defined by the formula

$$[b]'[x] \pm 2\sqrt{[x]'[S_{bb}][x]}$$

where

$$[x]' = [1 \quad d \quad d^2]$$

for $d = 4, 5, \dots, 26$

For example, if $d = 10$, $[x]' = [1 \quad 10 \quad 100]$ and the 95 percent confidence limits become

$$1047.487 \pm 39.813$$

for a lower and upper limit of 1007.674 and 1087.300 respectively for the ordinary least squares and

$$1028.666 \pm 188.294$$

for a lower and upper limit of 840.372 and 1216.960 respectively for the modified least squares.

A look at Table 2 shows that the confidence limits calculated by the modified least squares are much wider. We did expect to have them wider but not by that much. One of the reasons may be the sampling error; we happened to work with a highly unusual sample. Another reason may be that the simulated tree population from which the sample trees were selected is not representative of the real life populations; we have, in the simulated population a much larger variance "between" clusters (relative to the variance "within" clusters) than what we ordinarily have in real life. A third reason is the fact that the two sets of confidence limits refer to two different things; for the ordinary least squares the limits refer to the average biomass per tree for the class of all population trees that have the same diameter, while for the modified least squares they refer to the average biomass per plot (cluster) of all population plots that have one tree of the given diameter. As our sample has no such plot (and plots with one tree only are very rare), the biomass table constructed by the modified least squares contains regression estimates calculated for values of the independent variables s that fall outside the sample data.

It may be of interest to see the impact of the least squares estimation procedure on the estimates of the mean biomass per acre and its error when the biomass regression is applied to an actual forest inventory data. Let us refer to Example 1 of Cunia (1986a) where the ordinary weighted least squares biomass regression was applied to a sample of 926 one-fifth acre plots. He has then found the following results

$$w = 120480 \text{ pounds}$$

= estimate of the mean biomass per acre

$$S_{ww}^{(1)} = 12415782 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is ignored}$$

$$S_{ww} = 19059566 = \text{estimate of the variance of } w \text{ when the error of the biomass regression is accounted for}$$

As the reader can verify, 65.1 percent of the variance is associated with the error of the sample plots and 34.9 percent is associated with the error of the biomass regression.

Table 1 - The plot variables t = biomass (pounds), s_1 = number of trees, s_2 = sum of diameters d , s_3 = sum of squared diameters d^2 , $a = \sqrt{2}d^4$, $u = t/a$, $v_1 = s_1/a$, $v_2 = s_2/a$ and $v_3 = s_3/a$ used in Example 1.

Plot	t	s_1	s_2	s_3	a	u	v_1	v_2	v_3
1	11726	13	110.4	1018.98	333.19	35.19	.03902	.3313	3.058
2	795	2	15.0	118.98	92.39	8.60	.02165	.1624	1.288
3	16870	16	151.7	1545.77	437.49	38.56	.03657	.3467	3.533
4	6580	3	33.6	571.58	512.61	12.84	.00585	.0655	1.115
5	12705	12	97.3	854.09	296.55	42.84	.04046	.3281	2.880
6	20753	21	164.2	1605.20	595.35	34.86	.03527	.2758	2.696
7	9859	12	92.2	783.68	278.50	35.40	.04309	.3311	2.814
8	10083	10	86.1	827.35	315.37	31.97	.03171	.2730	2.623
9	355	1	6.6	43.56	43.56	8.15	.02296	.1515	1.000
10	3599	6	47.6	414.86	201.19	17.89	.02982	.2366	2.062
11	1813	2	15.4	133.16	111.01	16.33	.01802	.1387	1.200
12	7902	16	108.9	767.25	205.49	38.45	.07786	.5299	3.734
13	6220	6	57.0	671.34	398.56	15.61	.01505	.1430	1.684
14	16828	22	189.4	1752.98	434.61	38.72	.05062	.4358	4.033
15	16512	18	164.9	1627.29	443.43	37.24	.04059	.3719	3.670
16	20102	22	202.4	2094.80	550.56	36.51	.03996	.3676	3.805
17	15420	11	119.3	1604.55	712.55	21.64	.01544	.1674	2.252
18	4727	14	95.6	695.42	215.76	21.91	.06489	.4431	3.223
19	7270	15	108.9	825.61	229.83	31.63	.06527	.4738	3.592
20	18169	11	134.2	1912.90	767.99	23.66	.01432	.1747	2.491
21	10737	16	123.4	1040.30	321.68	33.38	.04974	.3836	3.234
22	9043	19	163.1	1505.33	400.50	22.58	.04744	.4072	3.759
23	12248	13	109.5	1096.81	404.10	30.31	.03217	.2710	2.714
24	8074	15	101.1	695.29	186.88	43.20	.08027	.5410	3.721
25	327	2	12.0	72.00	50.91	6.42	.03928	.2357	1.414
26	16994	13	135.8	1503.48	465.06	36.54	.02795	.2920	3.233
27	14575	9	99.1	1208.87	469.34	31.05	.01918	.2111	2.576
28	13405	6	75.9	1097.79	592.85	22.61	.01012	.1280	1.852
29	18541	17	160.0	1667.92	482.82	38.40	.03521	.3314	3.455
30	9699	10	96.0	995.70	358.77	27.03	.02787	.2676	2.775

Table 2 - The biomass tables and their 95 percent confidence limits as constructed in Example 1 by the ordinary weighted least squares (OWLS) and modified weighted least squares (MWLS); the above-ground tree biomass is given in pounds of green weight.

diameter inches	OWLS			MLS		
	lower limit	biomass estimate	upper limit	lower limit	biomass estimate	upper limit
4	67	110	154	0	158	541
5	118	202	225	3	234	465
6	304	319	334	223	337	453
7	443	462	482	396	469	542
8	605	631	657	513	628	743
9	794	826	859	655	814	974
10	1008	1047	1087	840	1029	1217
11	1244	1295	1345	1067	1271	1474
12	1500	1568	1635	1323	1540	1757
13	1775	1867	1958	1592	1837	2083
14	2070	2192	2313	1855	2162	2469
15	2385	2543	2701	2106	2515	2924
16	2719	2920	3121	2345	2895	3445
17	3073	3323	3573	2577	3303	4029
18	3446	3752	4057	2803	3738	4674
19	3840	4207	4573	3025	4202	5378
20	4254	4687	5121	3245	4693	6141
21	4687	5194	5701	3462	5211	6951
22	5141	5727	6314	3677	5757	7838
23	5614	6286	6958	3890	6331	8772
24	6108	6871	7634	4102	6933	9763
25	6622	7482	8341	4313	7562	10811
26	7155	8118	9081	4522	8219	11916

Using the biomass regression estimated by the modified weighted least squares method and the method and statistics of Cunia (1986a) example, we find

- $w = 120672$ pounds
 = estimate of the mean biomass per acre
- (1) $S_{ww} = 12280114$ = estimate of the variance of w when the error of the biomass regression is ignored
- $S_{ww} = 49544699$ = estimate of the variance of w when the error of the biomass regression is accounted for

In this case, 24.8 percent of the error (expressed as variance) is due to sample plots and 75.2 percent is due to the biomass regression.

In terms of the 95 percent confidence limits, an expression that is more meaningful to the layman, we have the following

- (1) When the ordinary weighted least squares method is applied
 and $w \pm 2\sqrt{S_{ww}^{(1)}} = (120480 \pm 7047)$ pounds
 $w \pm 2\sqrt{S_{ww}} = (120480 \pm 8731)$ pounds
- (2) When the modified weighted least squares method is applied
 and $w \pm 2\sqrt{S_{ww}^{(1)}} = (120672 \pm 7009)$ pounds
 $w \pm 2\sqrt{S_{ww}} = (120672 \pm 14078)$ pounds

As the reader can verify, the effect of the estimation procedure has a negligible effect on the value of the mean biomass per acre estimate, but may have a critical effect on the estimation of its error. Of course, the values above refer to a simulated population and the corresponding values for natural populations may be different. However, these results may show that the estimation may have a critical effect on the values of the estimates of the error.

Sample Tree Selection by Double Sampling

A common method used in the selection of sample trees (for biomass tables construction) consists of two main phases. In the first phase, a relatively small sample of trees is selected by some random procedure, and the trees are measured for diameter d , height h and biomass y . In the second phase, a relatively large sample of trees is selected, again by some random procedure and the trees are measured for diameter d and height h alone; these trees are not measured for biomass y . The data of the trees from the first phase sample are used to estimate a relationship between biomass, diameter and height; the data of the second phase sample trees are used to estimate a relationship between tree height and diameter; and finally, the two relationships are combined, in some way, to estimate the regression function of tree biomass on diameter.

There are two major computational procedures to combine the data from the two phases. They

are described, among others, by Clutter et al (1983) for forest mensurationists and by Cunia (1982) for forest biometricians. Here is a short description of these procedures.

Procedure 1 - The regression function of the tree biomass y on diameter d and height h , say $\hat{y} = r_1(d, h)$, is first estimated from the sample of tree data of the first phase. The regression of tree height h on tree diameter d , say $\hat{h} = r_2(d)$, is estimated next from the data of the second phase sample trees. Then, the estimate of the regression function of the biomass y on diameter alone d is defined as

$$\hat{y} = r(d) = r_1(d, \hat{h}) = r_1(d, r_2(d))$$

This means that the average biomass per tree of given diameter d_0 is estimated by the regression value $\hat{y}_0 = r_1(d_0, \hat{h}_0)$, where \hat{h}_0 is calculated as $\hat{h}_0 = r_2(d_0)$, the estimate of the average height of the tree of the given diameter d_0 .

This procedure has been used in forest inventory for a long time. Data from a sample of trees, usually selected from logging operations, are used to construct a standard two-way volume table (by diameter and height) by graphical first and by least squares techniques later. The trees from a second sample, usually selected from the sample plots of the current inventory are measured for diameter and height. Their data are used to estimate a relationship between tree diameter and height, again by graphical or least squares methods. The one-way local volume table (on diameter alone) is finally constructed from the standard two-way volume table as follows: the average volume of a tree of diameter d_0 is estimated as the volume given by the two-way table for (i) the given diameter d_0 and (ii) the height h_0 obtained from the relationship diameter-height of the second sample above.

Procedure 2 - As in Procedure 1 above, the sample trees of the first phase are measured for biomass y , diameter d and height h and their data are used to estimate the regression function $\hat{y} = r_1(d, h)$. The second phase sample trees are measured for diameter d and height h . Instead of estimating a relationship diameter-height, however, the one way tree biomass table is constructed as follows. First, the biomass y of each individual tree of the second phase sample is estimated by the regression function $\hat{y} = r_1(d, h)$ of the first phase. And then, the regression function $\hat{y} = r(d)$ is estimated from the measured values d and the estimated values \hat{y} of the second phase trees by the usual (graphical or) least squares techniques.

While the procedures for the estimation of the volume or biomass regressions above were known for a long time (and they are also of common use today) little was done to estimate the error of these regressions. An attempt was made lately by Cunia (1982) who proposed two approaches to estimate the error of the regression function of tree biomass on diameter, one for each procedure above, when (i) the samples of the two phases are statistically independent, (ii) all regression functions are assumed to be linear

and (iii) valid estimates of the covariance matrices of the regression coefficients of $\hat{y}=r_1(d,h)$ and $\hat{h}=r_2(d)$ for Procedure 1 and $\hat{y}=r_1(d,h)$ for Procedure 2 are given.

The approach for Procedure 1 was further described, and its applicability demonstrated, by Cunia and Michelakackis (1983). Using simulation techniques, Cunia and Michelakackis (1986) have shown that this approach leads to valid and reliable estimates of the error. It is this approach we shall discuss here. The other approach devised for Procedure 2 has never been applied, to my knowledge, and will not be considered here.

Modification of the Least Squares Regression
Method: Trees Selected by Double Sampling

To simplify the description of the procedure we shall assume that the three regressions are of the form

$$\hat{y} = r_1(d,h) = \alpha_1 + \alpha_2 d^2 h + \alpha_3 d + \alpha_4 h + \alpha_5 d h + \alpha_6 d^2$$

$$\hat{h} = r_2(d) = \gamma_1 + \gamma_2 d + \gamma_3 d^2$$

and

$$\hat{y} = r(d) = \alpha_1 + \alpha_2 d^2 (\gamma_1 + \gamma_2 d + \gamma_3 d^2) + \alpha_3 d + \alpha_4 (\gamma_1 + \gamma_2 d + \gamma_3 d^2) + \alpha_5 d (\gamma_1 + \gamma_2 d + \gamma_3 d^2) + \alpha_6 d^2$$

$$= \beta_1 + \beta_2 d + \beta_3 d^2 + \beta_4 d^3 + \beta_5 d^4 = [\beta]' [x]$$

where

$$\beta_1 = (\alpha_1 + \alpha_4 \gamma_1), \quad \beta_2 = (\alpha_3 + \alpha_4 \gamma_2 + \alpha_5 \gamma_1)$$

$$\beta_3 = (\alpha_2 \gamma_1 + \alpha_4 \gamma_3 + \alpha_5 \gamma_2 + \alpha_6)$$

$$\beta_4 = (\alpha_2 \gamma_2 + \alpha_5 \gamma_3), \quad \beta_5 = \alpha_2 \gamma_3$$

and the definition of [x] is straightforward. We have used, for convenience, different notation to denote the coefficients of the different regressions.

Let us assume that we are given estimates [a] of [a] and [c] of [Y] and estimates [S_{aa}] and [S_{cc}] of their covariance matrices. Then, the estimate [b] of [β] is easily derived as

$$[b] = \begin{bmatrix} a_1 + a_4 c_1 \\ a_3 + a_4 c_2 + a_5 c_1 \\ a_2 c_1 + a_4 c_3 + a_5 c_2 + a_6 \\ a_2 c_2 + a_5 c_3 \\ a_2 c_3 \end{bmatrix}$$

The objective now is to calculate [S_{bb}], an estimate of the covariance matrix of [b].

It can be shown, see Cunia (1982), Cunia and Michelakackis (1983) and Michelakackis and Cunia (1986) that [S_{bb}] can be defined as

$$[S_{bb}] = [C][S_{aa}][C]' + [A][S_{cc}^*][A]'$$

where

$$[A] = \begin{bmatrix} a_1 & a_4 & 0 & 0 \\ a_3 & a_5 & a_4 & 0 \\ a_6 & a_2 & a_5 & a_4 \\ 0 & 0 & a_2 & a_5 \\ 0 & 0 & 0 & a_2 \end{bmatrix}$$

$$[C] = \begin{bmatrix} 1 & 0 & 0 & c_1 & 0 & 0 \\ 0 & 0 & 1 & c_2 & c_1 & 0 \\ 0 & c_1 & 0 & c_3 & c_2 & 1 \\ 0 & c_2 & 0 & 0 & c_3 & 0 \\ 0 & c_3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[S_{aa}] = estimator of the covariance matrix of [a]

[S_{cc}] = estimator of the covariance matrix of [c]

and

$$[S_{cc}^*] = \begin{bmatrix} 0 & [0] \\ [0]' & [S_{cc}] \end{bmatrix}$$

= estimator of the covariance matrix of the extended vector [c*]' = [1 c₁ c₂ c₃]

It may be interesting to note that, in matrix notation

$$[b]' = [C][a] = [A][c^*]$$

Some of the coefficients a and c may be made equal to zero (when they are not significantly different than zero). Then, the formulae above still apply since the corresponding rows and columns of [S_{aa}] and [S_{cc}] will also be made equal to zero. As an example we shall assume that a₅=a₆=0, the case used in the example of the next section. Then

$$[a]' = [a_1 \quad a_2 \quad a_3 \quad a_4 \quad 0 \quad 0]$$

$$[A] = \begin{bmatrix} a_1 & a_4 & 0 & 0 \\ a_3 & 0 & a_4 & 0 \\ 0 & a_2 & 0 & a_4 \\ 0 & 0 & a_2 & 0 \\ 0 & 0 & 0 & a_2 \end{bmatrix}$$

$$[S_{aa}] = \begin{bmatrix} S_{a_1 a_1} & S_{a_1 a_2} & S_{a_1 a_3} & S_{a_1 a_4} & 0 & 0 \\ S_{a_1 a_2} & S_{a_2 a_2} & S_{a_2 a_3} & S_{a_2 a_4} & 0 & 0 \\ S_{a_1 a_3} & S_{a_2 a_3} & S_{a_3 a_3} & S_{a_3 a_4} & 0 & 0 \\ S_{a_1 a_4} & S_{a_2 a_4} & S_{a_3 a_4} & S_{a_4 a_4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $[c^*]$ and $[C]$ and $[S_{cc}^*]$ remain the same.

An Illustrative Example

To show how the procedure of the previous section can be applied we shall use an example already contained in a paper by Cunia and Michelakackis (1983). For more details the interested reader should refer to it.

Example 2 - Between 1967 and 1969, two hundred and eighty Norway Black Spruce trees were selected by cluster sampling from all over Finland by the Finnish Forest Research Institute and measured, among other things, for their diameter at breast height d (cm), total height h (m) and total biomass (kg). For illustration purposes, the 280 trees were divided into two groups; a first group of 56 trees (the first phase sample) were assumed measured for y , d and h and a second group of the remaining 224 trees (the second phase sample) were assumed measured for d and h alone. For convenience we shall also assume that the sample trees were selected by simple random (not cluster) sampling and that the trees of the first are statistically independent of the second group.

We start with the first phase sample of 56 trees. Assuming that (i) the true regression function of y on d and h is

$$\hat{y} = \alpha_1 + \alpha_2 d^2 h + \alpha_3 d + \alpha_4 h + \alpha_5 dh + \alpha_6 d^2$$

(ii) the conditional variance of y given d and h is proportional to $d^4 h^2$ and (iii) the null hypothesis $\alpha_5 = \alpha_6 = 0$ has been accepted, Cunia and Michelakackis (1983) have found the following statistics (where, for convenience, we have reported them with a relatively small number of significant digits)

$$[a]' = [-3.42213 \quad .0167680 \quad 3.72990 \quad -2.07425 \quad 0 \quad 0]$$

= estimate of the vector $[a]'$ of regression coefficients

and

$$[S_{aa}] = \begin{bmatrix} [S_{aa}^*] & [O_1] \\ [O_1]' & [O_2] \end{bmatrix}$$

= estimate of the covariance matrix of $[a]$,

where

$$[S_{aa}^*] = \begin{bmatrix} 12.6354 & .004429 & -.985188 & -.919102 \\ .004429 & .00000239 & -.000492 & -.000249 \\ -.985188 & -.000492 & .215943 & -.060859 \\ -.919102 & -.000249 & -.060859 & .203787 \end{bmatrix}$$

$$[O_1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad [O_2] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The data of the 224 trees of the second phase sample was used to estimate the regression function

of the tree height on diameter. It was assumed that the form of the regression function is

$$\hat{h} = \gamma_1 + \gamma_2 d + \gamma_3 d^2$$

and that the conditional variance of h given d is homogeneous. Cunia and Michelakackis (1983) report the following least squares statistics,

$$[c]' = [1.098167 \quad .862612 \quad -.005809]$$

= estimate of the vector $[Y]'$ of regression coefficients

and

$$[S_{cc}] = \begin{bmatrix} .855041 & -.0931381 & .0021821 \\ -.0931381 & .0107251 & -.00026127 \\ .0021821 & -.00026127 & .000006658 \end{bmatrix}$$

= estimate of the covariance matrix of $[c]$.

We can now apply our procedure to estimate the regression function of the tree biomass on diameter alone and its error. We start with the calculation of the vector $[b]$ of coefficients of the regression

$$\hat{y} = r(d) = r_1(d, r_2(d)) = b_1 + b_2 d + b_3 d^2 + b_4 d^3 + b_5 d^4$$

For convenience, the five elements of $[b]$ are listed individually as

$$\begin{aligned} b_1 &= a_1 + a_4 c_1 = -5.700010 \\ b_2 &= a_3 + a_4 c_2 + a_5 c_1 = 1.940622 \\ b_3 &= a_2 c_1 + a_4 c_3 + a_5 c_2 + a_6 = .0304628 \\ b_4 &= a_2 c_2 + a_5 c_3 = .01446426 \\ b_5 &= a_2 c_3 = -.00009740065 \end{aligned}$$

To estimate the covariance matrix of $[b]$ we write first the vectors and matrices

$$[c^*]' = [1 \quad 1.098167 \quad .862612 \quad -.005809]$$

$$[S_{cc}^*] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & .855041 & -.0931381 & .0021821 \\ 0 & -.0931381 & .0107251 & -.00026127 \\ 0 & .0021821 & -.00026127 & .000006658 \end{bmatrix}$$

$[a]$ and $[S_{aa}]$ as shown above

$$[C] = \begin{bmatrix} 1 & 0 & 0 & 1.098167 & 0 & 0 \\ 0 & 0 & 1 & .862612 & 1.098167 & 0 \\ 0 & 1.098167 & 0 & -.005809 & .862612 & 1 \\ 0 & .862612 & 0 & 0 & -.005809 & 1 \\ 0 & -.005809 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[A] = \begin{bmatrix} -3.42213 & -2.07425 & 0 & 0 \\ 3.72990 & 0 & -2.07425 & 0 \\ 0 & .0167680 & 0 & -2.07425 \\ 0 & 0 & .0167680 & 0 \\ 0 & 0 & 0 & .0167680 \end{bmatrix}$$

Now the reader can verify that the previous value of [b] is obtained by the matrix multiplications

$$[b] = [C][a] = [A][c^*]$$

and that the elements of $[S_{bb}]$, denoted here, for convenience, as

$$S_{ij} = \text{covariance of } b_i \text{ and } b_j$$

are the following

$$\begin{aligned} S_{11} &= 14.5413 & , & S_{12} = -2.05253 \\ S_{13} &= -.011749 & , & S_{14} = .0068236 \\ S_{15} &= -.000100030 & , & S_{22} = .0308731 \\ S_{23} &= .000671164 & , & S_{24} = -.000983024 \\ S_{28} &= .000013195 & , & S_{33} = .000130202 \\ S_{34} &= -.000013585 & , & S_{35} = .00000035829 \\ S_{44} &= .000004797 & , & S_{45} = -.00000008545 \\ S_{55} &= .000000001953 \end{aligned}$$

One can estimate now the biomass table and its 95 percent confidence limits by the formulae

$$\hat{y} = [b]'[x]$$

$$S_{\hat{y}\hat{y}} = [x]'[S_{bb}][x]$$

$$\hat{y} \pm 2\sqrt{S_{\hat{y}\hat{y}}}$$

where

$$[x]' = [1 \quad d \quad d^2 \quad d^3 \quad d^4]$$

for $d = 4, 5, \dots$

This table is reported by Cunia and Michelakackis (1983) and, thus, it is not repeated here.

Acknowledgements

This paper is based on research funded by the Research Foundation of the State University of New York, the United States Department of Agriculture Forest Service, and the Department of Energy, Grant No. 23-524.

Literature Cited

Briggs, E.F.; Cunia, T. Effect of cluster sampling in biomass tables construction: linear regression models. Canadian Journal of Forest Research 12: 255-263; 1982.

Cunia, T. On tree biomass tables and regressions: some statistical comments. In: 1979 forest resource inventories workshop proceedings, Vol. II, W.E. Frayer, (Ed.) Colorado State University, Fort Collins, CO, 1979a.

Cunia, T. On sampling trees for biomass tables construction: some statistical comments. In: 1979 forest resource inventories workshop proceedings, Vol. II, W.E. Frayer, (Ed.), Colorado State University, Fort Collins, CO; 1979b.

Cunia, T. Cluster sampling and tree biomass tables construction. In: Interdivisional Proceedings, 17th IUFRO World Congress, September 6-12, 1981, Kyoto, Japan; 1981.

Cunia, T. On the error of tree volume tables and its effect on the precision of forest inventory estimates. In: Statistics in theory and practice: essays in honor of Bertil Matern. B. Ranney (Ed.). Swedish University of Agricultural Sciences, Section of Biometry, S-90183, Umea, Sweden; 1982.

Cunia, T. Error of forest inventory estimates: its main components. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986a.

Cunia, T. Construction of tree biomass tables by linear regression techniques. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986b.

Cunia, T. Use of dummy variables techniques in the estimation of biomass regressions. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986c.

Cunia, T.; Gillespie, A.J. Cluster sampling and construction of biomass tables: results of a simulation study. In: Proceedings, third annual southern forest biomass workshop, March 12-14, 1985; University of Florida, Gainesville, FL; 1985.

Cunia, T.; Michelakackis J. On the error of tree biomass tables constructed by a two-phase sampling design. Canadian Journal of Forest Research. 13:303-313; 1983.

Gillespie, A.J.; Cunia, T. Estimation of tree biomass tables by cluster sampling: results of a simulated study. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.

Kotimaki, T.A.; Cunia, T. Effect of cluster sampling in biomass tables construction: ratio estimators models. Canadian Journal of Forest Research 11: 475-486; 1981.

Michelakackis, J.; Cunia, T. Error of biomass regressions: sample trees selected by double sampling. In: Proceedings of the Workshop on "Tree biomass regression functions and their contribution to the error of forest inventory estimates ", May 26-30, 1986, SUNY College of Environmental Science and Forestry, Syracuse, NY; 1986.