ANNALS OF FOREST RESEARCH
www.e-afr.org

Correction for bias of models with lognormal distributed variables in absence of original data

Bogdan Strimbu

Strimbu B., 2012. Correction for bias of models with lognormal distributed variables in absence of original data. Ann. For. Res. 55(2): 265-279, 2012.

Abstract. The logarithmic transformation of the dependent variables for models developed using regression analysis induces bias that should be corrected, regardless its magnitude. The simplest correction for bias was proposed by Sprugel (1983), which basically multiplies the back-transformed estimates with the constant value of exponential of half the variance of the errors of the logarithmically transformed variable. While this correction is fast and easy to implement does not supplies estimates of the variability existing in the original data. Consequently, a procedure based on generated data was developed to provide unbiased estimates for both attribute of interest and variability existing along the model. The procedure reveals that valid estimates can be obtained if large number of values is generated (e.g., 5000 values/x). The procedures supplies accurate estimates for the attribute of interest and its variability, but encounters significant data processing difficulties for models with more than one predictor variable. Nevertheless, irrespective the number of predictor of variables and magnitude of the correction factor computed by Sprugel, the estimates determined using logarithmic transformations should be corrected for bias, to avoid cumulated errors or chaotic effects associated with nonlinear models.

Keywords log-transformation, bios, correction, volum tables.

Author. Bogdan Strimbu (strimbu@latech.edu) - Louisiana Tech University, 1201 Reese Dr Ruston LA 71272 USA.

Manuscript received June 25, 2012; revised October 03, 2012; accepted October 10, 2012; online first October 15, 2012.

Introduction

The advent of information technology characterizing the last three decades led to development of accurate and unbiased models describing different forest processes, such as growth, yield, changes in species density, or carbon storage (Adams & Titus 2009, Nitschke & Innes 2008, Pretzsch 2009, Weiskittel et al. 2011). The impact of technological advancements was noticeable not only in development of complex models but also in addressing the bias, as it made possible the estimation of the coefficients for nonlinear models in an usable amount of time (Gentleman & Ihaka 2012, Gould 2012, Nie & Hull 2012, SAS Institute

2010). However, the nonlinear models were developed long before the computation support was widely available (Assmann 1970, Giurgiu 1979, Schumacher & Hall 1933, Spurr 1952). The lack of computational power available before 1980 lead to development of models on which the dependent variable is transformed, the case of site index equation (Clutter et al. 1983), or volume equations developed by Schumacher and Hall (1933) or Giurgiu (1979). Transformation of the dependent variable achieves two purposes: 1) fits the model better to the data, and 2) usually reduces the variability, consequently, increases the significance of statistical tests. In practice, nonlinear models are derived using linear regression on transformed variables. The linear regression analysis is recommended on transformed variables, as the transformation focuses on the linearization of the relationship between variables (Miller 1984). This approach to nonlinear models is presented extensively in most statistics textbooks (Montgomery et al. 2006, Neter et al. 1996, Rao 1973, Zar 1996). Nevertheless, when least square procedure , as developed by Cotes (Edleston 1850) and formalized by Legendre (1805), is used in estimating the parameters of interests the results are biased (Miller 1984). The presence of bias in the resulted models did not stop the need for models describing different environmental processes. Therefore, in time, a significant amount of equations were developed and implemented by different agencies, private, state or federal, in the day-to-day activities. Since these models are commonly present in routine computations of different entities acting in environmental area, there is of significant interest to develop a procedure that would maintain the equations in use but will correct their bias. This task is simple if the original data are available. Unfortunately, in many instances the original data used to develop the model cannot be used, as they are missing, protected by copyrights or privacy rights, or on hardcopy and difficult/expensive to digitize. The objective of the present research is 1) to prove the existence of bias in environmental models based on logarithmic transformation of predicted variables, 2) to provide at least one methodology of correcting the bias induced by the logarithmic transformation of the dependent variable in absence of original data, but with some information available (e.g. summary statistics or graphs showing the relationship between variables). The present research complements the work of Sprugel (1983), Beauchamp & Olson (1973), which provides bias corrections for log-normal variables but using the original data.

Methods

Bias of lognormal distributed variables developed using least square method

The least square (LS) method was used extensively in the last two centuries in almost all areas of science, and was likely first described by Cotes (Edleston 1850), which noticed that combination of observations in the estimation process leads to a decrease in errors. The observations of Cotes, confirmed by Gauss (Bjorck 1996) and proved by Legendre (Legendre 1805), led to the wide spread usage of the method on applied sciences in conjunction with regression analysis. However, LS method can lead to biased results when used in combinations with regression on transformed variables, as indicated by various authors (Finney 1941, Neter et al. 1996), and are based on the Cauchy–Bunyakovsky–Schwarz inequality (Poole 2005). The proof that models developed from linear regression of transformed variables are biased can be carried in three steps: 1) proof that the regression line contains the “average” point, namely the point with coordinates the arithmetic average of all values for each variable (i.e. axis), 2) proof that the arithmetic average is different than the geometric average, and 3) proof that LS method applied to lognormal variable leads to comparison of arithmetic mean with geometric mean. In this section a complete proof of the bias in log-normal variable is presented. The selection of logarithmic transformation was recommended by three factors: 1) logarithmic transformation is one of the most used transformations in environmental investigations (Schabenberger & Pierce 2002, Williams 1997), 2) log-normal distribution was extensively studied in the last 100 years (Aitchison & Brown 1957, Crow & Shimizu 1988), 3) being one of the distributions from the exponential family of distributions (Darmois 1935, Koopman 1936) it can be related with the LS method, as LS corresponds to the maximum likelihood criterion if residuals have a normal distribution (Neter et al. 1996).

To show that the "average" point [i.e., point with coordinates $(\bar{y},\bar{\mathbf{x}})$, where y is the dependent variable, and x is the vector of predictor variables, x_j, j from 1 to k, and k the number of predictor variables] lays on the regression line

$$y = b_0 + \sum_{j=1}^{k} b_j x_j \tag{1}$$

when LS method is used for parameter estimation, one can start from the aim of LS estimation, which is the minimization of the sum of squared residuals (i.e., a residual is the difference between the measured value and the value predicted by the regression equation), as required by the Gauss- Markov theorem (Plackett 1950):

$$\min\left(\sum_{i=1}^{n}(y_i-(b_0+\sum_{j=1}^{k}b_j x_{j,i}))^2\right) \tag{2}$$

where: y_i is the i^{th} measurement of the dependent variable y; $x_{j,i}$ is i^{th} measurement of the independent variable x_j; b_0 is the intercept; b_j is the coefficient of the independent variable x_j; n is the total number of measurement (aka observations); k is the total number of independent variables

The minimum of expression 2 is reached in a point that has all partial derivatives in respect with the coefficient b_j equal to 0, which transforms the minimization problem 2 in the set of equations 3:

$$\begin{aligned}
&\frac{\partial\left(\sum_{i=1}^{n}(y_i-(b_0+\sum_{j=1}^{k}b_j x_{j,i}))^2\right)}{\partial b_0} = \sum_{i=1}^{n}(y_i-(b_0+\sum_{j=1}^{k}b_j x_{j,i})) = 0\\
&\frac{\partial\left(\sum_{i=1}^{n}(y_i-(b_0+\sum_{j=1}^{k}b_j x_{j,i}))^2\right)}{\partial b_j} = \sum_{i=1}^{n}(y_i\sum_{j=1}^{k}x_{j,i}-b_0\sum_{j=1}^{k}x_{j,i}-\sum_{j=1}^{k}b_j x^2{}_{j,i}) = 0
\end{aligned} \tag{3}$$

The first equation from 3 can be simplified to:

$$\sum_{i=1}^{n} y_i = nb_0 + \sum_{j=1}^{k} b_j x_{j,i} \tag{4}$$

Dividing equation 4 by n it results:

$$\bar{y} = b_0 + \sum_{j=1}^{k} b_j \bar{x}_j \tag{5}$$

which proves that the "average" point is located on the regression line 1.

Several proofs can be found in the literature for the inequality between the arithmetic mean and geometric mean (Cauchy 1821, Hall & Knight 2005), some very simple, but based on fundamental observations (as the one that it will be produced in the following paragraph), while other more elaborated, such as the one produced by Polya (Steele 2004).

One of the simplest proofs of the inequality between the arithmetic mean and geometric mean is based on the observation that the two means supply different results when two values are involved in the computations. Specifically, if arithmetic and geometric mean of n

values are $\mu_{arithmetic}=\frac{\sum_{i=1}^{n}x_i}{n}$ for the former, and $\mu_{geometric}=\left(\prod_{i=1}^{n}x_i\right)^{1/n}$ for the latter, then by selecting any two values x_i, x_j, such that $x_i \neq x_j$, and replacing them in each mean with their arithmetic mean, namely x_i, x_j are substitute with $(x_i + x_j)/2$ the arithmetic mean will not change, while the geometric mean will increase. The proof of this observation is obvious for the arithmetic mean, while for geometric mean one can observe that by replacing x_i and x_j with their arithmetic mean, the geometric mean becomes

$$\mu_{geometric}=\left(x_1x_2..x_{i-1}\frac{x_i+x_j}{2}x_{i+1}..x_{j-1}\frac{x_i+x_j}{2}x_{j+1..}x_n\right)^{1/n}$$

The two geometric means differs only by the i^{th} and j^{th} term, which leads to x_ix_j for the former and to $\left(\frac{x_i+x_j}{2}\right)^2$ for the latter.

As it can be concluded that the latter geometric mean is larger or equal than the former, as

$\left(\frac{x_i-x_j}{2}\right)^2 \geq 0$, for any $x_i, x_j \in R$.

Therefore, the geometric mean will be the largest when all the terms are equal, while the arithmetic mean remains unchanged regardless the linear changes of the values used in computations. If the value of each term in the geometric mean is the arithmetic mean, then the geometric mean will be the largest among all combinations of xs from the arithmetic mean. Formally, the relationship between the geometric mean and arithmetic mean is expressed as

$$\mu_{geometric}=\left(x_1x_2..x_{i-1}x_ix_{i+1}..x_{j-1}x_jx_{j+1..}x_n\right)^{1/n}\leq\left(\frac{\sum_{i=1}^{n}x_i}{n}\frac{\sum_{i=1}^{n}x_i}{n}..\frac{\sum_{i=1}^{n}x_i}{n}\right)^{1/n}= \tag{6}$$

$$=(\mu_{arithmetic}\mu_{arithmetic}..\mu_{arithmetic})^{1/n}=\mu_{arithmetic}=\frac{\sum_{i=1}^{n}x_i}{n}$$

Finally, the third step in proving the bias of models developed from linear regression on log(y) consists in relating the estimates supplied by LS method for the transformed and untransformed y. According to equation 5 the regression line passes to the point of coordinates $(\bar{y},\bar{\mathbf{x}})$, which for the logarithmic transformed y implies that $\overline{\log(y)}$ is on the regression line. As the objective of the modeling exercise is to produce equations for the untransformed variable y, the antilog is taken from the estimates, namely $\log(y)=f(\mathbf{x})\rightarrow y=\exp(f(x))$ if the base of the logarithmic transformation is e. Considering that $\overline{\log(y)}$ is on the regression line, the untransformed y corresponding to $\overline{\log(y)}$ is:

$$\exp(\overline{\ln(y)}) = \exp\left(\frac{\sum\limits_{i=1}^{n}\ln(y_i)}{n}\right) = \exp\sum_{i=1}^{n}\ln(y_i^{1/n}) = \exp\left(\ln\prod_{i=1}^{n} y_i^{1/n}\right) = \qquad (7)$$

$$= \left(\prod_{i=1}^{n} y_i\right)^{1/n} = \mu_{geometric} \le \mu_{arithmetic}$$

Equation 7 shows that transformation of the logarithmic *y* in the original units is associated with a bias of $\delta = \mu_{arithmetic} - \mu_{geometric}$, as back-transformation line does not contains the arithmetic mean. The arguments presented until this point does not bring any element of novelty to the scientific body of knowledge, just illustrates in the same argument the complete proof of the bias resulted from the transformation of the dependent variable.

Method of correcting bias of the log-transformed dependent variable using generated data

The purpose of the logarithmic transformation of the dependent variable is to translate the nonlinear relationship between variables in a linear relationship. It is not the objective of this paper to assess the significance of the relationship between variables, as it is assumed that all assumptions associated with linear regression analysis are fulfilled, such as residuals are normally distributed and have the same variance, irrespective the predictor variable x (Fig. 1). Explicitly, it is assumed that for each x the residuals are normally distributed with mean and variance σ^2 (Montgomery et al. 2006).

In absence of original data, one does not have $\hat{y} = f(\mathrm{x})$ knowledge of each observation, posibly only summary statistics, namely mean, variance, range and number of observations. One can consider correcting the bias induced by the logarithmic transformation by replacing the original data with generated data that fulfills the same distributional assumptions as original data. Consequently,

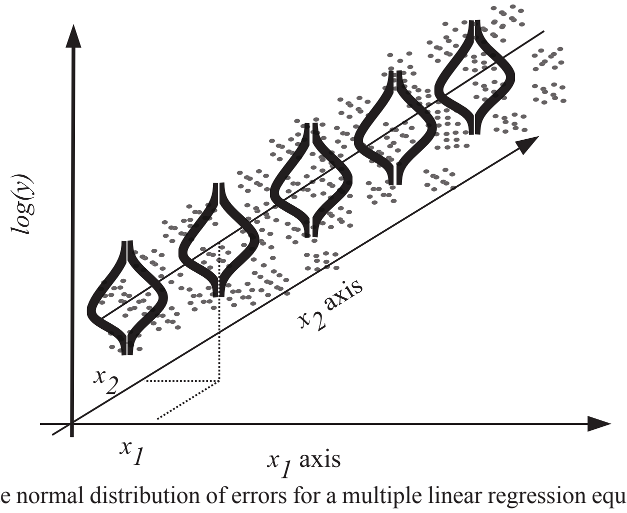

Figure 1 Bivariate normal distribution of errors for a multiple linear regression equation with two predictor variables

one could generate residuals that are normally distributed $N(0,\sigma^2)$ residuals that will be subsequently used to compute the difference between arithmetic mean and geometric mean, hence the bias. Random data generation has the advantage that is simple to implement regardless the number of independent variables, but is associated with two major issues: 1) data generated by different software are not random but pseudorandom, 2) generation of large amount of data can impose challenges to the subsequent data analysis. Therefore, one should determine how many values will be generated, such that the analysis will not be difficult or impossible, and should choose a methodology that reduces the chance of producing values that are not random. To select the number of values that should be generated for each x, one can use as an upper bound the value recommended by Liew et al (1985), and reduced, if possible, to a number that will produced insignificant changes in the resulted values, while not impeding the analysis by size. The approach used here to identify a threshold that ensures both requirements (i.e., insignificant difference between true and generated mean, and lack of computational issues) was simulation, in the sense that a set of values of size n_x, $n_x \geq 2$, were selected from a normally distributed random variables with mean μ and variance σ^2. The generation of the sample was performed using SAS 9.3 (SAS Institute 2010), four normal distributions means (i.e., 5, 10, 15, 20), and a unit variance (i.e., $\sigma^2 = 1$). The four means were chose such that the generated positive values (a requirement in the computation of the geometric mean) can be used to assess whether or not the samples size is depended on the magnitude of the values. The unit variance was chosen in conjunction with the mean (i.e., from 5 times smaller to 20 times smaller), such that the generated values will likely not be negative, considering that probability of generating a value smaller than -5 or larger, from a normal distribution with mean 5 or larger, and variance 1, is less than 10^{-4} (Feller 1968). The maximum sample size was selected to be 100, similar to Liew et al (1985), as commonly regression based models are developed from more than 100 measurements or observations. An upper limit of 100 was chosen to avoid possible computational issues, as more than 10000 values will be used in computations. In the case of multiple predictor variables the complexity of the problem is exponential in nature, which recommends less than 100 values to be used to assess the difference between arithmetic and geometric mean (McClave & Dietrich 1991, Tran 1997). An additional complication associated with multiple predictor variables is the possible existence of a significant correlation between the variables, which invalidates a factorial approach to the generated data approach. In absence of correlation between predictor variables, for each combination of *xs* the same number of values will be generated; for example if height is predicted as a function of site productivity and age the same number of values will be generated for all age – site productivity combinations, which can lead to more than 1800 combinations (i.e., 60 ages x 30 site productivities). However, when correlation between predictor variables is present (for instance the standard volume equations (Husch et al. 2002) that include diameter at breast height (dbh) and height as predictor variables), and extra conditioning should be included in the data generation process. For correlated variables, the linear equation quantifying the relationship between variables could be used to generate the number of values to be used in bias correction. However, as the residuals of the logarithmic transformed variables are multi-normal distributed the most values are located on the regression line, which precludes the usage of a factorial approach for possible combinations of predictor variables. A possible solution to the unbalanced repartition of values across combinations of predictor variables is to generate values decreasing from the presets *xs* for which $f(\mathbf{x}) = \hat{y}$ (i.e., the regression line) to *xs* that lead to *f(x)* multi-normal distributed around $\hat{y}$. As approximately 95% of the observations are located at two standard deviation from $\hat{y}$ (Grimmett & Stirzaker 2002), one can use in determining the set of *xs* for which values will be generated, a discrete set that contains values situated at most two root mean square errors from the predicated value (Neter et al. 1996). To avoid bias results induced by a balanced number of values across the set of identified xs, data generation will produce less values for combinations of *xs* that supplies $\hat{y}$ two root mean square errors from the regression line than for combination of *xs* that supplies $\hat{y}$ on the regression line. However, the smallest number of values generated has to be at least the number of values identified as providing arithmetic and geometric means unaffected by the size of the data generation. Consequently, the bias adjustment for multiple linear regressions with transformed dependent variable requires generation of more values than the minimum number of values for which the two means do not change significantly.

As data are generated, one can question whether or not the arithmetic mean and geometric mean will converge to the same value, which will render data generation approach as a biased procedure. Eq. 6 proves that geometric mean is smaller than arithmetic mean for more than two different values used in computations; consequently the lack of convergence to the same value of the two means. For finite number of bounded values, the arithmetic and geometric means are finite, as proven by Kolmogorv & Fomin (1999), which shows that the two means are between the smallest and the largest value.

Determination of the arithmetic and geometric mean from generated data is justified by the violation of the linear property of the arithmetic mean. The violation is found not only in the lack of equivalence between the two means, which operates basically on two different spaces, one on L^1 and one on L^2 (Grimmett & Stirzaker 2002), but also at distributions level, as one has normal residuals while the other has log-normal residuals. While violation of distributional assumptions were extensively addressed in the presence or absence of data (Aitchison & Brown 1957), the linearity violation was not studied in the absence of data. Formally, the breach of linearity is induced by the assumption that inversion function is a linear operator, which can be expressed as:

$$\begin{array}{l}\widehat{\ln y} = f(\mathrm{x}\quad) \xrightarrow{LS} \exists \mathrm{x} \quad \textit{such that } \overline{\ln y} = f(\mathrm{x}\quad) \xrightarrow{assumed} \bar{y} = \exp(f(\mathrm{x}\quad)) \\ \text{relationship facing two issues:} \\ 1)\, \overline{\ln y} \neq \ln \bar{y} \\ 2)\, \widehat{\ln y} \neq \ln \hat{y} : \textit{distributional assumptions, hence confidence}\end{array} \qquad (8)$$

Eq. 8 shows that the value predicted by *f(*x*)*, which is on the regression line, is also the mean of the normal distribution of the residuals located at x (Montgomery et al. 2006, Neter et al. 1996). The fact that $\ln(\hat{y})$ is situated on *f(*x*)* is at the center of the proposed correction for bias $\delta = \mu_{arithmetic} - \mu_{geometric}$, which is based on generated data. As generation of normally distributed data was dependent not only on the mean (i.e., *f(*x*)*) but also on the variance, one should compute the variance of the residuals from the summary statistics based on the original data. Montgomery et al (2006) proved the additive property of the sum of squares, from which can be deducted that variance of the residuals, also known as mean square error, $\sigma^2_{\varepsilon_{\ln(y)}}$, is:

$$\sigma^2_{\varepsilon_{\ln(y)}} = \frac{\sum_{i=1}^{n}(\ln(y_i)-\widehat{\ln(y_i)})^2}{n-k-1} = \frac{\sum_{i=1}^{n}(\ln(y_i)-\overline{\ln(y)})^2 - \sum_{i=1}^{n}(\widehat{\ln(y_i)}-\overline{\ln(y)})^2}{n-k-1} =$$
$$= \frac{(n-1)\times\left(\sum_{i=1}^{n}(\ln(y_i)-\overline{\ln(y)})^2\right)/(n-1)-(k-1)\times\left(\sum_{i=1}^{n}(\widehat{\ln(y_i)}-\overline{\ln(y)})^2\right)/(k-1)}{n-k-1} = \tag{9}$$
$$= \frac{(n-1)\sigma^2_{\ln(y)}-(k-1)\sigma^2_{\widehat{\ln(y)}}}{n-k-1} = \frac{(n-1)\sigma^2_{\ln(y)}}{n-k-1} - \frac{(k-1)\sigma^2_{\widehat{\ln(y)}}}{n-k-1} \xrightarrow[k\ small]{n\ increases} \frac{(n-1)\sigma^2_{\ln(y)}}{n-k-1} \to \sigma^2_{\ln(y)}$$

as $\sigma^2_{1\mathrm{n}(y)}$ is constant and independent of number of measured or observed data

$$\frac{(k-1)\sigma^2_{\widehat{\ln(y)}}}{n-k-1} \xrightarrow[k\ small\ \&\ constant]{n\ increases} 0 \tag{10}$$

where n is the total number of generated data, and k is number predictor variables

$\sigma^2_{\varepsilon_{\ln(y)}}$ is the variance of the residuals of the linear regression, $\varepsilon_{\ln(y)} = y-\hat{y}$ (10a)

$\sigma^2_{\ln(y)}$ is the variance of the logarithmic transformed data $\sigma^2_{\widehat{\ln(y)}}$ is the variance of the regression line.

In a large number of situations only the regression equation is presented, and both original data and summary statistics are no longer available (Bennett et al. 1959, Clutter et al. 1983, Giurgiu 1979). In this situation, information regarding variance can be obtained from the Chebyshev's theorem (Grimmett and Stirzaker 2002), which proved that for symmetric unimodal distribution, standard deviation can be determined from range, using the formula *σ = range/6*, as proved by McClave and Dietrich (1991). Range, in this case, can be determined either from the asymptotic behavior of the regression equation (Clutter et al. 1983), if the asymptote does exists, or from the charts that traditionally accompany the regression equations in the past (Avery & Burkhart 2001, Bennett et al. 1959; Bettinger et al. 2009, Husch et al. 2002).

Knowledge of the variance of the residuals allows the generation of a set of *ln(y)* values that are normally distributed with mean *f(*x*)* and variance $\sigma^2_{\varepsilon_{\ln(y)}}$ for each x. To avoid the assumption of linearity when it is not true, each generated data will be back-transformed in *y*'s

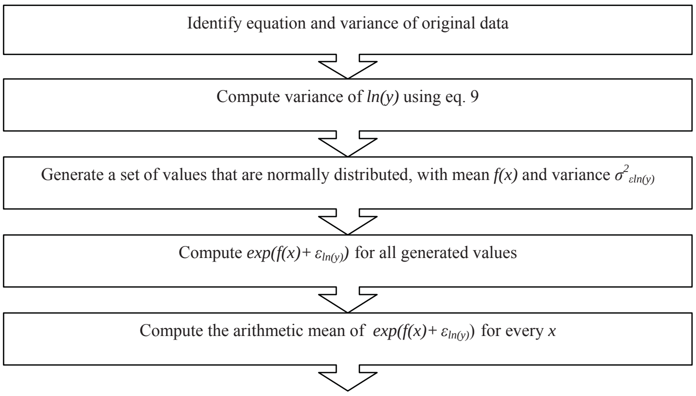
Identify equation and variance of original data
Compute variance of *ln(y)* using eq. 9
Generate a set of values that are normally distributed, with mean *f(x)* and variance $\sigma^2_{\varepsilon ln(y)}$
Compute $exp(f(x)+\varepsilon_{ln(y)})$ for all generated values
Compute the arithmetic mean of $exp(f(x)+\varepsilon_{ln(y)})$ for every *x*

Figure 2 Flowchart outlying the succession of steps needed to correct bias using generated data

original units, and subsequently the arithmetic mean of back-transformed values will be computed. This procedure adjusts for bias the predicted $y = e^{f(x)}$, as avoids the equivalence issue proven by Eq. 8 ($\overline{\ln y} \neq \ln \overline{y}$), and incorporates the difference between the geometric mean and arithmetic mean. The series of steps that correct for bias using generated data and presented above is summarized in the flowchart from Fig. 2.

Random data generation can require a large amount of values to ensure the desired estimation properties, as proven by Cochran (1977). For small sample sizes there is the possibility that the results would contradict the original data, as the resulted equation will have various derivatives, namely both positive and negative. In eventuality that random data generation requires a large number of values, then the corrections for bias proposed by Baskerville (1972), Sprugel (1983), or Beauchamp and Olson (1973) can be used. The correction basically adjusts the estimates provided by the back transformation from logarithmic units to original units according to first order moment of the lognormal distribution:

$$y\big|_{\mathbf{x}} = e^{f(\mathbf{x})+0.5\sigma^2_{\varepsilon_{\ln(y)}}}$$

If a logarithmic transformation uses a different base than *e*, then the correction proposed by Sprugel become:

$$y\big|_{\mathbf{x}} = e^{0.5\times\ln a\times\sigma^2_{\varepsilon_{l\,\mathrm{og}_a(y)}}}\,a^{f(\mathbf{x})}$$

where log$_a$*(y)* resents the logarithm of y in *a*, $a > 0$.

The approaches that correct bias only using expectation differ from two fundamental perspectives: one uses a constant correction factor, irrespective the predictor variable (Sprugel 1983), and one uses a correction that depends on the predictor variable (Beauchamp & Olson 1973). Here, only the approach proposed by Sprugel is shown, for consistency, as Beauchamp and Olson, developed their method only for simple linear regression, and this research presents corrections for multiple linear regression (Giurgiu 1979).

The change in distribution from normal to lognormal is not the focus of the present research, as corrections for the transformation of the depended variable were presented extensively in the literature in the last 50 years (Aitchison & Brown 1957). Formally, the distributional change is expressed as:

$$\begin{aligned}
&\widehat{\ln y} = f(\mathrm{x}\quad) \xrightarrow{assumed} \hat{y} = \exp(f(\mathrm{x}\quad)) \leftrightarrow y = \exp(f(\mathrm{x}\quad)) + \xi \\
&\text{where}\, \xi \sim F(\exp(\sigma^2/2), \sigma^2) \text{ and} \\
&\qquad F \text{ is the probability distribution function of the log-normal distribution} \quad (11) \\
&\text{but} \\
&\widehat{\ln y} = f(\mathrm{x}\quad) \leftrightarrow \ln y = f(\mathrm{x}\quad) + \varepsilon \leftrightarrow y = \exp(f(\mathrm{x}\quad) + \varepsilon) \\
&\text{where}\, \varepsilon \sim N(0, \sigma^2)
\end{aligned}$$

One of the objectives of the article is to illustrate two procedures available to correct bias of lognormal regression models: 1) one using data generation, 2) one based on moment estimations. The procedures are presented using two types of regression equations, whose presence is ubiquitous in forestry: 1) a simple linear regression, namely the guiding curve site index equation for slash pine (*Pinus elliottii E.*) developed by Clutter et al. (1983) based on Schumacher and Hall (1933) approach, 2) a multiple linear regression, namely the volume equation developed by Giurgiu (1979) for Norway spruce (*Picea abies L*). The bias was corrected using information on original data (e.g., summary statistics or graphs) and the final regression line. The site index equation presented in this paper was initially developed by Bennett et al (1959), and subsequently adjusted by Clutter et al (1983), is :

$$ln(height) = b_0 + b_1 age^{-1} . \qquad (12)$$

The two coefficients, determined using measurements executed in imperial units, are: b_0 = 4.6646 and b_1 = -12.4486. The coefficients were not converted to international units as the focus of the article is not the equation but the technique.

The tree volume equation developed by Giurgiu (1979) is:

$$log(volume) = b_0 + b_1 \cdot log(dbh) + b_2 \cdot log^2(dbh) + b_3 \cdot log\ (heigt) + b_4 \cdot log^2(height) \qquad (13)$$

where b_0= -4.0239, b_1= 1.9341, b_2= -0.0722, b_3= 0.6365, b_4= 0.1720, and the base of the logarithm is 10.

Results

The proposed procedure for correcting bias of log-transformed variable requires aprioric knowledge of the number of values to be generated. To identify the needed number of values, the simulation process using less than 100 values for each x and four arithmetic means, revealed significant variation for both means when less than 10 values are used in computation, but exhibit a significant reduction in variability for more than 20 values (Fig 3). Therefore, for each x 20 values will be randomly generated.

The site index equation (12) used as an example, has an asymptote at 106 feet (i.e., 32.31 m). However, for slash pine plantations the rotation age is seldom larger than 50 years, as recorded by Zarnoch & Feduccia (1984), and indicated by Carmean et al (1989), who draw site index curves based on Bennett et al (1959) data for ages less than 25 years. Therefore, it can be assumed that Bennett et all (1959) recorded stands with height of dominant trees at most 65 feet (i.e., 19.81 m). Carmean et al. (1989) showed that the minimum age for which data were collected is 10 years, indicating that the smallest height is 30 feet (i.e. 9.144 m). Consequently, the range of the original data was

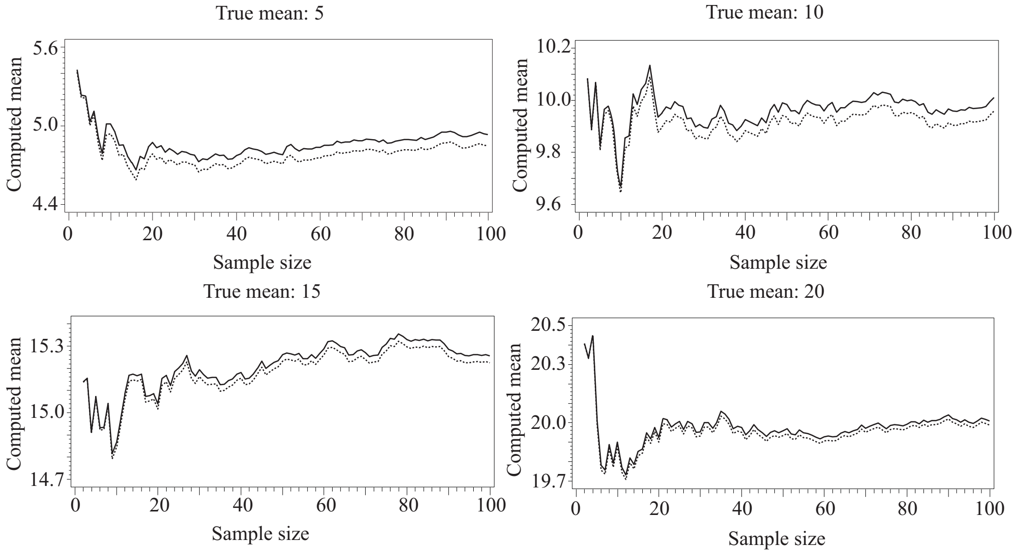

Figure 3 Flowchart outlying the succession of steps needed to correct bias using generated data

probably 35 feet (i.e., 10.67 m), which leads to a standard deviation of the logarithmic data of 0.128 $\left(i.e. \frac{\ln(65)-\ln(30)}{6}\right)$, and a variance of $\sigma^2_{\varepsilon_{\ln(y)}} = 0.0166$ (according to Eq. 9).

The generated data for the linear regression line using 20 values produced values that had both means changing unrealistically with age, in the sense that larger ages have smaller heights. This non-monotonic behavior (Rudin 1987) of the random generated data is not consistent with the growth and yield processes and requires additional investigation. As the law of large number, which is at the foundation of the estimation process, requires a large number of values, the solution to the unrealistic temporal variation was to increase the number of generated heights from 20/year to 5000/year. The 5000 values/year were selected following the same procedure that recommended the 20 values/year, but constrained not only to insignificant changes in the means but also to insignificant differences between the expected mean and computed mean. The large dataset built using 5000 values/year (i.e., 80 000 records) showed a possible representation of the original data that were used to develop the site index guiding curve. The bias correction computed using the generated data is consistently larger than the uncorrected guiding curve, as expected according to Eq. 7, confirming the bias induced by the log-transformation (Beauchamp and Olson 1973). To correct for bias Eq. 12 using the correction factor proposed by Sprugel (1983), one can multiply all the values with the $e^{0.5\sigma^2_{\varepsilon_{\ln(y)}}}$, which is 1.0083. Therefore, the Eq. 12 is

$$height = e^{4.6729-12.4486/age}$$

instead of

$$height = e^{4.6646-12.4486/age} \tag{14}$$

The bias determined using Sprugel's approach for the site index Eq. (12) is less than 1% (i.e., 0.823%).

The correction proposed by Sprugel (1983),

Table 1 Slash pine (*Pinus elliottii* E.) guiding curve statistics for corrected and uncorrected for bias values

Age	Uncorrected height (Clutter et al 1983)	Corrected height (Sprugel 1983)	Corrected height (generated data)	Variance of generated heights (original units)
10	30.56	30.82	30.56	0.063
11	34.22	34.51	34.23	0.081
12	37.61	37.92	37.61	0.097
13	40.73	41.07	40.73	0.113
14	43.62	43.98	43.60	0.133
15	46.28	46.66	46.28	0.15
16	48.74	49.15	48.75	0.159
17	51.03	51.45	51.03	0.177
18	53.14	53.59	53.14	0.196
19	55.11	55.57	55.12	0.213
20	56.95	57.42	56.96	0.212
21	58.66	59.15	58.67	0.237
22	60.27	60.77	60.27	0.249
23	61.77	62.28	61.77	0.264
24	63.18	63.70	63.16	0.275
25	64.50	65.03	64.50	0.285

is comparable with the values obtained from generated data (Table 1), indicating that both procedures could be successfully used to correct bias of logarithmic transformed variables. The generated data approach has the advantage that provides information on variance change along regression curve, which increases with age (Table 1).

Based on the findings from the guiding curve equation, the multivariable equation of Giurgiu (1979) should be corrected using 5000 values/dbh-height combination. However, for simplicity only the correction proposed by Sprugel is presented in this paper, as the results for simple linear regression indicated the consistency between the values obtained using generating data and Sprugel (1983) approach. According to Giurgiu et al (1972), the range of the original data is 12.453 (i.e., from 0.003 to 12.456), and according to Eq. 9 the range for the logarithmic volume is 3.618 [i.e., log(12.456) - log(0.003)], which renders a standard deviation of 0.603 and a variance of 0.3636. Consequently, Sprugel's correction factor is $e^{0.5\times 2.303\times\sigma^2_{\varepsilon_{\log(y)}}} = 1.5198 = 10^{0.1818}$, which should be used to multiply all the volumes obtained using the equation:

$$volume = 10^{-4.0239}\, dbh^{1.9341-0.0722\log(dbh)}\, height^{0.6365+0.172\log(height)}$$

Finally, the corrected Giurgiu's equation for the individual tree volume of Norway spruce is:

$$volume = 10^{-3.8421}\, dbh^{1.9341-0.0722\log(dbh)}\, height^{0.6365+0.172\log(height)} \qquad (15)$$

The bias determined using Sprugel's approach for the Norway spruce stem volume equation of Giurgiu (1979) is 34.2% = (1 - $10^{-0.1818}$).

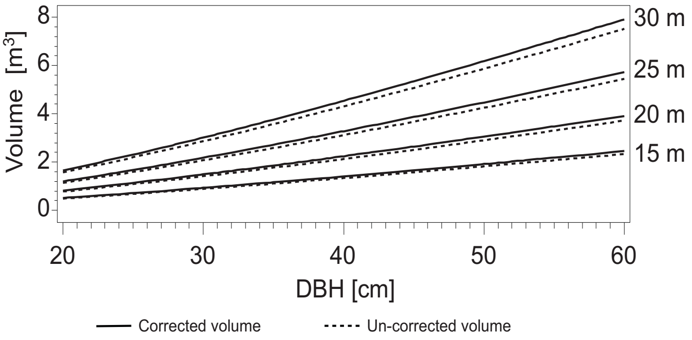

Figure 4 Impact of height and DBH on Norway spruce stem volume (corrected and uncorrected for bias) computed using Giurgiu (1979) equation for trees with DBH between 20-60 cm and four heights (i.e., 15 m, 20 m, 25m, 30 m)

Discussion

The common belief that significant results can be obtained based on a reduced amount of generated data proved not to be supported by the results. In contrary, it seems that a large amount of data has to be generated to ensure the achievement of the convergence properties, as stated by the laws of large numbers (Grimmett & Stirzaker 2002). The present results suggest that 5000 values for each x leads to the expected results, while data generated with less than 100 values/x does not necessarily produces the anticipated results. The requirement of large amount of data can create difficulties in analysis and increases the chance of computational errors. When the regression model contains several uncorrelated predictor variables that operates on a fine range of values (e.g., three predictor variables each with 100 values), the generated data produces more than 10 billion records. Consequently, the approach for bias correction based on data generation (Fig 2) encounters difficulties for models with multiple variables, which limit its general applicability. The results support this approach only for simple linear regression, situation on which the proposed procedure is justified, as it supplies not only unbiased estimates but also a measure of the variability existing in the original data. For multiple linear regressions with logarithmic dependent variable the correction proposed by Sprugel (1983), should be used, even that no information on the variability of the estimates would be available. However, the simplicity of Sprugel's approach recommends its implementation in these situations. Furthermore, Sprugel's correction changes only the intercept, as the coefficients of the predictor variables remain the same.

The correction for bias of models developed by linearising a relationship using the logarithmic function is required from two perspectives: 1) it does produces the correct results, even that sometimes from operational perspective the correction is insignificant [the case of site index equation of Clutter et al. (1983)], and 2) can lead to significant changes in estimated values [the case of Norway spruce stem volume equation of Giurgiu (1979)]. However, even for smaller biases, the correction increases with the magnitude of the predicted value, therefore on absolute scale it accretes to large values.

The bias magnitude for Giurgiu (1979) equation raises questions on the practical application of the equation, as underestimates of more than a 1/3 of the actual volume is reflected in large financial losses from both finite products perspective (i.e., seller of the raw material receives a reduced amount) as well as from management perspective (i.e., larger volumes leads to an earlier peak of the mean annual increment, consequently a shorter rotation age). However, the large bias is likely not induced by an inappropriate implementation of the double logarithmic equation, but possibly the results of the lack of data. As stem volume covers a large range of values, many under 1 m^3, the logarithmic transformation, especially base 10, lead to large negative values, as $log_{10}(0.001) = -3$. In the case of forest operators that harvest and sell wood fiber harvested from more than 1000 ha/year, the losses induced by the bias associated with Giurgiu (1979) equation can be more than 0.4 m^3/stem (Fig. 4), an estimated value of at least $1,000,000/year.

The case of Giurgiu's equation is emblematic for the situations on which the results are further inputted in other models, for example the development of stand and stock tables. In these situations, the bias can cumulates to produce results that are no longer defendable scientifically, operationally, as well as legally. Consequently, the correction for bias should be executed irrespective the magnitude of the correction, to avoid any subsequent errors, which can lead to chaotic models of the behavior of the managed forests (May 1977).

Conclusions

The logarithmic transformation of the dependent variables for models developed using regression analysis induces bias that should be corrected, regardless its magnitude. The simplest correction for bias was proposed by Sprugel (1983), which basically multiplies the back-transformed estimates with the constant value of $e^{0.5\sigma^2_{\varepsilon_{\ln(y)}}}$. While this correction is fast and easy to implement does not supplies estimates of the variability existing in the original data. Consequently, a procedure based on generated data was developed to provide unbiased estimates for both attribute of interest and variability existing along the model. The procedure reveals that valid estimates can be obtained if large number of values is generated (e.g., 5000 values/x). The procedures supplies accurate estimates for the attribute of interest and its variability, but encounters significant data processing difficulties for models with more than one predictor variable. Nevertheless, irrespective the number of predictor of variables and magnitude of the correction factor computed by Sprugel, the estimates determined using logarithmic transformations should be corrected for bias, to avoid cumulated errors or chaotic effects associated with nonlinear models. The correction for bias can be executed using either the proposed data generation procedure or Sprugel's correction factor, depending on the complexity of the model.

Acknowledgements

This work was supported by the UEFSCDI project No. PN-II-PT-PCCA-2011-3.2-1710, and USDA, NIFA, McIntire-Stennis Cooperative Forestry Research Program project No. LAZ00071.

References

Adams W.S., Titus S.J., 2009. Mixwood growth model. Department of Renewable Resources University of Alberta, Edmonton AB.

Aitchison J., Brown J.A.C., 1957. The lognormal distribution. Cambridge University Press, Cambridge UK.

Assmann E., 1970. The principles of forest yield study. Pergamon Press, Oxford UK. 506 p.

Avery T.E., Burkhart H., 2001. Forest Measurements. Mcgraw-Hill Ryerson, New York. 1-480 p.

Baskerville G.L., 1972. Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forest Research 2(1): 49-53.

Beauchamp J.J., Olson J.S., 1973. Corrections for bias in regression estimates after logarithmic transformation. Ecology 54(6): 1403-1407.

Bennett F.A., McGee C.E., Clutter J.L., 1959. Yield of old-field slash pine plantations. US Forest Service.

Bettinger P., Boston K., Siry J.P., Grebner D.L., 2009. Forest Management and Planning. Academic Press, Burlington MA, 360 p.

Bjorck A., 1996. Numerical Methods for Least Squares Problems. SIAM: Society for Industrial and Applied Mathematics, Philadelphia PA.

Carmean W.H., Hahn J.T., Jacobs R.D., 1989. Site index curves for forest tree species in the eastern United States. USDA - Forest Service. NC-128. 153 p.

Cauchy A.-L., 1821. Cours d'analyse de l'Ecole Royale Polytechnique. Imprimerie Royale, Paris FR.

Clutter J.L., Forston J.C., Pienaar L.V., Brister G.H., Bailey R.L., 1983. Timber management : a quantitative approach. Krieger Publishing Company, Malabar, FL. 352 p.

Cochran W.G., 1977. Sampling techniques. John Wiley and Sons, Singapore. 428 p.

Crow E.L., Shimizu K., 1988. Hystory, genesis and properties. Crow, E.L., and K. Shimizu (eds.) in The Lognormal Distribution, . Marcel Dekker, New York NY. 26 p.

Darmois G., 1935. Sur les lois de probabilites a estimation exhaustive. Comptes Rendus de l'Académie des Sciences 200: 1265-1266.

Edleston J., 1850. Correspondence of Sir Isaac Newton and Professor Cotes. John Parker, London UK.

Feller W., 1968. An Introduction to probability theory and its applications, volume 1. John Wiley and Sons, New York. 528 p.

Finney D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. Supplement to the Journal of the Royal Statistical Society 7(2): 155-161.

Gentleman R., Ihaka R., 2012. R. Comprehensive R Archive Network, Aukland NZ.

Giurgiu V., 1979. Dendrometrie şi auxologie forestieră. Ceres, Bucharest. 853 p.

Giurgiu V., Decei I., Armesescu S., 1972. Biometria arborilor şi arboretelor din România. Ceres, Bucharest RO.

Gould B., 2012. STATA. SataCorp, College Station TX.

Grimmett G.D., Stirzaker D.R., 2002. Probability and random processes. Oxford University Press, New York. 600 p.

Hall H.S., Knight S.R., 2005. Higher Algebra: A Sequel to elementary algebra for schools. Adamant Media Corporation, New York NY. 581 p.

Husch B., Beers T.W., Kershaw J.A., 2002. Forest mensuration. Wiley. 456 p.

Kolmogorov A.N., Fomin S.V., 1999. Elements of the theory of functions and functional analysis. Dover Publications Inc, Mineola NY., 257 p.

Koopman B.O., 1936. On distributions admitting a sufficient statistic. Transactions of the American Mathematical Society 39: 399-409.

Legendre A.-M., 1805. Sur la methode des moindres quarres. P. 72-75 in Nouvelles methodes pour la determination des orbites des cometes. Firmin Didot, Paris.

Liew C.K., Choi U.J., Liew C.J., 1985. A data distortion by probability distribution. ACM Transactions on Database Systems 10(3): 395-411.

May R.M., 1977. Thresholds and breakpoints in ecosystems with a multiplicity of stable states. Nature 269(5628): 471-477.

McClave J.T., Dietrich F.H., 1991. Statistics. Dellen Publishing Company, New York. 928 p.

Miller D.M., 1984. Reducing transformation bias in curve fitting. The American Statistician 38(2): 124-126.

Montgomery D.C., Peck E.A., Vining G.P., 2006. Introduction to linear regression analysis. Wiley, New York NY.

Neter J., Kutner M.H., Nachtsheim C.J., Wasserman W.. 1996. Applied linear statistical models. WCB McGraw-Hill, Boston. 1408 p.

Nie N.H., Hull C.H., 2012. SPSS. IBM, Armonk NY.

Nitschke C.R., Innes J.L., 2008. A tree and climate assessment tool for modelling ecosystem response to climate change. Ecological Modelling 210(3): 263-277.

Plackett R.L., 1950. Some Theorems in Least Squares. Biometrika 37(1-2): 149-157.

Poole D., 2005. Linear algebra. Thomson Brooks/Cole, Toronto. 712 p.

Pretzsch, H. 2009. Forest dynamics, growth and yield. Springer, Berlin. 664 p.

Rao C.R., 1973. Linear statistical inference and its applications. John Wiley & Sons, Inc, New York. 625 p.

Rudin W., 1987. Real and complex analysis. McGraw-Hill.

SAS Institute, 2010. SAS. SAS Institute, Cary NC.

Schabenberger O., Pierce F.J., 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton FL.

Schumacher F.X., Hall F.d.S., 1933. Logarithmic Expression of Timber-Tree Volume. Journal of Agricultural Research 47(9): 719-734.

Sprugel D.G., 1983. Correcting for Bias in Log-Transformed Allometric Equations. Ecology 64(1): 209-210.

Spurr S.H., 1952. Forest inventory. The Ronald Press Company, New York. 476 p.

Steele J.M., 2004. The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities. Cambridge University Press, Cambridge UK. 304 p.

Tran Z.V., 1997. Estimating Sample Size in Repeated-Measures Analysis of Variance. Measurement in Physical Education and Exercise Science 1(1): 89-102.

Weiskittel A.R., Hann D.W., Kershaw J.A., Vanclay J.K., 2011. Forest Growth and Yield Modeling. Wiley-Blackwell, Chichester UK. 415 p.

Williams G.P., 1997. Chaos theory tamed. Joseph Henry Press, Washington, D.C. 499 p.

Zar J.H., 1996. Biostatistical analysis. Prentice - Hall, Upper Saddle River NY. 662 p.