Science real and relevant conference 2010
Proceedings homepage://www.conference.csir.co.za/

Calculating the variance and prediction intervals for estimates obtained from allometric relationships

Authors: Alecia Nickless[1], Robert J. Scholes[1] and Sally Archibald[1]

[1]CSIR Natural Resources and the Environment
Corresponding Author: Alecia Nickless

Email: ANickless@csir.co.za

Reference: NE08-PO-F

Abstract

Often researchers are interested in obtaining estimates of variables which are quite difficult or expensive to measure. To obtain these estimates, relationships between those variables of interest and more easily measured variables are used. These relationships are referred to as allometric equations.

In science it is important to quantify the error associated with an estimate in order to determine the reliability of the estimate. Therefore, prediction intervals or standard errors are usually quoted with estimated values. In the case of allometric equations, information about the original fitting of the allometric relationship is needed in order to put a prediction interval around an estimated value. However, often all the information required to calculate this prediction interval is not provided with published allometric equations, forcing the users of these equations to use alternative, less rigorous methods of obtaining error estimates.

This paper will explain the method behind obtaining prediction intervals for allometric estimates, and what information is required from the original fitting of the allometric relationships. This information seeks to provide researchers with the necessary parameters which should be published with allometric relationships. In addition, a method is explained for how to deal with relationships which are in the power function form – a common form for allometric relationships.

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 stem diameter, and the metabolic rate of a mammal is related to its weight. The use of allometric equations to calculate the size of one body measurement based on another has been in existence since 1897 when Eugène Dubois first published his paper containing a quantitative formula relating the weight of the brain to the weight of a person's body (Gayon, 2000). Since this time, equations have been used to quantify the relationship between body measurements for a multitude of organisms and for a diverse set of objectives.

For estimates to be scientifically defensible, it is important that they are associated with prediction intervals (confidence intervals for predicted values – estimates based on new information). If standard regression theory is used to obtain allometric equations, formulae are available for these prediction intervals. Unfortunately, authors of allometric equations often do not include all the relevant information required to obtain these intervals.

The purpose of this paper is to provide some of the underlying theory behind fitting these equations and to explain how this theory can be extended to obtain prediction intervals for the estimated values.

Methodology

The general form of the simple linear regression equation is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ \ldots \ (1)$$

where *i* is the subject index, y_i 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. It is assumed 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.

In the case of many allometric relationships, the simple linear regression equation is often modified by assuming that the logarithm of the response variable can be explained by the linear equation:

$$\ln(y_i) = \beta_0^* + \beta_1^* \ln(x_i) + \varepsilon_i^* \ \ldots (2)$$

where the regression parameters superscripted by asterisks denote the log regression parameters. The assumptions previously mentioned 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 σ^{2*}. The fitting techniques which apply to simple linear regression can be used to obtain the regression parameters once the variables have been log transformed. This relationship is often expressed in the power form by taking exponents on both sides of the equation:

$$y_i = \exp(\beta_0^*) x_i^{\beta_1^*} \exp(\varepsilon_i^*) \ \ldots (3).$$

Often the error term is not shown. For example, this is the typical form of the equation when applied to plant woody biomass (Zianis, 2008). Another typical application of this form of the equation is the relationship between weight and height in humans (García-Berthou, 2001).

The ordinary least squares (OLS) fitting method assumes that the dependent variable is perfectly known, and only the independent variable is prone to measurement error. In most cases, both the parameter which can be easily measured and the parameter of interest are prone to some error when measured. The major axis and reduced major axis (RMA) fitting methods were developed for this situation (Niklas, 2004). McArdle (1988) studied the accuracy of regression estimates obtained under the three different fitting techniques and found that if the error in the independent variable was less than a third of the error in the dependent variable, the OLS method was more accurate. The RMA method was shown to be more accurate than the major axis method.

Since the measurement error of the parameter of interest is likely to be much larger than that of the easily measurable parameter, under most circumstances using OLS to fit regression models will be justified. Niklas (2004) also states that if the coefficient of determination (R^2) is more than 0.95, there will be very little difference between the slope estimates of different fitting methods. When dealing with strong relationships between body parameters, such as the relationship between a tree's biomass and its stem diameter, generally, R^2-values tend to be above 0.90 (e.g. Williams *et al*., 2003; Laclau *et al*., 2008), but there are always exceptions. For these reasons, and since OLS tends to be the most utilised method (Zianis, 2008), the OLS method was used for fitting regression relationships in this paper.

If it can be assumed that the natural logarithm of the response is normally distributed, it implies that the response itself must be log-normally distributed with mean

$$\mu = \exp(\mu^* + \sigma^{2*}/2) \ \ldots (4)$$

and variance

$$\sigma^2 = \exp(2\mu^* + 2\sigma^{2*}) - \exp(2\mu^* + \sigma^{2*}) \ \ldots (5) \text{ (Crow and Shimizu, 1988).}$$

Note that the estimate for μ is a function of both μ^* and σ^{2*}. This is why it is not possible to simply take the exponent of the estimates of the logged response from the linear regression to obtain the estimates in the required scale, as this would result in bias (Stow *et al*., 2006).

Statisticians and statistical software packages compute systems of linear equations (such as those that underlie linear regression analysis) using matrix algebra. A brief explanation of this convenient and efficient notation is included here for the non-specialists. The matrix of observations of the independent variable, referred to as the predictor or design matrix, is denoted *X*, and its transpose as X'. In the case of a simple linear regression, the first column of *X* is a column of ones (for the intercept), and the second column is the vector of observations of the independent variable (in the same scale as represented in the linear regression equation). A key calculation in the text that follows is $(X'X)^{-1}$, which is the matrix inversion of the matrix multiplication between *X* and its transpose.

From regression theory it is known that the expected value (or mean value) (E) and variance (Var) of $\ln(\hat{y}_i)$ is given by

$$\mathrm{E}(\ln(\hat{y}_i)) = \hat{\mu}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* \ln(x_i) \ldots (6) \text{ and}$$

$$\mathrm{Var}(\ln(\hat{y}_i)) = \hat{\sigma}_i^{2*} = \mathrm{MSE}(1 + \ln(x_i)(X'X)^{-1} \ln(x_i)) \ldots (7)$$

where $\ln(\hat{y}_i)$ 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 MSE can also be referred to as the residual sum of squares or the sum of squares of the error. In the case of a simple linear regression, the calculation of the variance of a predicted value can be simplified into easily derivable terms from the original regression data. The term $X'X$ can be simplified to $\begin{pmatrix} n & \sum \ln x_j \\ \sum \ln x_j & \sum (\ln x_j)^2 \end{pmatrix}$, 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*. Therefore if the MSE and the summary terms of $X'X$ are available, this closed form expression for the variance of a new predicted value can be used. A thorough explanation of linear regression theory can be found in Seber and Lee (2003).

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

$$\hat{y}_i = \exp(\ln(\hat{y}_i) + \hat{\sigma}^{2*}/2) \ldots (8).$$

Therefore it is necessary to have an estimate of the variance for the logged prediction in order to obtain an unbiased estimate for y_i. In addition the variance of the predicted value can be obtained from the following equation:

$$\hat{\sigma}^2 = \exp(2\hat{\mu}_i^* + 2\hat{\sigma}_i^{2*}) - \exp(2\hat{\mu}_i^* + \hat{\sigma}_i^{2*}) \ \ldots (9) \text{ (Crow and Shimizu, 1988).}$$

The approximate 100(1 – *α*)% prediction limits for a lognormal variable can be used:

$$\text{Lower Limit} = \hat{y}_i \exp[-(z_{1-\alpha/2}^2 \hat{\sigma}_i^2 + \{\hat{\sigma}_i^2/2\}^2)^{1/2}] \ \ldots (10)$$

$$\text{Upper Limit} = \hat{y}_i \exp[(z_{1-\alpha/2}^2 \hat{\sigma}_i^2 + \{\hat{\sigma}_i^2/2\}^2)^{1/2}] \ \ldots (11)$$

where $z_{1-\alpha/2}$ is the 1 - $\alpha/2$ quantile of the standard normal distribution. These equations were derived based on the method described in Zou, Huo and Taleban (2009) where confidence intervals were derived for the mean of a lognormal variable.

Application

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.

The area in which the tree census was carried out measured 200 m by 200 m. It was accurately demarcated into a 50 m by 50 m sampling grid. All the stems taller than 1.0 m within the entire area were measured to obtain diameter just above the basal swelling.

Before biomass estimates could be obtained, the available allometric datasets specific to the tree species at the site were collected and the appropriate regression parameters were calculated. To collect the allometric datasets, plant parameters, including stem diameter, were measured on the selected trees. Once these measurements were taken, the tree or branch of the tree was cut at the base. The biomass was then separated into woody and leaf biomass and then oven-dried for at least 48 hours. The sources for all the allometric datasets are from Scholes (1988) and Goodman (1990). Both of these authors gave access to their original data, and therefore the regression coefficients and required regression statistics could be derived. Originally only the regression coefficients, R^2 value, sample size and range of stem diameters were reported.

Table 1: Regression statistics obtained from allometric datasets. 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	4.24×10^{-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	3.80×10^{-3}	28	725.00	26583.00	0.92	2.8 – 10.2

For demonstration purposes, estimates will be obtained using the equation for *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). The statistics supplied in this table are sufficient to calculate the variance of biomass estimates obtained from the given regression equations. The units of the MSE are the squared units of the response, therefore in this example they depend on whether biomass was logged or not. The MSE is reported with the standard regression output, and the sums of predictor variable can easily be obtained using the sum function in R or in most spreadsheet applications.

The equation fitted to woody biomass was $\ln(y_{Wi}) = \beta^*_{W0} + \beta^*_{W1}\ln(x_i) + \varepsilon^*_{Wi}$, where y_{Wi} is the dried woody biomass in kg, x_i is the stem diameter in cm, β^*_{W0} and β^*_{W1} are the regression coefficients for the logged woody biomass, and ε^*_{Wi} is the error in the estimation of logged woody biomass. The equation fitted to leaf biomass was $y_{Li} = \beta_{L0} + \beta_{L1}x_i^2 + \varepsilon_{Li}$, where y_{Li} is the leaf biomass in kg, β_{L0} and β_{L1} are the regression coefficients for leaf biomass, and ε_{Li} is the estimation error. For leaf biomass it was found that a linear form of the relationship fit the data better than a power equation, and that a relationship with the

square of stem diameter fit better than the unsquared diameter, which can be explained since tree volume should scale with cross-sectional area of the trunk (Scholes (1988) and Chidumayo (1990) have also reported linear equations for leaf biomass).

To obtain the variance estimates for the leaf biomass, a similar approach as described above 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 $\hat{y}_{Li}$ when the relationship is in the form of a simple linear regression, with stem diameter squared as the predictor variable, is

$$\mathrm{Var}(\hat{y}_{Li}) = \hat{\sigma}^2_{Li} = MSE(1 + x_i^2 (X'X)^{-1} x_i^2) \quad \ldots (12)$$

where $X'X$ can now be simplified to $\begin{pmatrix} n & \sum x_j^2 \\ \sum x_j^2 & \sum (x_j^2)^2 \end{pmatrix}$. For this example, the estimate for $\hat{y}_{Li}$ will be in kilograms, and therefore the variance term will be in squared kilograms. The 95% prediction interval will then be $\hat{y}_{Li} \pm 1.96 \times \sqrt{\hat{\sigma}^2_{Li}}$.

The derived regression equations, as well as the additional regression statistics, were used to obtain biomass estimates and their variances based on the stem diameter measurements taken at the Skukuza flux site. As an example, a stem with diameter measured as 8.27 cm would have an estimated logged biomass of 2.65 with a variance of 0.044. Converting these logged estimates into kilograms, using the equation for unbiased estimates, gives an estimate 14.5 kg and a variance of 9.47 kg^2. Applying the formulae for the prediction intervals when the logged variable is modelled, the 95% prediction interval is then (9.63, 21.92) kg. The same stem diameter can be used to estimate the leaf biomass for that stem. Since the biomass was modelled directly (rather than log transformed), the estimate for leaf biomass is 0.67 kg with a variance of 0.0048 kg^2, and no further transformations are required. The 95% prediction interval for this estimate is (0.53; 0.80) kg.

A stem diameter of 28.32 cm results in an estimated woody biomass of 460.30 kg with a variance of 11524.68 kg^2. The corresponding prediction interval is then (292.9, 723.23) kg. Estimating leaf biomass gives an estimate of 9.52 kg with a variance of 0.29 kg and a 95% prediction interval of (8.45, 10.58) kg. Compared to the previous stem diameter, these estimates are much larger, and also have much wider prediction intervals. This diameter value lies outside of the original range of stem diameters used to derive the allometric equations (Table 1).

Large diameter values will give wide prediction intervals because, as the diameter of the trees moves further outside of the range of the original regression, the prediction interval becomes wider since the uncertainty in the estimate will be greater. In the case of a simple linear regression, the standard deviation of a predicted value can be written as:

$$\hat{\sigma}_i = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}\right)} \quad \ldots (13) \text{ (Chatterjee and Hadi, 2006).}$$

This formulation shows that the standard deviation of the predicted value from x_i, $\hat{\sigma}_i$, will get larger the further the new x value moves from the mean of the x's from the original regression, $\bar{x}$. The standard deviation is the component of the prediction interval for a predicted value which determines how wide the interval will be, and therefore the prediction interval becomes wider for larger x values. This has been illustrated by plotting diameter values against the corresponding estimates for biomass, including the prediction intervals (Fig. 1). This result implies that allometric relationships are best suited for obtaining estimates from predictor variables which are in the same range as those used to fit the original regression equation.

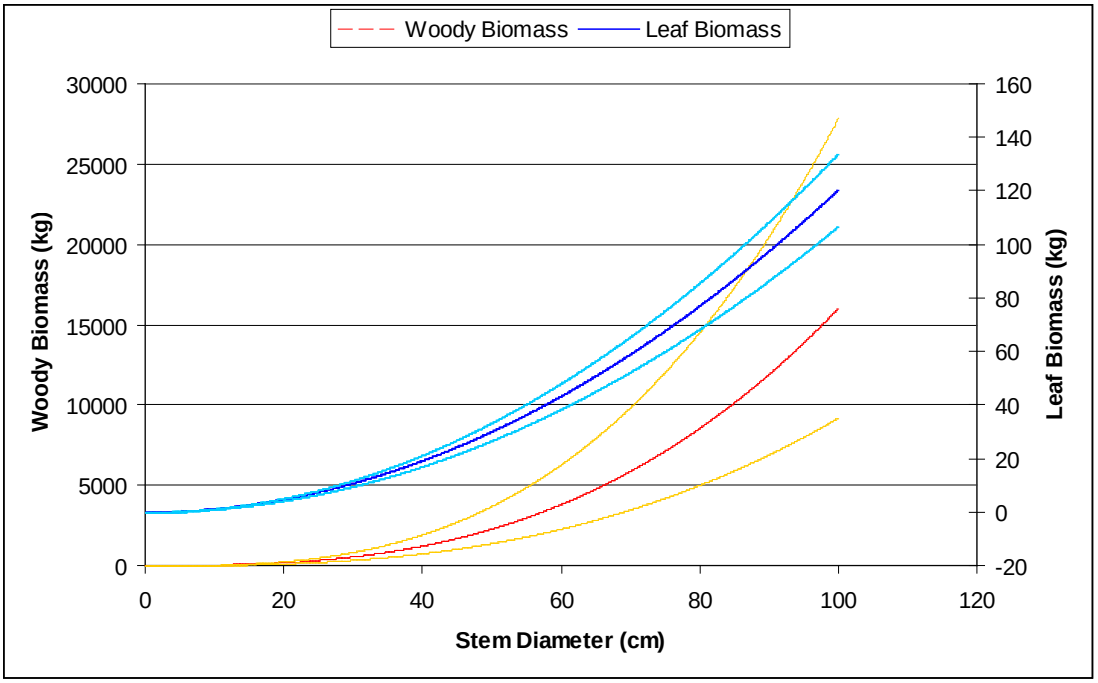

Fig.1: The best estimate and upper and lower 95% confidence limits for wood biomass (broken line, left axis) and leaf biomass (solid line, right axis) in *Combretum apiculatum*, as a function of stem diameter.

Conclusions

The method explained in this paper 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. Therefore no additional assumptions are made and no additional field work is required. For the widely used power-law formulation, lognormal distribution theory is used to obtain appropriate estimates and asymmetric prediction intervals for the estimates in the required units.

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. These statistics are readily available from current statistical analysis software. Together these parameters fully define the relationship. If other methods of fitting allometric relationships are used, such as non-linear regression methods, enough information must be reported to allow users of these relationships to construct prediction intervals for estimates.

This methodology can be extended to any application where a linear regression equation obtained from historic data is used to predict on new data, including those applications where a logged dependent variable is used.

References

Chatterjee, S. and Hadi, A.S., 2006. *Regression Analysis by Example*. John Wiley & Sons Inc.: New Jersey, 375 pages.

Chidumayo, E.N., 1990. Above-ground woody biomass structure and productivity in a Zambezian woodland, *Forest Ecology and Management*, vol. 36, pp. 33-46.

Crow, E.L. and Shimizu, K., 1988. *Lognormal Distributions: Theory and Applications*. Dekker: New York, 387 pages.

García-Berthou, E., 2001. On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance, *Journal of Animal Ecology*, vol. 10, pp. 708-711.

Gayon, 2000. History of the concept of allometry, *American Zoologist*, vol. 40, pp. 748-758.

Goodman, P.S., 1990. Soil, vegetation and large herbivore relations in Mkuzi Game Reserve, Natal, *PhD Thesis*, University of the Witwatersrand.

Laclau, J.-P., Bouillet, J.-P., Gonçalves, J.L.M., Silva, E.L., Jourdan, C., Cunha, M.C.S., Moreira, M.R., Saint-André, L., Maquère, V., Nouvellon, Y. and Ranger, J., 2008. Mixed-species plantations of *Acacia mangium* and *Eucalyptus grandis* in Brazil 1. Growth dynamics and aboveground net primary production, *Forest Ecology and Management*, vol. 255, pp. 3905-3917.

McArdle, B.H., 1988. The structural relationship: regression in biology, *Canadian Journal of Zoology*, vol. 66, 2329-2339.

Niklas, K.J., 2004. Plant allometry: is there a grand unifying theory?, *Biological Reviews*, vol. 79, pp. 871-889.

Seber, G.A.F. and Lee, A.J. 2003. *Linear Regression Analysis.* John Wiley & Sons Inc.: New Jersey, 557 pages.

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.

Stow, C.A., Reckhow, K.H., and Qian, S.S., 2006. A Bayesian approach to retransformation bias in transformed regression, *Ecology*, vol. 87, pp. 1472-1477.

Williams, C.J., LePage, B.A., Vann, D.R., Tange, T., Ikeda, H., Ando, M., Kusakabe, T., Tsuzuki, H. and Sweda, T., 2003. Structure, allometry, and biomass of plantation *Metasequoia glyptostroboides* in Japan, *Forest Ecology and Management*, vol. 180, pp. 287-301.

Zianis, D., 2008. Predicting mean aboveground forest biomass and its associated variance, *Forest Ecology and Management*, vol. 256, pp. 1400-1407.

Zou, G.Y., Huo, C.Y., and Taleban, J., 2009. Simple confidence intervals for lognormal means and their differences with environmental applications, *Environmetrics*, vol. 20, pp. 172-180.