BIAS CORRECTION IN LOGARITHMIC REGRESSION AND COMPARISON WITH WEIGHTED REGRESSION FOR NON-LINEAR MODELS

WEI SHENG ZENG
Academy of Forest Inventory and Planning, State Forestry Administration, #18 Hepingli East Street
Eastern District, Beijing 100714, China
zengweisheng@sohu.com
SHOU ZHENG TANG
Institute of Forest Resources Information, Chinese Academy of Forestry, #1 Dongxiaofu, Xiangshan Street
Haidian District, Beijing 100091, China
stang@caf.ac.cn

Non-linear models with heteroscedasticity are commonly used in ecological and forestry modeling, and logarithmic regression and weighted regression are usually employed to estimate the parameters. Using the single-tree biomass data of three large samples, the bias correction in logarithmic regression for non-linear models was studied and comparison between logarithmic regression and weighted regression was discussed in this paper. Firstly, the immanent cause producing bias in logarithmic regression was analyzed, and a new correction factor was presented with which three commonly used bias correction factors were examined together, and the results showed that the correction factors presented here and derived by Baskerville (1972) should be recommended, which could insure the corrected model to be asymptotically consistent with that fitted by weighted regression. Secondly, the fitting results of weighted regression for non-linear models, using the weight function based on residual errors of the model estimated by ordinary least squares (OLS) and the general weight function ($w=1/f(x)^2$) presented by Zeng (1998) respectively, were compared with each other that showed two weight functions worked well and the general function was more applicable. It was suggested that the best way to fit non-linear models with heteroscedasticity would be using weighted regression, and if the total relative error of the estimates from the model fitted by the general weight function was more than a special allowance such as ±3%, a better weight function based on residual errors of the model fitted by OLS should be used in weighted regression.

Keywords: non-linear model; biomass model; logarithmic regression; weighted regression; bias correction; heteroscedasticity

1. Introduction

Many models used in ecology and forestry are non-linear models, such as volume equations and biomass equations. The ecological and forestry models usually exhibit heteroscedasticity, that is, the error variance is not constant over all observations. To eliminate the influence of heteroscedasticity, it is necessary to estimate the parameters using logarithmic transformation or weighted regression[1-8].

Finney[1] noticed early the problem of biased estimation from logarithmic transformation, and derived an unbiased estimator for bias correction. Baskerville[2] indicated that while the use of a logarithmic transformation was valid, the retransformation was to the median rather than the mean; and presented an unbiased estimate of the mean based on the sample variance s^2, where the bias correction factor was $\exp(s^2/2)$ which subsequently became widely used[6,9-12]. Beauchamp and Olson[13] considered that the correction factor presented by Baskerville was still biased because of the sample variance s^2 being only the unbiased estimate of population variance σ^2 of which the true value was unknown, and then derived another correction factor expressed by $\Psi(t)$ function. Flewelling and Pienaar[3] reviewed all the kinds of bias correction factors, and presented some guidelines for the selection of estimators: for moderate to large sample sizes (usually more than 30), the maximum relative difference between any two estimates under most circumstances except extreme extrapolation is $\exp(3/2s^2)$; and if this magnitude of error is not of consequence, any of the correction factors presented by Baskerville, Finney or

Beauchamp and Olson should be adequate; if the sample size is small, it is necessary to use more complicated estimators. Snowdon[4] presented a ratio estimator for bias correction in logarithmic regressions, and concluded that the new method could give more reliable results than the methods presented by Baskerville and Finney respectively, and would be less sensitive to departures from the assumption of a lognormal distribution than the other two methods.

In addition to logarithmic transformation, weighted regression was usually used to remove the influence of heteroscedasticity on parameter estimation. Zeng[14] and Zeng *et al*[15] presented the general weighting function, that is $w = 1/f(x)^2$, based on the studies of heteroscedasticity of volume equations and biomass equations; Zhang *et al*[16], Xu[17] and Parresol[5-6] all involved the studies of heteroscedasticity of biomass data, and presented some weighting functions. What is the difference between weighted regression and logarithmic regression? Which situation is the logarithmic regression suitable for? And what is the immanent cause producing bias in logarithmic regression? In this paper, based on the analysis of observed biomass data, the logarithmic regression of nonlinear models and the bias correction will be studied again at first; Secondly, relationship between logarithmic regression and weighted regression will be analyzed; and finally, conclusions and suggestions will be presented for regression estimation of nonlinear models in practice.

2. Data and Method

2.1. *Data*

The data used in this paper are aboveground biomass data from destructive sampling, including three parts: (i) 132 sample trees collected by the South Term of the National Biomass Modeling Program in 1997 from the Lizhai Forest Farm of Dexing County in Jiangxi Province, involving Chinese fir (*Cunninghamia lanceolata*), Masson pine (*Pinus massoniana*), and several broadleaved species such as *Quercus*, *Phoebe*, and *Cinnamomum*. (ii) 237 sample trees collected in 2009 from the National Biomass Modeling Program for Continuous Forest Inventory (NBMP-CFI) involving two species, larch (*Larix* spp.) and Masson pine. The number of trees for larch is 119 which were located in four provinces of north-eastern China, and the number of trees for Masson pine is 118 which were located in nine provinces of southern China. (iii) 79 sample trees from published papers, involving two species: green weight data for 39 willow oak (*Quercus phellos*) trees from the State of Mississippi, USA[5]; green mass data for 40 slash pine (*Pinus elliottii*) trees from the State of Louisiana, USA[6]. The studies in this paper are mainly based on the first two parts of data (the statistics are listed in Table 1, and the relationship between aboveground biomass and tree diameter is showed in Fig.1), and the third part of data are served for additional analysis.

Table 1. The statistics of above-ground biomass data

Data sources	Sample trees	Variables	Mean	Max	Min	S.D.
Mixed Species in Jiangxi, China	132	Diameter at breast height/cm	13.2	25.6	5.0	5.7
		Above-ground biomass/kg	66.00	329.70	3.42	68.76
Larch in North-Eastern China	119	Diameter at breast height/cm	19.8	41.8	5.3	10.8
		Above-ground biomass/kg	213.16	847.66	3.99	227.58
Masson pine in Southern China	118	Diameter at breast height/cm	19.5	41.4	5.2	10.8
		Above-ground biomass/kg	197.34	797.05	4.08	222.47

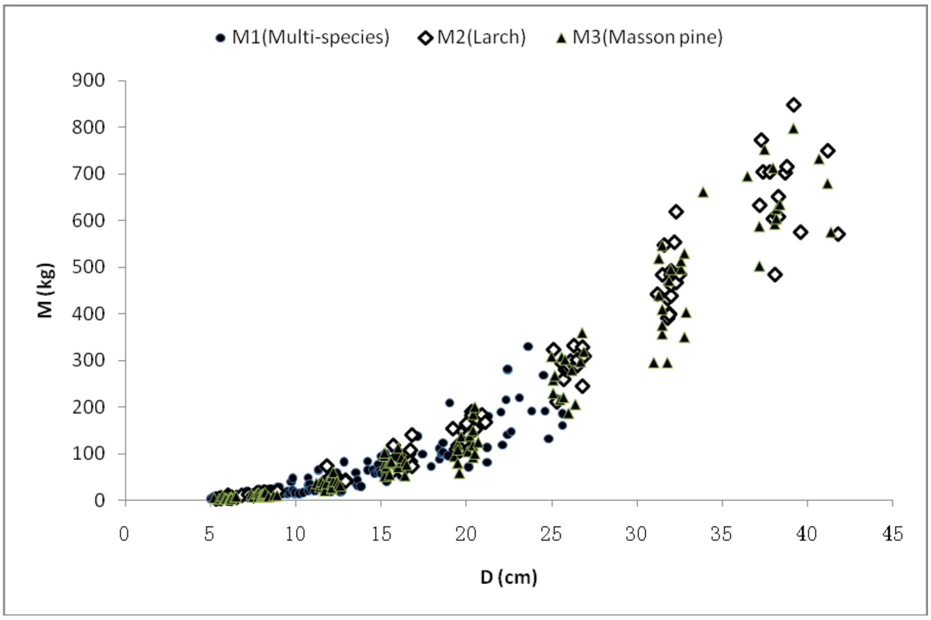

Fig.1 Relationship between aboveground biomass and tree diameter

2.2. *Method*

To eliminate the influence of heteroscedasticity, it is usually to use logarithmic regression or weighted regression for the parameter estimation of nonlinear models in ecology and forestry. The simplest and most commonly used tree biomass model is the power function as follows:

$$M=aD^b+\varepsilon \tag{2.1}$$

where M is the tree biomass, D is the diameter at breast height, a, b are parameters, and ε is the error term. Eq. (2.1) was sometimes expressed as the following form[5]：

$$M=aD^b\varepsilon \tag{2.2}$$

The error term in eq. (2.1) is additive, which usually means that ε is random, independent and normally distributed, and its variance is homoscedastic, that is, E(ε)=0 and Var(ε)=σ^2. And, the error term in eq. (2.2) is multiplicative, which usually means that ε is dependent upon diameter and its variance is heteroscedastic. However, based on the expression of eq. (2.2), it will not be sound to take E(ε)=0 and Var(ε)=σ^2, instead we should take E(ζ)=0 and Var(ζ)=σ'^2 where ζ=ln(ε). For the sake of simplicity and un-confusion, we think that the power function of tree biomass can be expressed as eq. (2.1) whether the error variance is homoscedastic or heteroscedastic. If the variance of error term ε is heteroscedastic and the relative error ε' has homoscedastic variance, then eq. (2.1) can be expressed as follows:

$$M=aD^b(1+\varepsilon') \tag{2.3}$$

where $\varepsilon'=\varepsilon/aD^b$. The logarithmic form is：

$$\ln M=\ln a+b\ln D+\ln(1+\varepsilon') \tag{2.4}$$

where ln is the natural logarithm. Eq.(2.4) seems to be equivalent to the following standard form of linear model:

$$y=a_0+b_0x+\xi \tag{2.5}$$

where y=ln$M$, x=ln$D$, a_0, b_0 are parameters, and ξ is error term. Using the ordinary least squares (OLS) method to fit eq.(5), the estimate of biomass can be obtained by the following equation:

$$\hat{M} = \exp(a_0 + b_0 \ln D) \qquad (2.6)$$

But in fact, this is not correct. Because using OLS method should meet the need that the mathematical expectation of error ξ be equal to zero, that is $E(\xi)=0$, but the mathematical expectation of the term $\ln(1+\varepsilon')$ in eq. (2.4), corresponding to the error term ξ in eq. (2.5), is certainly not equal to zero, that is $E[\ln(1+\varepsilon')] \neq 0$. Let us discuss it in detail.

Since the ε' in eq. (2.3) is relative error, the theoretical distribution range is $(-\infty, +\infty)$, and the empirical distribution range is usually between ±0.5. We assume that relative error ε' is random, independent and normally distributed, and its variance is homoscedastic, then $E(\varepsilon')=0$ and $\mathrm{Var}(\varepsilon')=\sigma^2$. Because $\ln(1+\varepsilon')$ is always less than ε', thus $E[\ln(1+\varepsilon')] \neq 0$, which should be equal to some negative value, assuming to be $-c$. We know the OLS estimation of eq. (2.5) must meet $E(\xi)=0$. Then, the parameters of eq. (2.4) and (2.5) have the following relationships:

$$\xi = \ln(1+\varepsilon') + c \qquad (2.7)$$

$$a_0 = \ln a - c \qquad (2.8)$$

That is to say, to meet the need $E(\xi)=0$ of the OLS estimation for linear eq. (2.5) from logarithmic transformation, some part (i.e., c) of the value was separated from parameter a_0 to error term ξ, thus the parameter a_0 would be underestimated. This is the immanent cause that the estimate of logarithmic regression needs to be corrected. Now, let us analyze the size of c-value which results in the bias directly.

From eq. (2.7), we can obtain:

$$1+\varepsilon' = \exp(\xi - c) = \exp(-c)\exp(\xi) \qquad (2.9)$$

Since $E(\varepsilon')=0$, from the mathematical expectation of the two sides of eq. (2.9), we can obtain $\exp(-c)=1/E[\exp(\xi)]$, then:

$$\exp(c) = E[\exp(\xi)] = \int \sum_{k=0}^{\infty} \frac{1}{k!} x^k f_{\xi}(t)dt \approx \int (1 + t + 0.5t^2) f_{\xi}(t)dt = 1 + 0.5\sigma^2$$

That is to say, the estimate of c-value is nearly equal to $\ln(1+s^2/2)$, where s is the standard error of estimate for eq. (2.5) fitted by the OLS method. From eq. (2.8), we known that the unbiased estimate of $\ln a$ should be a_0+c, then the corrected eq. (2.6) would be expressed as:

$$\hat{M} = \exp(a_0 + c + b_0 \ln D) = (1+s^2/2)\exp(a_0 + b_0 \ln D) \qquad (2.10)$$

It means that the bias correction factor is $1+s^2/2$. In this paper, we call it the first correction factor:

$$CF_1 = 1 + s^2/2 \qquad (2.11)$$

As for comparison, other three correction factors are considered:

$$CF_2 = \exp(s^2/2) \qquad (2.12)$$

$$CF_3 = \exp\{s^2/2[1 - s^2(s^2+2)/4n + s^4(3s^4+44s^2+84)/96n^2]\} \qquad (2.13)$$

$$CF_4 = \sum M / \sum \hat{M} \qquad (2.14)$$

where CF_2 is the most commonly used correction factor which was presented by Baskerville[2]; CF_3 is the approximate expression of the correction factor $g(s^2/2)$ presented by Finney[1]; CF_4 is the ratio correction factor presented by Snowdon[4].

Besides the afore-mentioned logarithmic regression, the weighted regression can be directly used to estimate the parameters of eq. (2.1) or (2.3). Three situations are considered:

(i) the OLS method, being equivalent to the weighting function $w=1$;

(ii) the weighted regression method using the general weighting function $w=1/f(D)^2$

which is the best weighting function if the relative error has homoscedastic variance;

(iii) the weighted regression method using the special weighting function $w = 1/D^{c_1}$ which is from the regression relationship between diameter D and residual squares (e^2) of the OLS estimates, $e^2 = c_0 D^{c_1}$, where c_0, c_1 are parameters.

According to the analysis above, if we assume that the variance of errors is heteroscedastic and that of relative errors is homoscedastic, then the corrected estimate of logarithmic regression should be the same as the estimate of weighted regression based on the general weighting function. But the correction factors are approximate estimates, and the estimates of weighted regression for nonlinear models are also asymptotic values from iterative algorithm, so the two kinds of results are not exactly the same.

To compare the results from logarithmic regression and weighted regression, five statistics are selected for evaluating the goodness-of-fit, which are mean error, mean absolute error, total relative error, average systematic error, and mean percent standard error. They are calculated as follows[5-6,15,18]:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{2.15}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{2.16}$$

$$TRE = \sum_{i=1}^{n}(y_i - \hat{y}_i) / \sum_{i=1}^{n}\hat{y}_i \times 100 \tag{2.17}$$

$$ASE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) / \hat{y}_i \times 100 \tag{2.18}$$

$$MPSE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| / \hat{y}_i \times 100 \tag{2.19}$$

where n is sample number, y_i is observed value of biomass (i.e., M), $\hat{y}_i$ is estimated value of biomass (i.e., $\hat{M}$). The smaller the values of statistics above, better the prediction of the models.

3. Results and Analysis

3.1 *Logarithmic regression*

The single-tree aboveground biomass data mentioned in Table 1 were used in this paper. Firstly, convert the data of diameter D and biomass M into logarithmic forms; then through the procedure of regression analysis in Excel, eq. (2.5) was fitted by using the OLS method. Finally, based on the OLS estimates, four correction factors were calculated from eq. (2.11) to (2.14). The results are listed in Table 2. Then, according to eq. (2.15) to (2.19), the evaluation statistics of logarithmic regression model (2.5) and the corrected models were computed out, which are listed in Table 3.

Table 2. The fitting results of tree above-ground biomass models from logarithmic regression

Data sources	Sample	Parameter estimates			Fit statistics		Correction factors			
	number	a_0	exp(a_0)	b_0	R^2	s^2	CF_1	CF_2	CF_3	CF_4
Mixed Species in Jiangxi, China	132	-2.42611	0.088380	2.42929	0.90233	0.12901	1.06450	1.06663	1.06660	1.05761
Larch in North-Eastern China	119	-2.41271	0.089573	2.46339	0.98685	0.03113	1.01556	1.01569	1.01568	0.98878
Masson pine in Southern China	118	-2.63440	0.071762	2.50478	0.97675	0.05804	1.02902	1.02944	1.02944	1.01402

Table 3. The evaluation statistics of tree above-ground biomass models from logarithmic regression

Data sources	*CF*	*ME*	*MAE*	*TRE*	*ASE*	*MPSE*
Mixed Species in Jiangxi, China	CF_0	3.60	17.52	5.76%	6.67%	30.91%
	CF_1	-0.43	17.98	-0.65%	0.21%	29.95%
	CF_2	-0.56	18.00	-0.85%	0.01%	29.92%
	CF_3	-0.56	18.00	-0.84%	0.01%	29.92%
	CF_4	0.00	17.92	0.00%	0.86%	30.02%
Larch in North-Eastern China	CF_0	-2.42	26.51	-1.12%	1.58%	13.72%
	CF_1	-5.77	26.62	-2.64%	0.02%	13.42%
	CF_2	-5.80	26.63	-2.65%	0.01%	13.42%
	CF_3	-5.80	26.63	-2.65%	0.01%	13.42%
	CF_4	0.00	26.58	0.00%	2.73%	14.02%
Masson pine in Southern China	CF_0	2.73	32.25	1.40%	2.87%	19.88%
	CF_1	-2.92	32.67	-1.46%	-0.03%	19.34%
	CF_2	-3.00	32.68	-1.50%	-0.07%	19.34%
	CF_3	-3.00	32.68	-1.50%	-0.07%	19.34%
	CF_4	0.00	32.42	0.00%	1.45%	19.61%

Note: Correction factor CF_0=1 means the model not being corrected.

It is shown in Table 3 that the biases of uncorrected models from logarithmic regression are not ignorable in some extent. The model for mixed species in Jiangxi was most typical, and the *TRE* and *ASE* of the fitted model reached to 5.76% and 6.67% respectively. However, the *TRE* of the model fitted with data for larch in North-eastern China was negative, that is, the estimated value was larger than the measured value, which probably resulted from the dissatisfactory sample structure. The correction factors CF_2 and CF_3 are almost the same, and are very close to CF_1. So the evaluation statistics of the corrected models with the three factors almost have no difference, and the *ASE*'s are all close to zero (for model with homoscedastic variance of relative errors, the *ASE* is theoretically equal to zero), and the *TRE*'s are about ±3%. As for the model corrected by CF_4, the *ME* and *TRE* are exactly equal to zero, the *ASE* reaches to about ±2%, and the other two statistics have no significant differences among the models.

3.2 *Weighted regression*

The eq. (8.1) was fitted by using nonlinear regression method through Marquardt iterative algorithm. For comparison, the OLS method was used at first; then the weighted regression method was used by using the two afore-mentioned weighting functions. The results are listed in Table 4. Then, according to eq. (2.15) to (2.19), the evaluation statistics of the OLS regression model (2.1) and the weighted regression models were computed out, which are listed in Table 5. In addition, the residual plot of aboveground biomass models by the OLS method is shown in Fig.2.

Table 4. The fitting results of tree above-ground biomass models from weighted regression

Data sources	Sample numbers	Weighing functions	Parameter estimates (*t*-values)		Fit statistics	
		w	*a*	*b*	R^2	s^2
Mixed Species in Jiangxi, China	132	1	0.134835 (2.45)	2.30490 (17.17)	0.81644	867.95
		$1/f(D)^2$	0.098586 (5.43)	2.41130 (33.07)	0.81536	873.08
		$1/D^{3.90}$	0.098995 (4.93)	2.40972 (32.01)	0.81541	872.85
Larch in North-Eastern China	119	1	0.252916 (4.22)	2.15767 (32.16)	0.95716	2218.41
		$1/f(D)^2$	0.092345 (12.56)	2.45813 (88.86)	0.94383	2909.20
		$1/D^{2.16}$	0.141378 (6.74)	2.32523 (53.49)	0.95463	2349.83
Masson pine in Southern China	118	1	0.147935 (3.47)	2.29927 (28.24)	0.94764	2591.12
		$1/f(D)^2$	0.076325 (9.75)	2.49284 (69.64)	0.94408	2767.92
		$1/D^{3.72}$	0.079449 (7.65)	2.47971 (61.09)	0.94488	2727.97

Note: Weighting function *w*=1 means the OLS method being used. It is the same in Table 5.

Table 5. The evaluation statistics of tree above-ground biomass models from weighted regression

Data sources	*w*	*ME*	*MAE*	*TRE*	*ASE*	*MPSE*
Mixed Species in Jiangxi, China	1	-0.72	18.15	-0.98%	-4.62%	29.82%
	$1/f(D)^2$	-0.12	17.95	-0.16%	0.00%	29.93%
	$1/D^{3.90}$	-0.09	17.95	-0.12%	-0.02%	29.93%
Larch in North-Eastern China	1	-3.81	27.19	-1.76%	-13.65%	20.28%
	$1/f(D)^2$	-5.16	26.49	-2.36%	0.00%	13.42%
	$1/D^{2.16}$	-0.02	25.61	-0.01%	-4.80%	15.21%
Masson pine in Southern China	1	-2.89	32.07	-1.46%	-10.73%	21.08%
	$1/f(D)^2$	-1.42	32.44	-0.72%	0.00%	19.34%
	$1/D^{3.72}$	-0.55	32.24	-0.28%	-0.34%	19.30%

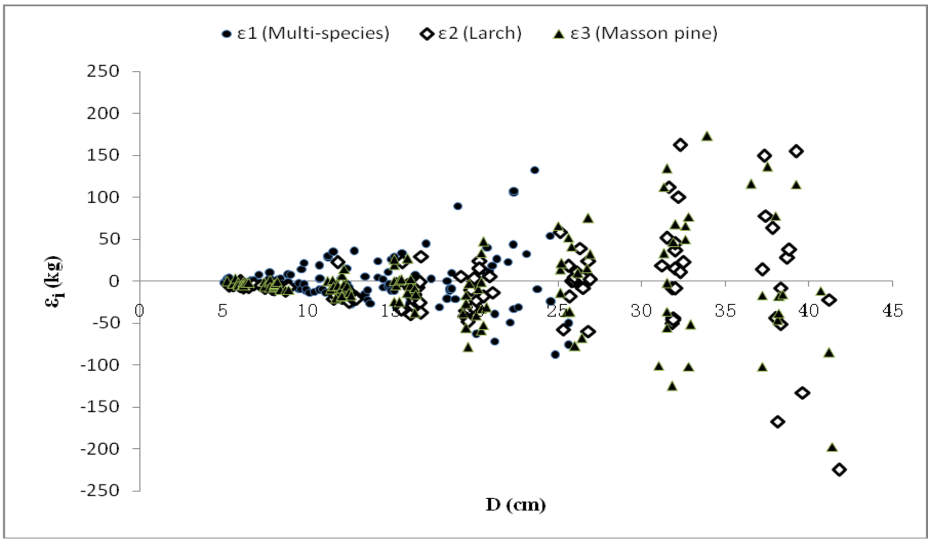

Fig.2 Residual plot of aboveground biomass models by the OLS method

It is obvious in Fig.2 that the residual errors, which were random and independent, increased with the growing diameter, and consequently exhibited heteroscedasticity. Then, the Park test[19] was used to construct the following linear equation:

$$\ln\varepsilon^2 = \gamma_0 + \gamma_1 \ln D + \delta \tag{3.1}$$

where ε is residual error of the OLS regression model, γ_0 and γ_1 are parameters. The t-test was utilized to examine whether or not the parameter γ_1 in eq. (3.1) was equal to zero. The results showed that the estimates of parameter γ_1 in eq. (3.1) for the three models by the OLS method in Table 4 were 3.90, 2.16 and 3.72 respectively, and the t-values were 11.66, 6.59 and 12.38 respectively, which indicated that parameter γ_1 was very significantly different from zero, and heteroscedasticity was obviously existed.

In the nonlinear regression method, the nonlinear model is expressed with linear Taylor's series approximation, so it is not like the fitting result of linear model where the *ME* and *TRE* are both equal to zero. This feature is shown in Table 5. Because the influence of heteroscedasticity was not eliminated in the OLS method, the three models for *w*=1 have obvious systematic errors. For example, the *ASE* of the model fitted with data for larch in North-eastern China reached to -13.65%; even the smallest *ASE* of the model fitted with data for mixed species in Jiangxi also reached to -4.26%. But, the models estimated by weighted regression had no distinctive difference whether using the general or special weighting functions, where the *ASE* values were equal or close to zero (only the *ASE* value of the model fitted for larch in North-eastern China using special weighting function reached to -4.80%), and the *TRE* values were all less than ±3%. However, it is shown in Table 4 that the t-values of parameter estimates fitted with general weighting function were larger than those fitted with special weighting function, which meant that using general weighting function could eliminated the heteroscedasticity better, and obtained more reliable parameter estimates. Furthermore, from the comparison between Table 5 and Table 3, it is shown that the fitting results from weighted regression with general weighting function were very close to those from logarithmic regression corrected by correction factors calculated from eq. (2.11) to (2.13), which is consistent with our expectation.

4. Discussions

It is well known that when a nonlinear model was transformed to linear logarithmic form and estimated by the OLS method, an inherent negative bias would produced. As for bias correction, many researchers have studied and presented several correction factors among which the correction factor exp(s^2/2) recommended by Baskerville[2] have being widely used in practice. Snowdon[4] had some doubt on the applicability of the correction factor and recommended a simple ratio estimator for bias correction. In this paper, a new correction factor was derived from the analysis of logarithmic transformation, and comparison with the two commonly used correction factors was made based on three datasets of single-tree aboveground biomass. The results showed that the correction factors had no evident difference and the corrected models were similar to those fitted by weighted regression. The ratio estimator for bias correction recommended by Snowdon[4] was also feasible and effective, but it was presented from another viewpoint, not from the property itself of logarithmic transformation. We suggest that the correction factors derived from the logarithmic transformation are used for bias correction which would result in consistent models with weighted regression.

As for weighted regression, the key point is to determine the weighting function. According to the review paper by Parresol[5], the commonly used approach is to fit a variance function with the OLS residuals. From the comparison between Table 4 and Table 5, the estimates of weighted regression from general weighting function and special weighting function based on the OLS residuals had no obvious difference, and the estimates from general weighting function were more reliable. The authors used the green weight data for 39 willow oak trees from Mississippi[5] and for 40 slash pine trees from Louisiana[6] to compare the fitting results of weighted regression for two weighting functions (see Table 6).

Table 6. The comparison of fitting results of weighted regression for two weighting functions

Data	Models	Using special weighting function (from Parresol)					Using general weighting function (from Zeng)				
		b_0	b_1	b_2	R^2	*SEE*	b_0	b_1	b_2	R^2	*SEE*
Willow oak *n*=39	Wood	25.74948	0.02731		0.98	182.32	26.244	0.027307		0.98	182.31
	Bark	-0.51532	0.10253		0.94	41.01	-0.35661	0.10241		0.94	41.01
	Crown	117.195	0.057502	-4.61687	0.81	21.15	104.68	0.056644	-4.0816	0.80	21.27
	Total	46.381	0.031558		0.98	217.37	52.653	0.031380		0.98	214.68
Slash pine *n*=40	Wood	0.016363	1.0585		0.98	27.44	0.016366	1.0585		0.98	27.44
	Bark	0.046277	2.2093		0.96	5.11	0.044298	2.2246		0.96	5.16
	Crown	0.027378	3.6804	-1.2624	0.89	14.22	0.033816	3.6536	-1.3139	0.89	14.42

Note: (i) *SEE* is the standard error of estimate, that is the *s* in Table 2 and Table 4; (ii) The expression of models could be found in the Ref. 5 and 6; (iii) The fitting results of weighted regression for special weighting function were directly cited from Ref. 5 and 6, except for R^2 and *SEE* of slash pine which were calculated from the model parameters.

It is obvious from Table 6 that the fitting results of weighted regression, whether using general weighting functions or special ones from Parresol[5-6], were not very different from each other. As you know, it is not easy to derive the optimal weighting function from the OLS residuals. When the number of variables increases and the expression of model becomes more complicated, modeling the relationship between independent variables and the OLS residuals would be very difficult. For example, the weighting function of crown green weight model for willow oak presented by Parresol[5] was complicated, which was expressed as:

$$w=(D^2H*LCL/1000)^{1.646}*\exp(-0.00406H^2)$$

where *LCL* is the live crown length, *D* is the diameter at breast height, and *H* is the tree height. The general weighting function ($w=1/f(x)^2$) derived from the model itself is simple and convenient in application, and is also the optimal one if the relative errors have homoscedastic variance, which should be applied to weighted regression estimation.

From analyzing the criterion of parameter estimation, we know that it is the *ME* and *TRE* that would be equal to zero for the models from the OLS regression (for nonlinear models, they are not exactly equal to zero because of the linear Taylor's series approximation), and it is the *ASE* that would be zero for the models from the weighted regression with general weighting function. This feature is shown in Table 5. For a desirable dataset of samples or the fit data without obvious heteroscedasticity, both the *TRE* and *ASE* of the models should be close to zero, that is, the models from the OLS regression and weighted regression tend to be the same[15]. If we use the green weight data of willow oak published by Parresol[5] for modeling, the *TRE* and *ASE* of the models from the OLS regression and weighted regression would be all less than ±0.5%, and the *MPSE* less than 10%. Because it is generally difficult to obtain desirable data, the fitting results from the OLS regression and weighted regression would be different in some extent. When using weighted regression to estimate the model parameters, the general weighting function should be selected at first; if the *TRE* of the model is a bit large, for example exceeding ±3%, then a special weighting function derived from the OLS residuals would be necessary for the weighted regression.

In addition, it should be pointed out that for the generalized models such as tree volume or biomass equations, the evaluation of models would not be limited to the statistics of the total sample. In general, the prediction effects in the range of size classes of independent variables should be taken into account. Thus, we usually need to evaluate the statistics of size classes, but here we would not like to discuss in detail.

5. Conclusions

Using the single-tree aboveground biomass data of three large samples, the logarithmic regression and bias correction of nonlinear models were studied, and the immanent cause producing bias in logarithmic regression was pointed out, and comparison between logarithmic regression and weighted regression was made. From the study, we can present the following conclusions:

(i) For the nonlinear models exhibiting heteroscedasticity, they could be transformed to linear logarithmic forms and estimated by the OLS method, but bias correction is necessary. It is recommended to use the correction factors presented in this paper or presented by Baskerville[2], which are derived from the property itself of logarithmic transformation, and can assure the corrected models to be consistent with those from the weighted regression.

(ii) As for weighted regression, the key point is to determine the weighting function. From the comparison of weighted regressions between the commonly used special weighting function derived from the OLS residuals and the general weighting function presented by Zeng[14], it is showed that the general weighting function has perfect applicability, produces more reliable parameter estimates, and is optimal if relative errors of the model have homoscedastic variance. In the case of estimating complicated model, the general weighting function will be more outstanding.

(iii) Both logarithmic regression and weighted regression could be used to eliminate the influence of heteroscedasticity, and if the model from logarithmic regression was corrected properly, it would be almost the same as that from weighted regression. Thus, if the nonlinear model could be converted into linear logarithmic form, then not only weighted regression but also logarithmic regression can be used for parameter estimation. Since the application of nonlinear regression method is very common in ecology and forestry, for the nonlinear models exhibiting heteroscedasticity, it is recommended to use weighted regression directly with the general weighting function. If the *TRE* of the model from weighted regression with the general weighting function is a bit large, for example exceeding ±3%, then the reasons resulting in large *TRE* value should be analyzed, and a special weighting function derived from the OLS residuals may be necessary for the weighted regression.

Acknowledgements

This work was initiated as part of the National Biomass Modeling Program for Continuous Forest Inventory (NBMP-CFI) funded by the State Forestry Administration of China. We acknowledge the NBMP-CFI teams of the Academy of Forest Inventory and Planning and Central South Forest Inventory and Planning Institute of SFA for data collection.

References

1. Finney, D. J. On the distribution of a variate whose logarithm is normally distributed [J]. *J. R. Statist. Soc*., Suppl.7 (1941):155-161.
2. Baskerville, G. L. Use of logarithmic regression in the estimation of plant biomass [J]. Can. *J. For. Res.*, **2** (1972):49-53.
3. Flewelling, J. W.; Pienaar, L. V. Multiplicative regression with lognormal errors [J]. *For. Sci.,* **27** (1981): 281-289.
4. Snowdon, P. A ratio estimator for bias correction in logarithmic regressions [J]. *Can. J. For. Res.*, **21** (1991): 720-724.
5. Parresol, B. R. Assessing tree and stand biomass: a review with examples and, critical comparisons [J]. *For. Sci.*, **45** (1999): 573-593.
6. Parresol, B. R. Additivity of nonlinear biomass equations [J]. *Can. J. For. Res.*, **31** (2001): 865-878.
7. Zeng, W.S.; Luo, Q. B. On Method of Mathematical Modeling for Forestry Tables [J]. *Central South Forest Inventory and Planning*, **20** (2001):1-4 (in Chinese).
8. Wang, Z. F. 2006. *On the forest biomass's modeling and precision analysis* [D]. Beijing Forestry University, China (in Chinese).
9. Wiant, Jr., H. V.; Harner, E. J. Percent bias and standard error in logarithmic regression [J]. *For. Sci.*, **25** (1979):167-168.
10. Sprugel, D. G. Correcting for bias in log-transformed allometric equations [J]. *Ecology*, **64** (1983):209-210.
11. Lehtonen, A.; Mäkipää, R.; Heikkinen, J.; Sievänen, R.; Liski, J. Biomass expansion factors (BEFs) for Scots pine, Norway spruce and birch according to stand age for boreal forests [J]. *Forest Ecology and Management*, **188** (2004):211-224.
12. Fatemi, F. R. 2007. *Aboveground biomass and nutrients in developing northern hardwood stands in New Hampshire*, *USA* [D]. College of Environmental Science and Forestry, State University of New York, USA.
13. Beauchamp, J. J.; Olson, J. S. 1973. Corrections for bias in regression estimates after logarithmic transformation [J]. *Ecology,* **54** (1973):1403-1407.
14. Zeng, W. S. Another discussion on selection of weight function in weighted least squares [J]. *Central South Forest Inventory and Planning*, **17** (1998):9-11 (in Chinese).
15. Zeng, W.S.; Luo, Q. B.; He, D. B. Research on weighting regression and modeling [J]. *Scientia Silvae Sinicae*, **35** (1999):5-11 (in Chinese).
16. Zhang, H.R.; Tang, S. Z.; Xu, H. On the heterosceasticity in biomass model [J]. *Forest Resources Management*, (1999): 46-49 (in Chinese).
17. Xu, H. A study on the heterosceasticity in tree biomass model [J]. *Journal of Northwest Forestry College*, **19** (1999):73-77 (in Chinese).
18. Zabek, L. M.; Prescott, C. E. Biomass equations and carbon content of aboveground leafless biomass of hybrid poplar in Coastal British Columbia [J]. *Forest Ecology and Management*, **223** (2006):291-302.
19. Huang, S. M. 2004. *Introduction to Econometrics* [M]. Peking University Press, China (in Chinese).