Use of Logarithmic Regression in the Estimation of Plant Biomass[1]

G. L. BASKERVILLE

Canadian Forestry Service, P.O. Box 4000, Fredericton, New Brunswick.

Received August 24, 1971

BASKERVILLE, G. L. 1972. Use of logarithmic regression in the estimation of plant biomass. Can. J. Forest Res. 2, 49–53.

The basic assumptions of regression analysis are recalled with special reference to the use of a logarithmic transformation. The limitations imposed on inference-making by failure to comply with these assumptions are discussed and ways to avoid the limitations indicated. A systematic bias of the order of 10 to 20% which is inherent in most, if not all, prior uses of the logarithmic equation to estimate plant biomass is noted as is the correction for the bias.

BASKERVILLE, G. L. 1972. Use of logarithmic regression in the estimation of plant biomass. Can. J. Forest Res. 2, 49–53.

Les hypothèses de base de l'analyse par régression sont énoncées avec référence spéciale à l'emploi de la transformation logarithmique. Les restrictions imposées sur l'inférence, à défaut de satisfaire les hypothèses, sont discutées et les moyens d'éviter les restrictions sont indiqués. Un biais systématique de 10 à 20% qui est inhérent dans la plupart, sinon tous, les emplois antérieurs de l'équation logarithmique pour l'estimation de la biomasse des plantes, est noté comme la correction pour le biais.

Introduction

The most common procedure for estimating biomass in forest stands is through the use of regressions and stand tables. A few stems are destructively sampled and the weight of each component determined and related by regression to some dimension of the standing tree. A stand table which classifies stems per unit area by units of the dimension used in the regression is then expanded to an estimate of biomass by multiplying the number of stems in each dimension class by the weight (estimated from regression) for that class. This general approach has been common for many years and had been called allometry in Europe and Japan (Kira and Shidei 1967) and dimensional analysis in North America (Whitaker and Woodwell 1968).

The weight of a plant component usually can be plotted over some dimension (*e.g.* diameter, height, or a combination thereof) to yield a straight line on double-log paper. Thus it has been expedient to calculate regressions as linear in the logarithms of the variables and to transform back to arithmetic units by determining the antilogarithm for the expansion of the stand table to biomass. Occasionally regressions have been calculated in terms of combinations of x-variables (usually D^2H) which give a linear relation in arithmetic units. Avoidance of the logarithm may be dangerous when it leads to violation of necessary assumptions of regression analysis.

This paper briefly reviews the assumptions of regression and the reasons for using a transformation and calls attention to the appropriate way of converting estimates from a logarithmic equation back to arithmetic units. These considerations are inherent in any use of a logarithm transformation and not limited to calculations of plant biomass. However, it is shown that in the past, misinterpretation of estimates from logarithmic equations has resulted in underestimates of biomass in most, if not all, cases where the logarithmic transformation has been used. While the ready access to computers today makes perpetuation of the error almost automatic, it is not difficult to seek and use methods that are appropriate to each data set and will remove the error.

The Problem

In the general case, we have two variables Y and X such that, on double-log paper, the plot of Y on X yields a straight line. The relationship suggested is that of the allometric equation

[1]Research sponsored by the U.S. Atomic Energy Commission under contract with the Union Carbide Corporation.

[1] $$Y = \beta X^{\alpha}$$

We require an efficient and unbiased expression of this relation which will permit (a) the estimation of $\hat{Y}_i$ with limits of uncertainty given X_i and, (b) the comparison of the parameters β and α among independent data sets.

Solution for the parameters β and α can be accomplished in arithmetic units by computer programs using an iterative least-squares technique which minimizes the sum of squares

[2] $$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

where N is the number of paired observations. Alternatively, equation [1] can be written in logarithmic form, either base e or base 10,

[3] $$\mathrm{LN}(Y) = \mathrm{LN}\,\beta + \alpha\,\mathrm{LN}\,(X)$$

which is linear. The parameters of this equation can be estimated by solving as in ordinary linear regression minimizing the sum of squares

[4] $$\sum_{i=1}^{N} \left\{\mathrm{LN}(Y_i) - \widehat{\mathrm{LN}(Y_i)}\right\}^2$$

The sums of squares given in [2] and [4] are not equivalent and the importance of choosing the proper regression model when solving for β and α has recently been emphasized (Zar 1968).

The Right Model

There are three assumptions fundamental to a least-squares regression:

1) It is assumed that for each X there is a normally distributed population of Y from which the sample Y's used in the regression are taken as a random sample. Failure to comply with this assumption will limit the inferences that can be made regarding the original population.
2) It is assumed that the true means, μ, of all the sampled populations fall along a given path, for example in the linear model $\mu = a + bX$. Failure to comply with this assumption will result in a systematic bias in estimated values of Y.
3) It is assumed that the variance, σ^2, is the same for all the populations. That is, the populations of Y at every X are normally distributed about their respective μ_i with common variance, σ^2. Failure to comply with this assumption results in an "averaged" estimate of σ^2 and invalidates estimates of uncertainty and comparisons of β and α among data sets.

Since we often wish to set limits of uncertainty and to compare sets of β and α to determine the feasibility of pooling data (for which purpose σ^2 must be uniform), it is desirable that the uniformity of σ^2 be ensured, if necessary by transformation. The procedures for checking the uniformity of variance do not lend themselves to an approach with pass-or-fail tests of significance and judgment is an important factor (Draper and Smith 1966). A sequence of steps which the author has found useful is as follows:

A—the variance of Y is calculated for each X class and plotted over the X-class centers on arithmetic paper

1) If the variance shows a definite trend, in plant material commonly increasing with increasing X, proceed as in step B. See Draper and Smith (1966) or other standard references for equivocal cases.
2) If the plot of variance of Y over X-class yields a horizontal band (often with wide, but random, scatter), this indicates that the variance of Y_i is independent of X_i and it is reasonable to assume a model of the form

[5] $$Y_i = \beta X_i^{\alpha} + \varepsilon_i$$

where ε_i is a random error. The appropriate sum of squares to minimize is that given by equation [2]. There are several iterative least-squares methods available for such a solution, for example see Hull (1967) and Zar (1968).

B—If the variance of Y is not uniform across the domain of X, this indicates that the variance of Y is not independent of X_i. In this case a possible model would be

[6] $$Y_i = (\beta X_i^{\alpha})\varepsilon_i$$

which, when transformed to logarithms yields

[7] $$\mathrm{LN}\,(Y_i) = \mathrm{LN}(\beta) + \alpha\,\mathrm{LN}(X_i) + \mathrm{LN}(\varepsilon_i)$$

To test this possibility, the variance of $\mathrm{LN}(Y)$ is calculated and plotted over X-class as before.

1) If the variance shows a trend away from the horizontal, proceed as in step C.
2) If the plot of variance of LN(Y) over X-class is essentially horizontal with random deviations, then the model is indeed of the form of [7] and the appropriate sum of squares to minimize is that given by equation [4]. The solution procedure is to transform each Y and X variate to its logarithm and compile as in linear regression.

C—If both the arithmetic and logarithmic variances fail to show uniformity, it will be necessary to weight each Y_i observation, commonly by the inverse of the variance of each $\overline{Y}_i$ (Draper and Smith 1966) and then solve for the regression constants using the weighted logarithms.

For determining the correct model, I have found a FORTRAN program which calculates the variance of Y and of LN(Y) by X-classes and plots the variance of Y over X and variance of LN(Y) over X useful (Baskerville 1970). Plots are also obtained of Y over X, LN(Y) over LN(X) and of the deviations from regression in both arithmetic and logarithmic units. Such a display makes it relatively easy to evaluate the validity of the assumptions discussed above and to determine the appropriate model.

The above is a minimal, but often sufficient, procedure for ensuring the correct choice of model. The reader is referred to standard references for definitive treatments.

Interpretation

If the model is of the form [5] and if the solution for the parameters is by iterative techniques that minimize the sum of squares in equation [2], then using the proper degrees of freedom: a) The sample variance (*i.e.*, the variance yielded by the $(\varepsilon_i)^2$) is an unbiased estimate of σ^2 and is the appropriate value to use in the comparison of regression parameters; b) the estimate $\hat{Y}_i$ is an unbiased estimate of μ at X_i; and c) The limits of uncertainty about Y can be calculated in the usual way using $\hat{\sigma}^2$.

If the model is of the form of [7] and if the solution is by linear regression after transformation to logarithms thus minimizing the sum of squares given by equation [4], then: a) The sample variance (*i.e.*, in terms of (LN$(\varepsilon_i))^2$) is an unbiased estimate of σ^2 at LN(X_i); b) The estimate $\widehat{\text{LN}(Y_i)}$ is an unbiased estimate of μ at LN(X_i); and c) the limits of uncertainty about LN(Y) are calculated in the usual way using $\hat{\sigma}^2$.

Conversion of Logarithmic Estimates to Arithmetic Units

When the logarithmic transformation is used, it is usually desirable, indeed necessary, to be able to express estimated values of Y in arithmetic (*i.e.*, untransformed) units. However, the conversion of the unbiased logarithmic estimates of the mean and variance back to arithmetic units is not direct. This results from the fact that if the distribution of LN(Y) at a given X is normal, the distribution of Y cannot be normal but will certainly be skewed. In fact, if the distribution is normal in logarithms, the solution of [3] for a given X_i and the determining of the antilogarithm of LN(Y_i) yields the median of the skewed arithmetic distribution rather than the mean (Brownlee 1967; Finney 1941)! The corrections for skewness are given by Brownlee (on p. 62) as follows:

$$\text{if } \hat{\mu} = \widehat{\text{LN}(Y)} = \hat{\beta} + \hat{\alpha}\text{LN}(X)$$

and $\hat{\sigma}^2$ = sample variance of the logarithmic equation;
Then

$$[8] \qquad \hat{Y} \doteq e^{(\hat{\mu} + \hat{\sigma}^2/2)}$$

$$[9] \qquad \hat{\sigma}_A^2 \doteq e^{(2\hat{\sigma}^2 + 2\hat{\mu})} - e^{(\hat{\sigma}^2 + 2\hat{\mu})},$$

where $\hat{Y}$ is the estimated mean in arithmetic units of the (skewed) Y distribution at X and $\hat{\sigma}_A^2$ is the estimated variance (for the skewed Y distribution) in arithmetic units. Uncertainty limits can be retransformed from logarithms in a manner similar to $\hat{Y}$ and these will be asymmetric about the regression line but the asymmetry will be in a direction appropriate to account for the skewness.

An Example

As an example of the difference between retransformation to the median and mean, Table 1 shows the estimated weight of foliage on balsam fir trees (*Abies balsamea* (L.)

Mill.) from 1 to 14 in. (2.54 to 35.56 cm) in diameter at breast height: a) determined from the median $\hat{Y}_i$, that is the antilog of $[\beta + \alpha \text{ LN}(X_i)]$; b) determined from the mean $\hat{Y}_i$ as calculated by [8]; and c) determined from a weighted mean $\hat{Y}_i$. The last is an adjustment for the fact that the slope of the allometric curve is continuously increasing over the domain of X and therefore the $\hat{Y}$ at the X-class mid-point is always a slight underestimate of the mean for all the possible $\hat{Y}$'s for the class. The regressions on which this table is based contained 102 observations and by virtue of the scheme outlined above required transformation to logarithms for compilation. Further, examination of the plottings of the Y variable and deviations from the model over X showed that the distribution of Y at a given X was normal in logarithm form and skewed in arithmetic form.

TABLE 1. Comparison of three solutions of the allometric equation $\text{LN}(Y) = \beta + \alpha[\text{LN}(X)]$

DBH class (inches)	Weight (kg) of foliage determined from—		
	Median	Mean	Weighted mean
1		0.03	
2		0.26	
3		0.97	
4		2.45	
5		5.04	
6		9.09	
7		14.95	
8		23.02	
9		33.69	
10		47.35	
11		64.43	
12		85.35	
13		110.54	
14		140.46	

The differences in Table 1 are seen to be appreciable, particularly between the median and mean estimates. For some cases it may be reasonable to use the median value, but in estimating biomass (and the chemical inventories which depend upon it) it is clear that the centroid of the class is the desired value and this is given by the mean. The literature contains many estimates of plant biomass based on logarithmic relationships, but I am not aware of any case (including my own data) in which the median was not inadvertently used in place of the mean in the expansion to biomass per unit area. The problem was recognized by Madgwick (1970) although he did not pursue the matter.

It is evident that the error introduced by the use of the median Y where the mean Y is appropriate increases with the average size in the X dimension. The effect could be devastating when stands of different structure are being compared since a differential error is introduced. For example, when the stand table for a young stand was expanded by means of appropriate logarithmic equations to biomass per hectare determined by each of the above three estimating procedures, it was apparent that retransformation of regression estimates to median values as opposed to mean values introduced an error of the order of 10–20% of the total biomass for a tree component. This error will always be in the nature of an underestimate.

I have examined some 40 regressions for various components of four broad-leaved and two coniferous tree species each having some 70 to 100 observations. In every case, the variance was highly unstable in arithmetic units and the logarithmic transformation rectified this problem. In every case, the plotted data $(Y/X, \text{LN}(Y)/\text{LN}(X), (Y-\hat{Y})/X)$ indicated the distribution of Y to be normal in logarithms and markedly skewed in arithmetic units. Thus, in every case it was necessary to apply equation [8] in the retransformation. Casual inspection of several similar data sets in the literature indicates that while the use of a logarithmic transformation was valid, the retransformation was to the median when it was intended to have been to the mean.

Conclusions

Proper use of regression techniques often makes it necessary to transform data to their logarithms since failure to do so invalidates limits of uncertainty and the comparison of regression constants (for example to examine the possibility of pooling data for stands or for a group of species). However, the transformation from the logarithmic form back to arithmetic units by simply determining the antilogarithm has, by failing to account for the skewness of the distribution in arithmetic

units, yielded the median rather than the mean value of Y_i for a given X_i. This has resulted in a systematic underestimation of biomass whenever the logarithmic transformation has been used. Simply to avoid the logarithm is not the solution since this will, in most cases, retain the unstable variance and associated doubts about limits of uncertainty.

Inasmuch as there is ready access to computers at virtually all centers of investigation, it is not a great burden to choose the regression model and retransformation appropriate to the data at hand. The bias is sufficiently large that tests for its existence and, where necessary, its correction are well worth the effort.

The author acknowledges the assistance and advice of Dr. J. J. Beauchamp of the Oak Ridge National Laboratory, Biometrics Group, in the preparation of the material for this paper.

BASKERVILLE, G. L. 1965. Dry matter production in immature balsam fir stands. Forest Sci. Mono. 9. 42 pp.

——— 1970. Testing the uniformity of variance in arithmetic and logarithmic units of a *Y*-variable for classes of an *X*-variable. Oak Ridge Nat. Lab. Publ. ORNL-IBP-70-1. 38 pp.

BROWNLEE, K. A. 1967. Statistical theory and methodology in science and engineering. Second Edition. John Wiley and Sons, N.Y. 400 pp.

DRAPER, N. R., and SMITH, H. 1966. Applied regression analysis. John Wiley and Sons, N.Y. 405 pp.

FINNEY, D. I. 1941. On the distribution of a variate whose logarithm is normally distributed. J. Royal Stat. Sci., Series B **7**, 155–161.

HULL, Norma C. 1967. STATPAK. U.S. At. Energy Comm. Doc. 1863, (see NONLIN).

KIRA, T., and SHIDEI, T. 1967. Primary production and turnover of organic matter in different forests ecosystems of the western Pacific. Jap. J. Ecol. **17**, 70–87.

MADGWICK, H. A. I. 1970. Biomass and productivity models of forest canopies. *In* Analysis of temperate forest ecosystems. *Edited by* D. E. Reichle, Springer Verlag, Heidelberg and New York, pp. 47–54.

WHITAKER, R. H., and WOODWELL, G. M. 1968. Estimating primary productivity in terrestrial ecosystems. Amer. Zool. **8**, 19–30.

ZAR, J. H. 1968. Calculation and miscalculation of the allometric equation as a model in biological data. BioSci. **18**, 1118–1120.