A method for calculating the variance and prediction intervals for biomass estimates obtained from allometric equations

A KIRTON B SCHOLES S ARCHIBALD
CSIR Ecosystem Processes and Dynamics, Natural Resources and the Environment P.O. BOX 395, Pretoria, 0001, South Africa
Email: anickless@csir.co.za - www.csir.co.za

INTRODUCTION

Because of the way that living things grow and develop their physical characteristics often follow simple rules. For example, the mass of a tree is strongly related to its trunk diameter; or the metabolic rate of a mammal to its weight. These relationships (called allometric relationships) can be used to help estimate things that are quite difficult to measure.

Whenever one is estimating a value it is often important to know how much error is involved. In the case of allometric equations, this information is often not published, forcing the users of these equations to use alternative, less rigorous, methods of obtaining error estimates.

This poster explains how prediction intervals (confidence intervals for predicted values) for allometric estimates can be obtained using an example of estimating tree biomass from stem diameter. It explains how to deal with relationships which are in the power function form - a common form for allometric relationships - and identifies the information that needs to be provided with the allometric equation if it is to be used with confidence.

Correct estimation of tree biomass with known error is very important when trees are being planted for carbon credits. This method can be used in many other scientific disciplines as well.

METHODOLOGY

The general form of the simple linear regression equation is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where i is the subject index, yi is the response variable, x_i is the predictor variable, β_0 and β_1 are the regression coefficients, and ε_i is the error.

We make the assumptions that the error is normally distributed with zero mean and constant variance. The constant variance assumption implies that across the range of x values, the variability in the error does not change (i.e. no heteroscedasticity).

Often the power function in allometry is used: $y = ax^b\varepsilon$

This can be converted to: $ln(y_i) = \beta_0 + \beta_1 ln(x_i) + \varepsilon_i$

The above assumptions now apply to the regression relationship with the logged variables. Therefore $ln(y_i)$ is assumed to be normally distributed with mean $\mu=\beta_0+\beta_1 ln(x_i)$ and variance $\sigma 2^*$.
From regression theory it is known that the expected value (E) and variance (Var) of $ln(y_i)$ is given by:

$$\mathrm{E}(\ln(\hat{y}_i)) = \hat{\mu}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* \ln(x_i)$$

$$\mathrm{Var}(\ln(\hat{y}_i)) = \hat{\sigma}_i^{2*} = \mathrm{MSE}(1 + \ln(x_i)(X'X)^{-1}\ln(x_i))$$

where $ln(y_i)$ is the predicted value from a new x_i value, MSE is the mean square error obtained from the original regression analysis, and X is the design matrix of the original regression. The term $x'X$ can be simplified to:

$$\begin{pmatrix} n & \sum lnx_j \\ \sum lnx_j & \sum (lnx)^2 \end{pmatrix}$$

where n is the sample size of the original regression, and the summations are applied to the predictor vector used to derive the original regression relationship, indicated by the subscript j.

Figure 1: Tree parameters, such as stem diameter, are measured in the field in order to obtain biomass estimates for these trees.

Next, obtain the predicted value of y_i (in the correct units) from $\hat{\mu}_i^* = \ln(\hat{y}_i)$ Since the log of y_i is normally distributed, by definition, y_i is lognormally distributed. Using the theory of the lognormal distribution the value for the estimate of y_i can be obtained by applying the following transformation:

$$\hat{y}_i = \exp(\ln(\hat{y}_i) + \hat{\sigma}^{2*}/2)$$

Therefore it is necessary to have an estimate of the variance for the logged prediction in order to obtain an unbiased estimate for yi. In addition the variance of the predicted value can be obtained from the following equation: $\hat{\sigma}^2 = \exp(2\hat{\mu}_i^* + 2\hat{\sigma}_i^{2*}) - \exp(2\hat{\mu}_i^* + \hat{\sigma}_i^{2*}))$ (Crow and Shimizu, 1988).

Prediction intervals are then available for these lognormal variables:

$$\text{Lower Limit} = \hat{y}_i \exp[-(z^2_{1-\alpha/2}\hat{\sigma}_i^2 + \{\hat{\sigma}_i^2/2\}^2)^{1/2}]$$

$$\text{Upper Limit} = \hat{y}_i \exp[(z^2_{1-\alpha/2}\hat{\sigma}_i^2 + \{\hat{\sigma}_i^2/2\}^2)^{1/2}]]$$

where $z_{1-\frac{a}{2}}$ is the $(1 - \frac{a}{2})$ quantile of the standard normal distribution (Zou, Huo and Taleban, 2009).

To demonstrate this method, the stem diameters collected during a field campaign characterising the vegetation structure at the Skukuza flux site, located in the Kruger National Park, South Africa, were used to obtain the estimated biomass at the site, along with the variance estimates and 95% prediction intervals. In this example both woody biomass and leaf biomass were estimated.

APPLICATION

To obtain biomass estimates, biomass allometric equations needed to be obtained. These relationships were fitted to data from destructive samples, provided by Scholes (1988) and Goodman (1990).

For demonstration purposes, estimates will be obtained using the equation for the tree Combretum apiculatum. **Table 1** gives the equation and a summary of the regression results. The regression coefficients were derived using R open-source statistical software (http://www.r-project.org). For leaf biomass it was found that a linear form of the relationship fitted the data better than a power equation, and that a relationship with the square of stem diameter fitted better than the unsquared diameter, which can be explained since leaf area should scale with cross-sectional area of the trunk (Scholes (1988) and Chidumayo (1990) have also reported linear equations for leaf biomass).

Table 1: Regression statistics obtained from allometric datasets. The MSE is reported with the standard regression output, and the sums of the predictor variable can easily be derived using the sum function in R or any spreadsheet application. The data used to fit these equations are from Scholes (1988).

Woody Biomass: $\ln(\hat{y}_{Wi}) = \hat{\beta}^*_{W0} + \hat{\beta}^*_{W1}\ln(x_i)$

Species	$\hat{\beta}^*_{W0}$	$\hat{\beta}^*_{W1}$	MSE	N	$\sum(\ln x_j)$	$\sum(\ln x_j)^2$	R^2	Range of diameter (cm)
Combretum apiculatum	-3.27	2.80	4.24 $\times 10^{-2}$	30	61.37	133.39	0.98	2.1 – 18.2

Leaf Biomass: $\hat{y}_{Li} = \hat{\beta}_{L0} + \hat{\beta}_{L1}x_i^2$

Species	$\hat{\beta}^*_{L0}$	$\hat{\beta}^*_{L1}$	MSE	N	$\sum(x_j^2)$	$\sum(x_j^2)^2$	R^2	Range of diameter (cm)
Combretum apiculatum	-0.156	0.012	3.80 $\times 10^{-3}$	28	725.00	26583.00	0.92	2.8 – 10.2

To obtain the variance estimates for the leaf biomass, a similar approach as described earlier for the logged regression equation can be implemented, but it is now not necessary to make the adjustments for the lognormal distribution. The estimate for the variance of when the relationship is in the form of a simple (i.e. just one predictor variable) linear regression, with stem diameter squared as the predictor variable, is:

$$\mathrm{Var}(\hat{y}_{Li}) = \hat{\sigma}^2_{Li} = MSE(1 + x_i^2(X'X)^{-1}x_i^2)$$

where $x'X$ can now be simplified to $\begin{pmatrix} n & \sum x_j^2 \\ \sum x_j^2 & \sum (x_j^2)^2 \end{pmatrix}$ and the 95% prediction interval will then be $\hat{y}_{Li} \pm 1.96 \times \sqrt{\hat{\sigma}^2_{Li}}$. The results are displayed in **Figure 2**.

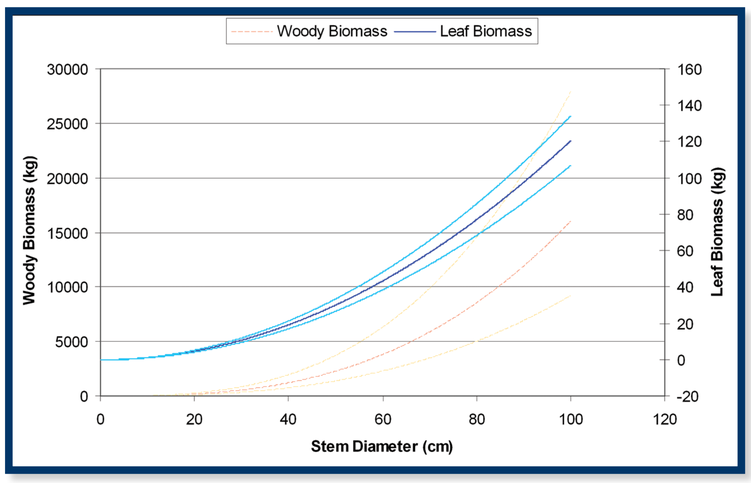

Figure 2: Plot of the estimated woody (red dashed line, left axis) and leaf (blue solid line, right axis) biomass for a Combretum apiculatum tree at different stem diameters. The prediction intervals are plotted as well. These plots show that as the diameter values moves further from the original range of diameters to which the regression equation was fitted, the confidence interval becomes wider. The plot of the woody biomass equation also indicates the asymmetrical nature of the prediction interval due to the lognormal distribution assumed to underlie the data.

The method explained in this poster provides a straightforward means of obtaining allometric estimates and their variances, along with prediction intervals. This method makes use of the regression theory already universally used to obtain allometric relationships, and goes further into the theory to extract the variance of predicted values.

If this method is to be widely implemented, publications on allometric relationships based on standard regression theory must report, in addition to the regression coefficients, the sum of the squared and unsquared predictor variable, the mean square error, and the sample size.

Therefore no additional assumptions are made. If this method is to be widely implemented, then publications on allometric relationships based on standard regression theory must report, in addition to the regression coefficients, the sum of the squared and unsquared predictor variable, the mean square error, and the sample size.

LITERATURE CITED

1. Chidumayo, E.N. 1990. Above-ground woody biomass structure and productivity in a Zambezian woodland. Forest Ecology and Management, 36: 33-46.
2. Crow, E.L. and Shimizu, K. 1988. Lognormal Distributions: Theory and Applications. Dekker: New York.
3. Goodman, P.S. 1990. Soil, vegetation and large herbivore relations in Mkuzi Game Reserve, Natal. PhD Thesis, University of the Witwatersrand.
4. Scholes, R.J. 1988. Response of three semi-arid savannas on contrasting soils to the removal of the woody component. PhD Thesis, University of the Witwatersrand.
5. Zou, G.Y., Huo, C.Y., and Taleban, J. 2009. Simple confidence intervals for lognormal means and their differences with environmental applications. Environmetrics, 20:172-180.