林业科学研究 2011,24(2):137～143
Forest Research

文章编号:1001-1498(2011)02-0137-07

非线性模型对数回归的偏差校正及与加权回归的对比分析*

曾伟生, 唐守正**

(中国林业科学研究院资源信息研究所, 北京 100091)

摘要:本文结合大样本的立木生物量实测数据,对非线性模型对数回归的偏差校正问题进行了探讨,并与加权回归结果进行了对比分析。首先,分析了对数回归产生偏差的内在原因,并提出了一个新的校正因子,同时对另外 3 个偏差校正因子一并进行了检验,结果表明本文和 Baskerville(1972)提出的校正因子,能保证与加权回归估计结果趋于一致;然后,对非线性加权回归中基于普通回归残差推导的权函数与通用权函数($W=1/f(x)^2$)的拟合效果进行了对比分析,结果表明二者基本相当,而通用权函数更具有广泛的适应性。建议对带有异方差的非线性模型,最好直接采用加权回归进行估计;当按照通用权函数进行估计其总相对误差超出一定范围时,应该根据普通回归估计的残差推导效果最佳的权函数后再进行加权回归。

关键词:非线性模型;生物量模型;对数回归;加权回归;偏差校正;异方差

中图分类号:S757.2 **文献标识码**:A

Bias Correction in Logarithmic Regression and Comparison with Weighted Regression for Non-linear Models

ZENG Wei-sheng, TANG Shou-zheng

(Research Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: Non-linear models with heteroscedasticity are commonly used in forestry modeling, and logarithmic regression and weighted regression are usually employed to estimate the parameters. Using the single-tree biomass data of large samples, the bias correction in logarithmic regression and comparison with weighted regression for non-linear models are studied in this paper. The immanent cause producing bias in logarithmic regression is analyzed, and a new correction factor is presented with which three commonly used bias correction factors are examined together, and the results show that the correction factors presented here and 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, are compared with each other that show two weights works well and the general function is more applicable. It is suggested that the best way to fit non-linear models with heteroscedasticity would be using weighted regression, and when the total relative error of the estimates from the model fitted by the general

收稿日期: 2010-04-26

基金项目: 林业数据分析技术及工具软件的完善与推广(05EFN216700395)和国家林业局专题"基于清查资料的中国森林植被生物量和碳储量评估"

作者简介: 曾伟生(1966—),湖南涟源人,在读博士,教授级高工,主要从事森林资源监测和林业数表研制工作.

*致谢:贵州省林业厅资源管理站夏忠胜博士为本研究提供了该省杉木、马尾松的立木生物量调查数据,在此深表谢意。

**通讯作者:E-mail:stang@caf.ac.cn

weight function is more than a special limit such as ±3%, a better weight function based on residual errors of the model fitted by OLS should be used in weighted regression.

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

在林业中用到的很多模型都是非线性模型,如材积方程和生物量方程。而且因为这些模型一般普遍存在着异方差性,因此,求解模型参数时通常采用对数转换或加权回归来消除异方差的影响[1-8]。

Finney 最早注意到采用对数转换产生有偏估计的问题,并提出了校正偏差的一个无偏估计量[1]。Baskerville 论证了对数转换后的估计值只是中位数无偏估计量,并提出了基于回归方程样本方差 s^2 的一个均值无偏估计量[2],其校正因子 $\exp(s^2/2)$ 后来得到了广泛应用[6,9-12]。Beauchamp 和 Olson 认为 Baskerville 提出的校正因子在实践中仍然是有偏的,因为样本方差 s^2 只是总体方差 σ^2 的无偏估计值,其真值是未知的,并提出了另外一个以 $\Psi(t)$ 函数表示的校正因子[13]。Flewelling 和 Pienaar 对提出的各类校正因子进行了综述,认为在中到大样本前提下(样本单元数一般在30以上),任何两个估计量之间的最大相对差异多数情况下(极端情况除外)为 $\exp(3/2s^2)$,如果不必太在意这些偏差的话,可以不校正或采用 Baskerville、Finney 或 Beauchamp 和 Olson 提出的校正因子进行校正;在小样本情况下,一般需要采用其它一些更复杂的估计量[3]。Snowdon 提出了一个比值估计量作为校正因子,并与应用最广的 Baskerville 和 Finney 提出的校正因子进行了比较,认为比值估计量简单实用且效果最好[4]。

除了采用对数转换以外,为了消除异方差对参数估计的影响,常采用加权回归估计。曾伟生[14-15]等在对材积方程、生物量方程的异方差性进行研究的基础上,提出了通用的权函数形式($W = 1/f(D)^2$);张会儒等、胥辉、Parresol 等都对生物量的异方差性进行了研究,也提出了各种权函数形式[5-6,16-17]。加权回归与对数回归之间到底有什么差异?对数回归适用于什么条件?其偏差校正的内在原因是什么?本文将结合对生物量实测数据的分析,对非线性模型的对数回归及其偏差校正问题再作探讨,并与加权回归结果进行对比分析,从而提出在实践中对非线性模型进行回归估计的结论和建议。

1 数据与方法

1.1 数据

本文所用数据为立木地上生物量实测数据,具体包括3个部分:第一部分数据来自1997全国生物量课题组南方片子项目,共计152株样木,采集地点在江西省德兴市李宅林场,包括杉木(*Cunninghamia lanceolata* (Lamb.) Hook.)、马尾松(*Pinus massoniana* Lamb.),以及栎类(*Quercus* spp.)、樟(*Cinnamomum* spp.)、楠(*Phoebe* spp.)等阔叶树种;第二部分数据来自2007年贵州省的生物量调查建模项目,包括杉木和马尾松两个树种组,共694株样木,其中杉木399株,马尾松295株,样本采集地点涉及贵州省杉木和马尾松分布区各县,对全省具有广泛的代表性。第三部分数据来自发表的文献[5-6],涉及两个树种:一是来自美国密西西比州的栎类鲜质量生物量数据,包括39株样木;二是来自美国路易斯安纳州的湿地松(*Pinus elliotii* Engelm.)鲜质量生物量数据,包括40株样木。本文的研究以前两部分数据为主(基本情况见表1),第三部分数据仅用于辅助分析。

表1 立木生物量数据的基本统计结果

数据源	统计量	平均值	最大值	最小值	标准差
南方片	胸径/cm	12.6	32.6	1.6	6.8
$n=152$	生物量/kg	67.424	541.163	0.429	87.013
贵州杉	胸径/cm	15.3	36.4	4.1	6.7
$n=399$	生物量/kg	71.646	529.992	1.329	78.118
贵州松	胸径/cm	16.0	44.8	4.0	8.1
$n=295$	生物量/kg	118.553	808.978	2.656	141.666

注:南方片数据指第一部分数据,贵州杉和贵州松指第二部分数据的杉木和马尾松数据;生物量指单株地上部分生物量。

1.2 方法

对于非线性模型的拟合,通常采用对数回归和加权回归方法,以消除异方差的影响。立木生物量模型最简单也最常用的结构为以下幂函数形式:

$$M = aD^b + \varepsilon \quad (1)$$

式中 M 为生物量,D 为胸径,a、b 为参数,ε 为误差。(1)式也有表示为以下形式的[5]:

$$M = aD^b\varepsilon \quad (2)$$

其中(1)式的误差称为加性(additive)误差,一般表示 ε 不随直径大小而变化,具有等方差性质;(2)式的误差称为乘积(multiplicative)误差,一般表示 ε 随直径大小而变化,具有异方差性质。实际上(2)式的表达方式是有问题的,因为如果 $E(\varepsilon)=0$,则大约有一半 M 可能出现负值。王仲锋已经提及这个问题[8]。作者认为,不管误差项 ε 的方差是否为等方差,一般都应该表示为(1)式;如果误差项 ε 的方差为异方差,但其相对误差 ε' 为等方差,则(1)式可以表述为如下形式:

$$M = aD^b(1+\varepsilon') \tag{3}$$

式中 $\varepsilon' = \varepsilon/aD^b$。将(3)式转换为对数形式后,变成:

$$\ln M = \ln a + b\ln D + \ln(1+\varepsilon') \tag{4}$$

式中 ln 为自然对数。(4)式看起来,似乎相当于以下标准形式的线性模型:

$$y = a_0 + b_0 x + \xi \tag{5}$$

式中 $y=\ln M, x=\ln D, a_0、b_0$ 为参数,ξ 为误差。采用普通最小二乘法拟合(5)式后,可按下式得到生物量的估计值:

$$M_{估} = \exp(a_0 + b_0\ln D) \tag{6}$$

但事实上这是错误的。因为普通最小二乘法要求误差 ξ 的数学期望为 0,即 $E(\xi)=0$,而实际上(4)式中与(5)式的 ξ 对应的 $\ln(1+\varepsilon')$ 的期望值不等于 0,即 $E[\ln(1+\varepsilon')]\neq 0$。下面来做具体分析。

由于(4)式中的 ε' 为相对误差,其理论分布区间为 $(-1,+\infty)$,现实分布范围一般在 ±0.5 之间。由于具有等方差性质,从而 $E(\varepsilon')=0$。因为 $\ln(1+\varepsilon')$ 恒小于 ε',所以,$E[\ln(1+\varepsilon')]\neq 0$,它应该等于某一个负数(假设为 $-c$)。而(5)式的最小二乘估计必须满足 $E(\xi)=0$。即,(4)、(5)两式之间的参数存在如下关系:

$$\xi = \ln(1+\varepsilon') + c \tag{7}$$

$$a_0 = \ln a - c \tag{8}$$

也就是说,经过对数形式转换后,为了满足线性最小二乘估计的条件 $E(\xi)=0$,从参数 a_0 中分离出了一部分常数 c 到误差项 ξ,从而使参数 a_0 系统偏小,这就是对数回归结果需要校正的内在原因。下面再来分析影响偏差大小的 c 值。

根据(7)式得到:

$$1+\varepsilon' = \exp(\xi - c) = \exp(-c)\exp(\xi) \tag{9}$$

再两边取数学期望,因为 $E(\varepsilon')=0$,从而 $\exp(-c)=1/E[\exp(\xi)]$,即:

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

也就是说,参数 c 的估计值近似等于 $\ln(1+s^2/2)$,s 为用普通最小二乘法拟合(5)式所得到的标准差估计值。从(8)式知 $\ln a$ 的无偏估计值应该为 a_0+c,从而校正后的(6)式应为:

$$\begin{aligned}M_{估} &= \exp(a_0 + c + b_0\ln D)\\ &= (1+s^2/2)\exp(a_0 + b_0\ln D)\end{aligned} \tag{10}$$

相当于其校正因子为 $1+s^2/2$。在这里,把它作为第一个校正因子:

$$CF_1 = 1 + s^2/2 \tag{11}$$

作为对比,同时还考虑以下 3 个校正因子:

$$CF_2 = \exp(s^2/2) \tag{12}$$

$$CF_3 = \exp\{s^2/2[1 - s^2(s^2+2)/4n + s^4(3s^4+44s^2+84)/96n^2]\} \tag{13}$$

$$CF_4 = \sum M / \sum M_{估} \tag{14}$$

其中 CF_2 是应用最多的一个校正因子,由 Baskerville 提出[2];CF_3 是 Finney 提出的校正因子 $g(s^2/2)$ 的近似表达式[1];CF_4 是 Snowdon 提出的比值校正因子[4]。

除了以上对数回归方法以外,可以对(1)或(3)式直接采用非线性回归进行估计。也考虑 3 种情况:一是普通最小二乘回归,相当于权函数为 1;二是按通用权函数($W=1/f(D)^2$)进行加权回归,它是相对误差为等方差时的最优权函数,能得到参数的无偏估计值[14-15];三是根据普通最小二乘回归结果的残差平方(e^2)拟合与 D 的回归关系($e^2=c_0D^{c_1}$,$c_0、c_1$ 为参数),再用 $W=1/D^{c_1}$ 作为权函数进行加权回归。

根据前述分析,在生物量数据的误差为异方差而相对误差为等方差的前提下,经过校正后的对数回归估计结果,应该与加权回归结果一样是无偏的。但是,由于上述校正因子本身只是一个近似值,且非线性模型的加权回归估计也是通过迭代算法得到的渐近估计值,故二者的估计结果实际上不可能完全相等。

为了比较分析对数回归和加权回归结果,本文采用以下 5 个评价指标:平均误差、平均绝对误差、总相对误差、平均系统误差和平均百分标准误差。

其计算公式如下[5-6,15,18]：

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (15)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (16)$$

$$TRE = \sum_{i=1}^{n}(y_i - \hat{y}_i)/\sum_{i=1}^{n}\hat{y}_i \times 100 \quad (17)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)/\hat{y}_i \times 100 \quad (18)$$

$$MPSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|/\hat{y}_i \times 100 \quad (19)$$

式中，n 为样本单元数，y_i 为生物量实测值(即 M)，$\hat{y}_i$ 为生物量预估值(即 $M_{估}$)。上述评价指标越小，表示模型预估的效果越好。

2 结果与分析

2.1 对数回归

利用第一、二部分立木生物量数据(以下简称南方片和贵州松、贵州杉数据)，将直径 D 和生物量 M 进行对数转换后，通过 Excel 的回归分析程序，采用普通最小二乘法拟合(5)式，然后根据拟合结果按照(11)～(14)式分别计算 4 个校正因子，其结果见表 2。

再根据(15)～(19)式计算对数回归模型(5)及其按 4 个校正因子校正后的模型的各种评价指标，结果见表 3。

表 2 立木地上生物量数据按对数回归的拟合结果

数据	样本量	参数估计值			统计指标		校正因子			
		a_0	$\exp(a_0)$	b_0	R^2	s^2	CF_1	CF_2	CF_3	CF_4
南方片	152	-2.199 06	0.110 907	2.343 23	0.943 64	0.130 67	1.065 34	1.067 52	1.067 49	1.090 02
贵州松	295	-2.405 94	0.090 181	2.438 18	0.939 84	0.111 46	1.055 73	1.057 31	1.057 29	1.037 75
贵州杉	399	-2.554 46	0.077 735	2.382 42	0.937 73	0.088 66	1.044 33	1.045 33	1.045 32	1.046 35

表 3 立木地上生物量数据对数回归模型的评价指标

数据	CF	ME	MAE	TRE/%	MSE/%	$MPSE$/%
南方片	CF_0	5.57	17.43	9.00	6.72	31.43
	CF_1	1.53	17.64	2.32	0.18	30.34
	CF_2	1.39	17.65	2.11	-0.03	30.31
	CF_3	1.39	17.65	2.11	-0.03	30.31
	CF_4	0.00	17.79	0.00	-2.09	30.09
贵州松	CF_0	4.31	24.85	3.78	5.69	26.32
	CF_1	-2.05	25.31	-1.70	0.11	25.10
	CF_2	-2.23	25.33	-1.85	-0.04	25.07
	CF_3	-2.23	25.33	-1.85	-0.04	25.07
	CF_4	0.00	25.08	0.00	1.84	25.40
贵州杉	CF_0	3.17	14.11	4.63	4.06	21.33
	CF_1	0.14	13.70	0.19	-0.35	20.34
	CF_2	0.07	13.70	0.10	-0.45	20.33
	CF_3	0.07	13.70	0.10	-0.45	20.33
	CF_4	0.00	13.70	0.00	-0.55	20.31

注：校正因子 $CF_0 = 1$ 指未校正的回归模型。

从表 3 可以看出，未校正的对数回归模型确实存在不可忽略的偏差，其中用南方片数据拟合的模型，其总相对误差 TRE 达到 9.00%，平均系统误差 MSE 达到 6.72%；用贵州杉木和马尾松数据拟合的模型，其总相对误差也分别达到 4.63% 和 3.78%，平均系统误差分别达到 4.06% 和 5.69%。校正因子 CF_2 和 CF_3 近似相等，且与 CF_1 差异也不大，因此用其校正后的模型，各项评价指标几乎没有差异，其中平均系统误差 MSE 均接近于 0(对于相对误差为等方差的模型，该指标理论上应该等于 0)，总相对误差 TRE 在 ±2% 左右，其中样本量最大的贵州杉木生物量模型，TRE 仅为 0.10%。而用 CF_4 校正的模型，其出发点是保证平均误差 ME 和总相对误差 TRE 为 0，但平均系统误差 MSE 却达到 ±2% 左右，其它 2 项指标则差异不大。

2.2　加权回归

对模型(1)式直接按非线性回归进行估计,采用 Marquardt 迭代算法。为了对比,首先按普通最小二乘法进行拟合,然后按前述 2 种权函数用加权最小二乘法进行拟合,结果见表 4。

表 4　立木地上生物量数据按加权回归的拟合结果

数据	样本量	权函数	参数估计值		统计指标	
		W	a	b	R^2	s^2
南方片	1 152	1	0.065 024	2.547 63	0.882 60	888.86
		$1/f(D)^2$	0.121 848	2.330 95	0.873 98	954.17
		$1/D^{4.57}$	0.119 794	2.338 07	0.875 05	946.04
贵州松	295	1	0.210 042	2.188 25	0.905 35	1 906.12
		$1/f(D)^2$	0.102 807	2.409 40	0.896 54	2 083.56
		$1/D^{3.24}$	0.106 643	2.397 77	0.897 57	2 062.73
贵州杉	399	1	0.050 534	2.535 77	0.918 20	500.45
		$1/f(D)^2$	0.081 085	2.381 52	0.914 29	524.36
		$1/D^{4.61}$	0.081 125	2.381 33	0.914 27	524.47

注:权函数 $W=1$ 指普通最小二乘回归估计,表 5 亦同。

再根据(15)～(19)式计算非线性回归模型(1)式及其加权回归模型的各种评价指标,结果见表 5。

表 5　立木地上生物量数据加权回归模型的评价指标

数据	W	*ME*	*MAE*	*TRE*/%	*MSE*/%	*MPSE*/%
南方片	1	1.17	17.20	1.77	13.37	34.62
	$1/f(D)^2$	1.87	17.64	2.85	0.00	30.29
	$1/D^{4.57}$	1.62	17.65	2.45	0.02	30.31
贵州松	1	−2.71	25.83	−2.23	−11.69	27.04
	$1/f(D)^2$	−0.36	25.01	−0.30	0.00	25.02
	$1/D^{3.24}$	−0.34	24.98	−0.29	−0.59	24.95
贵州杉	1	0.79	13.56	1.12	7.38	22.72
	$1/f(D)^2$	0.42	13.72	0.58	0.00	20.40
	$1/D^{4.61}$	0.43	13.72	0.60	0.01	20.40

从表 5 可以看出,由于非线性回归估计是用泰勒级数的线性展开式近似表示非线性模型来求解参数,因此其拟合结果并不像线性模型一样,能保证平均误差 *ME* 和总相对误差 *TRE* 为 0。由于普通回归估计没有消除异方差的影响,从而存在明显的系统误差,其中南方片数据拟合的普通回归模型,其平均系统误差 *MSE* 达到 13.37%,最小的贵州杉木模型也达到 7.38%。而按通用权函数与根据普通回归结果的方差拟合的权函数进行回归估计,其结果都相差不大,其中平均系统误差 *MSE* 都等于 0 或接近于 0,总相对误差 *TRE* 贵州杉木和马尾松模型也都在 ±1% 以内,仅南方片的模型有点偏大,达到近 3%,原因可能是涉及的树种多而对建模结果造成影响。另外,根据与表 3 的对比不难看出,按通用权函数的加权回归估计与按(11)～(13)式校正后的对数回归估计结果非常接近,与预期的效果完全一致。

3　讨论

非线性模型按对数转换为线性形式后,采用普通最小二乘法进行估计,其结果会存在偏差,这一点已经达成广泛共识。对偏差的校正问题,有很多学者做过研究,并提出过不少校正因子,其中 Baskerville 提出的校正因子 $\exp(s^2/2)$ 至今仍得到应用[2]。Snowdon 对该校正因子的实用性曾提出质疑,并建议用一个简便实用的比值估计量作为校正因子[4]。本研究结合 3 个大样本的立木生物量数据,对常用的校正因子进行了对比分析,结果表明按(12)、(13)式及本文提出的(11)式进行校正后的模型与加权回归的结果非常接近;Snowdon 提出的比值校正因子也是可行的,但他完全是从另一角度提出的。对于对数回归的偏差校正,建议采用基于对数转换特点提出的校正因子,从而保持与加权回归估计之

间的一致性。

关于加权回归估计,根据 Parresol 的综述文章[5],国外一般都是按照普通最小二乘估计的残差来推导权函数。从表 4 和表 5 的结果看,由残差推导的权函数和通用权函数的加权回归估计结果没有显著差异。笔者还利用 Parresol 发表的栎类和湿地松鲜质量数据[5-6],对两种权函数的拟合结果也进行了对比分析,结果见表 6。

表 6 通用权函数与残差推导权函数的拟合结果对比

数据	模型	按 Parresol 推导的权函数拟合					按通用权函数拟合				
		b_0	b_1	b_2	R^2	*SEE*	b_0	b_1	b_2	R^2	*SEE*
栎类 $n=39$	干材	25.749 48	0.027 31		0.98	182.32	26.244	0.027 307		0.98	182.31
	干皮	-0.515 32	0.102 53		0.94	41.01	-0.356 61	0.102 41		0.94	41.01
	树冠	117.195	0.057 502	-4.616 87	0.81	21.15	104.68	0.056 644	-4.081 6	0.80	21.27
	总量	46.381	0.031 558		0.98	217.37	52.653	0.031 380		0.98	214.68
湿地松 $n=40$	干材	0.016 363	1.058 5		0.98	27.44	0.016 366	1.058 5		0.98	27.44
	干皮	0.046 277	2.209 3		0.96	5.11	0.044 298	2.224 6		0.96	5.16
	树冠	0.027 378	3.680 4	-1.262 4	0.89	14.22	0.033 816	3.653 6	-1.313 9	0.89	14.42

注:*SEE* 指估计值的标准误,即表 2 和表 4 中的 s;模型表达式见 Parresol 的文章[5-6];按 Parresol 推导的权函数拟合结果,除湿地松的 R^2、*SEE* 数据是根据模型参数推算的以外,其它数据均直接引自 Parresol 的文章[5-6]。

从表 6 的结果不难看出,按 Parresol 推导权函数的拟合结果[5-6]与按通用权函数的拟合结果差异很小。众所周知,根据普通最小二乘估计的残差推导效果最佳的权函数不是一件容易的事,而且随着变量的增加和模型复杂化程度的提高,构建残差与自变量之间回归关系的难度也将增大。如 Parresol 推导的栎类树冠模型的权函数就非常复杂,其表达式为[5]:

$$W = (D^2H * LCL/1\ 000)^{1.646} * \exp(-0.004\ 06H^2)$$

式中 *LCL* 为活树冠长度。根据模型自身构建的通用权函数($W=1/f(x)^2$)简单实用,当相对误差为等方差时在理论上也是最优的,应该成为各类加权回归估计推荐使用的权函数。

从求解模型参数的准则分析,普通回归估计会保证模型的总相对误差及平均误差为 0(非线性回归因为是通过迭代算法近似求解,所以不完全为 0;线性模型则等于 0),而加权回归估计(在采用通用权函数的情况下)会保证模型的平均系统误差为 0,这一点可以参见表 5 的结果。对于理想的建模样本或异方差不明显的建模样本,总相对误差和平均系统误差都应该趋向于 0,即普通回归和加权回归的结果趋于一致[15]。如利用 Parresol 发表的栎类鲜质量数据[5],其普通回归和加权回归模型的总相对误差及平均系统误差都在 ±0.5% 以内,平均百分标准误差都在 10% 以内。由于建模样本通常难以达到理想状态,普通回归和加权回归的结果一般存在较大差异。如果按通用权函数进行加权回归得到的总相对误差比较大(如超过 ±3%),则应该根据普通最小二乘估计的残差推导效果最佳的权函数后再进行加权回归。

需要补充指出的一点是,作为生物量方程或材积方程等通用性模型,对模型的评价指标一般不能仅针对整体而言,还要考虑模型在整个自变量范围内的预估效果,因此,通常还需要分径阶进行检验和评价,在此不详述。

4 结论

本文结合大样本的立木生物量实测数据,对非线性模型的对数回归及其偏差校正问题进行了探讨,揭示了存在偏差的内在原因,并与加权回归结果进行了比较分析。可以得到以下结论:

(1)对于存在异方差的非线性模型,可以按对数转换为线性形式后,采用普通线性回归进行估计,但必需对模型存在的偏差进行校正;建议采用本文或 Baskerville 提出的校正因子,因为它是基于对数转换的特点而提出的,校正后的模型与加权回归估计结果是趋于一致的。

(2)在加权回归估计中,权函数的确定是非常重要的环节。通过对国内外常用的由残差推导的权函数与曾伟生提出的通用权函数所做的对比分析,表明通用权函数具有广泛的适应性,而且对于相对误差为等方差的模型,通用权函数在理论上也是最优的;尤其对于模型形式复杂的情况,通用权函数会更加显现出其优越性。

(3)对数回归和加权回归都能有效消除异方差的影响,在对数回归模型得到有效校正后,二者的预

估效果几乎等同。因此,对于能通过对数转换进行线性化的非线性模型,既可采用对数回归方法,也可直接采用加权回归方法。由于非线性回归方法已经得到普遍应用,对于存在异方差的非线性模型,建议直接采用加权回归估计。如果按通用权函数进行加权回归估计,其总相对误差超出了一定范围(如 ± 3%),则应该根据普通回归估计的残差推导效果最佳的权函数后,再进行加权回归。

参考文献:

[1] Finney D J. On the distribution of a variate whose logarithm is normally distributed [J]. J R Statist Soc, Suppl. 1941,7:155 - 161

[2] Baskerville G L. Use of logarithmic regression in the estimation of plant biomass [J]. Can J For Res, 1972,2:49 - 53

[3] Flewelling J W, Pienaar L V. Multiplicative regression with lognormal errors [J]. For Sci, 1981, 27(2):281 - 289

[4] Snowdon P. A ratio estimator for bias correction in logarithmic regressions [J]. Can J For Res, 1991, 21(5): 720 - 724

[5] Parresol B R. Assessing tree and stand biomass: a review with examples and, critical comparisons [J]. For Sci, 1999, 45(4): 573 - 593

[6] Parresol B R. Additivity of nonlinear biomass equations [J]. Can J For Res, 2001,31: 865 - 878

[7] 曾伟生,骆期邦. 论林业数表模型的研建方法[J]. 中南林业调查规划, 2001,20(2):1 - 4

[8] 王仲锋. 森林生物量建模与精度分析[D]. 北京: 北京林业大学, 2006

[9] Wiant Jr H V, Harner E J. Percent bias and standard error in logarithmic regression [J]. For Sci, 1979,25(1):167 - 168

[10] Sprugel D G. Correcting for bias in log - transformed allometric equations [J]. Ecology, 1983,64(1):209 - 210

[11] Lehtonen A, Mäkipää R, Heikkinen J, *et al*. Biomass expansion factors (BEFs) for Scots pine, Norway spruce and birch according to stand age for boreal forests [J]. Forest Ecology and Management, 2004,188:211 - 224

[12] Fatemi F R. Aboveground biomass and nutrients in developing northern hardwood stands in New Hampshire, USA [D]. USA: College of Environmental Science and Forestry, State University of New York, 2007

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

[14] 曾伟生. 再论加权最小二乘法中权函数的选择[J]. 中南林业调查规划, 1998,17(3):9 - 11

[15] 曾伟生,骆期邦,贺东北. 论加权回归与建模[J]. 林业科学, 1999,35(5):5 - 11

[16] 张会儒,唐守正,胥　辉. 关于生物量模型中的异方差问题[J]. 林业资源管理, 1999(1):46 - 49

[17] 胥　辉. 生物量模型方差非齐性研究[J]. 西北林学院学报, 1999,19(2):73 - 77

[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, 2006, 223: 291 - 302