密度—直径关系研究中变量变换形式的选择和变量误差矩阵的估计*

王琫瑜　　李希菲

摘要　在林分自稀疏过程研究中，利用度量误差模型的形式估计林分断面积平均直径 Dg 与单位面积株数 N 关系模型的参数时，需要确定 Dg 和 N 的抽样误差 e、ϵ 方差的比值。本文利用 20 块人工落叶松标地的材料，经过再抽样，计算并比较了各种数据变换方式下，断面积平均直径与单位面积株数的方差比。结果表明 $\ln(N)$ 和 Dg 的方差比接近于常数，其均值为 $3.024\,8\times10^{-2}$。由此，可确定变量的变换形式和构造误差矩阵，从而为精确估计模型参数提供依据。

关键词　自稀疏、度量误差模型、抽样误差

应用度量误差模型的形式估计林分因子间关系的模型参数，既可解决变量间相互预报的问题，又可提高模型参数的估计精度。本文研究直径 D 和株数 N 关系建模时，如何构造模型的误差结构，并确定二元随机误差的误差结构矩阵。

1　引　　言

应用数学模型描述林分中各变量之间的关系已成为林业研究工作的重要方法。用通常方法建立数学模型时，总是把变量分为因变量(预报变量)和自变量(说明变量)。在实际问题中，有些模型的预报变量和说明变量是不可以互换的，例如：用气象因子来预报农作物产量。在有些模型中，各因子是要互相预报的，例如：完满立木度林分中，直径 D_f 和密度 N_f 的关系[1]

$$\ln \alpha' = \beta \ln(D_f) + \ln(N_f) \tag{1}$$

或林分优势高 H_d 和平均高 H 的关系[2]

$$H_d = a + bH$$

对这种问题，当自变量和因变量位置互换后，用通常最小二乘法估出的回归方程不是同一条直线。另一方面，通常最小二乘法的模型是将误差加在因变量上，因此当自变量和因变量都存在较大量测误差时，直接用通常最小二乘法来估计模型参数，并不是一个好方法。Fuller[3] 系统地介绍了度量误差模型，这个模型与 Kendall[4] 介绍的函数关系与结构关系模型或 Cramer[5] 介绍的正交回归模型本质上是一致的。唐守正[2] 提出的对偶回归也属于度量误差模型。利用度量误差模型来建立林分因子间关系的模型既可以解决变量之间相互预报的问题，又可以提高模型参数的估计精度。现在用这个观点来考察模型(1)的参数估计问题。假定完满立木度林分直

1995—12—12 收稿。

王琫瑜助理研究员，李希菲(中国林业科学研究院资源信息研究所　北京　100091)。

* 本文是 1992 年自然科学基金重点项目‘我国主要人工林生长模型、经营模型和优化控制’部分内容。唐守正研究员给予指导，王雪峰同志给予帮助，特此致谢。

径 D_f 和 N_f 满足关系式(1), 为了估计参数 α和 β, 设置 m 个标准地, 其观测值为 n_i 和 $d_i(i=1, \cdots, m)$ 。由于抽样或量测误差, (n_i, d_i) 与其真值 N_i 和 D_i 不同。

目前, 构造模型的误差结构时, 一般有加法或乘法结构两种, 加模型认为:

$$N_i = n_i + \epsilon_i$$
$$D_i = d_i + e_i$$

乘模型认为:

$$N_i = n_i \times \epsilon_i' \text{或} \ln(N_i) = \ln(n_i) + \epsilon_i$$
$$D_i = d_i \times e_i' \text{或} \ln(D_i) = \ln(d_i) + e_i$$

对于模型(1) 的误差结构, 可能出现 4 种组合的模型:

模型 1∶D 和 N 都是加结构, 这时, 方程(1) 变为

$$\ln \alpha = \beta \ln(d_i + e_i) + \ln(n_i + \epsilon_i) \tag{2}$$

模型 2: D 和 N 都是乘结构, 方程(1) 变为

$$\ln \alpha = \beta \ln(d_i) + e_i + \ln(n_i) + \epsilon \tag{3}$$

模型 3: D 是加结构, N 是乘结构, 方程(1) 变为

$$\ln\alpha = \beta \ln(d_i + e_i) + \ln(n_i) + \epsilon_i \tag{4}$$

模型 4: D 是乘结构, N 是加结构, 方程(1) 变为

$$\ln\alpha = \beta \ln(d_i) + e_i + \ln(n_i + \epsilon_i) \tag{5}$$

在上述 4 种组合中, 如果 d_i 的度量没有误差, 即 $e_i = 0$, 这时方程(2) 和方程(5) 变成以 N 为因变量, D 为自变量的非线性模型。方程(3) 和方程(4) 变成以 N 为因变量, D 为自变量的对数线性化模型, 反之若 $\epsilon_i = 0$, 则方程(1) 和(3) 变成以 N 为自变量, D 为因变量的非线性模型, 方程(2)、(4) 变成以 D 为因变量, N 为自变量的对数线性化模型。本文的第一个目的就是通过对抽样误差的分析来确定, 当研究 D 和 N 的关系时, 究竟应当用哪一种模型。

其次, 当确定模型类型后, 由[2] 知, 还必须知道, σ_e^2 和 σ_ϵ^2 的比值, 才能使用度量误差模型来估计参数 α 和 β。由 Fuller[3] 知, 对于更一般的情况, 需要知道二元随机误差(e, ϵ) 的误差结构矩阵, 所谓误差结构矩阵 Φ是指矩阵:

$$\Phi = \begin{pmatrix} \varphi_1^2 & r\varphi_1\varphi_2 \\ r\varphi_1\varphi_2 & \varphi_2^2 \end{pmatrix}$$

它与(e, ϵ) 的协方差阵 Σ 差一个未知倍数 k, 即 $\Sigma = K\Phi$, 本文的第二目的是求出由抽样误差造成的 e(即 D 或 $\ln D$) 和 ϵ(即 N 或 $\ln N$) 的误差结构矩阵的数值。本文的结果很容易推广到多元情况, 例如[1] 中提出的一个估计自稀疏方程参数 S_f (最大密度指数)、β(自稀疏率)、γ(自稀疏指数) 的方程:

$$\ln N_2 = \ln S_f - \ln[(D_2/D_0)^{\beta\gamma} + (S_f/N_1)^{\gamma} - (D_1/D_0)^{\beta\gamma}]/\gamma \tag{6}$$

其中 D_1、D_2、N_1、N_2 为同一林分前期和后期直径和密度。

2　数据采集和处理

2. 1　数据

原始数据采用吉林省汪清林业局的 20 块人工落叶松(*Larix olgensis* Henry) 标地, 在选择

标地时取没有经过很大人为破坏的、面积不小于 1 350 m^2 的标地, 标地均有多次测树资料, 1992、1994 年按研究要求复测, 重绘了样地树木位置图。各样地的基本情况见表 1。

2. 2 抽样调查

为了计算变量方差比并分析变量方差比是否与抽样面积大小有关, 需要大量不同面积的样地数据, 为此设计了从 100～600 m^2 的 6 个不同面积大小的抽样方案, 对上述的 20 块标准地进行室内抽样调查, 抽样原则是: (1) 每级方案的小样地均覆盖原始样地的全部面积。(2) 每级抽样方案中任意两块相邻的小样地的重复部分均不超过其本身面积的 50%。按照以上抽样原则和方案由计算机进行抽样统计调查, 共计得到 1 565 块小样地调查数据(数据略)。

表 **1** 原始样地基本情况

样地编号	样地横向边长(m)	样地纵向边长(m)	株数	断面积平均直径(cm)
101	50	50	209	21. 233 6
102	50	50	208	20. 178 1
103	50	40	103	17. 617 1
204	100	25	164	17. 010 2
208	100	25	211	16. 129 5
209	25	100	180	16. 969 7
210	30	80	131	19. 540 4
211	25	100	112	19. 694 0
212	100	25	159	16. 434 5
213	100	25	153	16. 112 5
305	40	50	214	13. 630 8
306	50	40	306	16. 849 9
307	40	50	220	18. 111 9
308	50	40	129	19. 534 9
309	50	50	293	16. 782 6
310	50	50	198	19. 591 6
311	50	50	195	20. 478 5
312	50	50	219	18. 797 4
313	45	45	144	18. 193 0
314	45	45	167	17. 973 4

2. 3 数据处理

将所获小样地单位面积株数 N 和断面积平均直径 Dg 的数据进行自然对数变换, 并分别计算每块标准地每级抽样方案下断面积平均直径 Dg 和株数 N 测量值的方差及它们的自然对数的方差, 共有 120 组数据, 然后再分别计算变量方差比 S^2_{Dg}/S^2_N、S^2_N/S^2_{Dg}、$S^2_{\ln(Dg)}/S^2_{\ln(N)}$、$S^2_{\ln(N)}/S^2_{\ln(Dg)}$、$S^2_{Dg}/S^2_{\ln(N)}$、$S^2_{\ln(N)}/S^2_{Dg}$、$S^2_{\ln(Dg)}/S^2_N$、$S^2_N/S^2_{\ln(Dg)}$ 的值。

3 数据分析

3. 1 变量方差比值分析

从各组变量方差比的数据中可以看出, 各方差比基本上都趋近于各自不同的一个常数。求出各组方差比值序列的变动系数和均值, 见表 2。从表 2 中可以看出在各组方差比值中, 以 $S^2_{\ln(N)}/S^2_{Dg}$、$S^2_{\ln(N)}/S^2_{\ln(Dg)}$、$S^2_{Dg}/S^2_{\ln(N)}$、$S^2_{\ln(Dg)}/S^2_{\ln(N)}$ 的值的变动系数较小。

表 **2** 各组方差比值的变动系数和均值

方差比值	S^2_{Dg}/S^2_N	S^2_N/S^2_{Dg}	$S^2_{\ln(Dg)}/S^2_{\ln(N)}$	$S^2_{\ln(N)}/S^2_{\ln(Dg)}$	$S^2_{Dg}/S^2_{\ln(N)}$	$S^2_{\ln(N)}/S^2_{Dg}$	$S^2_{\ln(Dg)}/S^2_N$	$S^2_N/S^2_{\ln(Dg)}$
样地数	1 565	1 565	1 565	1 565	1 565	1 565	1 565	1 565
变动系数	1. 036 4	1. 268 9	0. 981 1	0. 784 2	0. 979 0	0. 750 7	1. 151 9	1. 138 8
均 值	1.13×10^{-4}	1.07×10^{-9}	0. 183 4	10. 090 4	59. 466 4	0. 030 248	3.55×10^{-7}	6.55×10^{-6}

3. 2 变量方差比与抽样面积关系分析

计算不同的抽样方案下各方差比数据序列的均值和变动系数, 对各组变量方差比与抽样面积进行方差分析, 结果见表 3。从表 3 中可见, 各组方差比在不同抽样方案下的变动系数和均值都较为稳定, 但方差分析的结果显示 S^2_{Dg}/S^2_N 的值在不同抽样方案下的差异较为明显, 而其余各组方差比在不同的抽样方案下的差异不明显。即可以认为 S^2_N/S^2_{Dg}、$S^2_{\ln(Dg)}/S^2_{\ln(N)}$、$S^2_{\ln(N)}/$

$S^2_{\ln(Dg)}$、$S^2_{Dg}/S^2_{\ln(N)}$、$S^2_{\ln(N)}/S^2_{Dg}$、$S^2_{\ln(Dg)}/S^2_N$ 和 $S^2_N/S^2_{\ln(Dg)}$ 的值与抽样面积无关。

表 **3**　变量方差比与抽样面积关系的方差分析

方差比	抽样方案	100 m^2	200 m^2	300 m^2	400 m^2	500 m^2	600 m^2	$F_{0.05}$	F 值及显著性
	样地数	646	302	181	166	143	127		
S^2_{Dg}/S^2_N	变动系数	0.606 3	0.695 0	0.993 7	0.892 5	1.099 3	0.931 1	2.29	2.576*
	均　值	7.10×10^{-5}	8.10×10^{-5}	9.00×10^{-5}	1.19×10^{-4}	1.35×10^{-4}	1.80×10^{-4}		
S^2_N/S^2_{Dg}	变动系数	0.774 1	0.702 5	1.550 2	0.817 2	0.842 1	1.248 6	2.29	1.313
	均　值	2.11×10^{-4}	1.99×10^{-4}	3.27×10^{-4}	1.70×10^{-4}	1.61×10^{-4}	1.46×10^{-4}		
$S^2_{\ln(Dg)}/S^2_{\ln(N)}$	变动系数	0.631 4	0.858 0	1.006 7	1.047 2	1.038 6	0.968 8	2.29	0.571
	均　值	0.150 7	0.162 5	0.164 6	0.194 3	0.193 1	0.235 1		
$S^2_{\ln(N)}/S^2_{\ln(Dg)}$	变动系数	0.552 6	0.758 9	0.789 4	0.791 4	0.824 4	0.774 4	2.29	0.932
	均　值	9.317 3	10.827 6	12.992 5	9.851 1	9.694 1	7.860 0		
$S^2_{Dg}/S^2_{\ln(N)}$	变动系数	0.658 6	0.699 8	0.854 9	0.937 4	0.977 2	0.936 5	2.29	2.106
	均　值	39.082 7	45.758 9	50.580 7	63.590 3	67.366 3	90.419 6		
$S^2_{\ln(N)}/S^2_{Dg}$	变动系数	0.571 6	0.679 9	0.785 0	0.722 3	0.715 6	0.755 2	2.29	2.264
	均　值	3.72×10^{-2}	3.53×10^{-2}	3.78×10^{-2}	2.62×10^{-2}	2.50×10^{-2}	2.00×10^{-2}		
$S^2_{\ln(Dg)}/S^2_N$	变动系数	0.658 4	0.768 6	1.162 6	1.056 8	1.493 4	1.024 4	2.29	0.649
	均　值	2.94×10^{-7}	2.89×10^{-7}	2.96×10^{-7}	3.67×10^{-7}	4.21×10^{-7}	4.36×10^{-7}		
$S^2_N/S^2_{\ln(Dg)}$	变动系数	0.907 0	0.760 0	1.357 2	0.725 0	0.708 3	1.039 7	2.29	1.329
	均　值	5.73×10^{-6}	6.21×10^{-6}	1.04×10^{-7}	6.01×10^{-6}	5.78×10^{-6}	5.19×10^{-6}		

4　结　论

(1) 综合表 2 和表 3 知, $\ln(N)$ 和 Dg 的方差比值与样地面积无关, $\ln(N)$ 和 $\ln(Dg)$ 的方差比值及 N 和 $\ln(Dg)$ 的方差比值也与样地面积无关, 但是 N 与 $\ln(Dg)$ 的方差比值的变动系数显著偏大, 而 $\ln(N)$ 的方差与 Dg 的方差比值又稍小于 $\ln(N)$ 的方差与 $\ln(Dg)$ 的方差比值, 这个事实说明 N 的误差是乘结构, 而 Dg 的误差为加结构。因而模型 3(即公式 4) 是四种模型中最好的模型。

(2) 由表 3 知 $S^2_{\ln(N)}/S^2_{Dg}$ 平均等于 $3.024\,8\times10^{-2}$, 进一步计算知 Dg 和 $\ln N$ 误差的相关系数为 0.621 7, 这样可得 (e, ϵ) 即 Dg 和 $\ln N$ 的度量误差结构矩阵如下:

$$\varphi = \begin{pmatrix} 1.0 & 0.108\,1 \\ 0.108\,1 & 3.024\,8\times10^{-2} \end{pmatrix}$$

参　考　文　献

1　唐守正. 同龄纯林自然稀疏规律的研究. 林业科学, 1993, 29(3): 234～241.

2　唐守正. 利用对偶回归和结构关系建立林分优势高和平均高模型. 林业科学研究, 1991, 4(增刊): 57～62.

3　Fuller W A. Measurement error models. New York: John Wiley and Sons, 1987. 1～440.

4　Kendall M G, Stecart A. The advanced theory of statistics. New York: Hafner Pulishing Com., 1973, 2: 1～758.

5　Cramer H. Mathematical method of statistics. Princeton: Princeton University Press, 1946. 1～416.

The Choice of Variables Transformation and Estimate of Measurement Error Matrix for the Study on Density-diameter Relationship

Wang Fengyu　Li Xifei

Abstract　It is necessary to determine the ratio between variance of sampling error e of quadratic average diameter Dg and that of sampling error ϵ of the tree numbers N per unit area when measurement error models is applied to estimate the parameters of a model between Dg and N, in a study on stand self-thinning processes. This paper calculated and compared the variance ratios between transformd quadratic average diameter and transformd number of trees per unit area by using resampling from 20 larch plantation plots. The results show that the ratio between variance of $\ln(N)$ and Dg approaches to a constant $3.024\,8\times10^{-2}$. From the results, the best form of variable transformation is determined and the structure of error matrix is constructed, which can be used to get more accurate estimate of model parameters.

Key words　self-thinning, measurement error model, sampling error

Wang Fengyu, Assistant Professor, Li Xifei(The Research Institute of Forest Resources Information and Technique, CAF　Beijing　100091).