林业科学研究　1997, 10(2) : 215～219
Forest Research

用哑变量法求算立地指数曲线族的研究*

李希菲　　洪玲霞

关键词　立地指数曲线　哑变量方法　导向曲线法

立地质量是表示生长在某一立地上既定树种或林分类型林木的生产潜力[1]。立地质量不同, 林木的经营措施也不同。描述立地质量的方法很多, 现在多采用地位指数, 即以造林年数和优势木平均高为基础的地位指数方程[1]。导向曲线法[1,2]是建立地位指数方程最简单常用的方法, 而在实际应用中发现导向曲线法估计斜率时过分依赖于样本的分布关系。为此介绍一种哑变量方法, 可以取得较好的结果。

1　引言

导向曲线法适用于研制合成地位指数方程, 一般可用临时样地测定, 提供成对的各样地优势木平均高及年龄数据, 求得一条导向曲线。这条导向曲线对于原始数据来说是一条平均线, 因此导向曲线的解与样地分布关系很密切。当采用导向曲线法时, 必须要求样地均匀分布在不同年龄、不同立地上, 这在实际上是很难做到的。当所取样地的年龄、立地分布不均匀时, 必然导致导向曲线的斜率偏高或偏低。如图 1-a, 表示当多数小年龄样地立地较差, 而多数大年龄的样地立地较好时, 导向曲线斜率偏大。反之, 图 1-b 说明当多数小年龄样地立地偏好, 而多数大年龄样地立地较差时, 导向曲线斜率偏小。而哑变量方法可以解决求导向曲线斜率时, 过分依赖样地的年龄和立地分布问题。

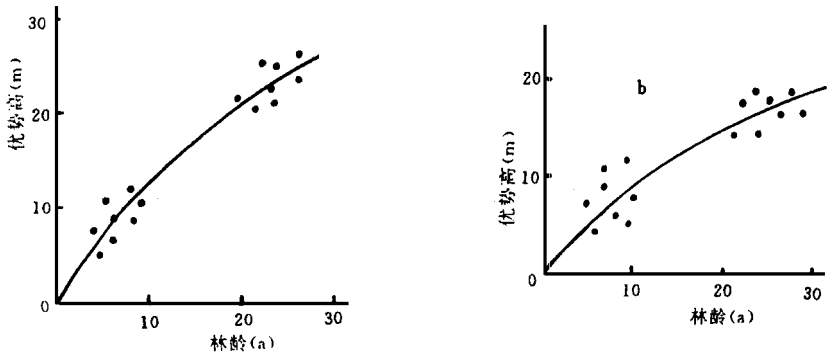

图 1　样地分布与导向曲线斜率关系

1996—10—16 收稿。
李希菲副研究员, 洪玲霞(中国林业科学研究院资源信息研究所　北京　100091)。
* 本文属 1992 年国家自然科学基金项目“我国主要人工林生长模型、经营模型和优化控制”部分研究内容。唐守正院士给予指导, 并提供计算机程序, 特此致谢。

本文以常见的舒马克曲线[2,3]: $H = k\ \exp(-b/A)$ (1.1)
(其中 H 为林分优势高, A 为林分年龄) 为例, 说明如何利用哑变量方法来确定曲线的斜率 b 值, 同时确定样地的立地指数。取 20 a 为立地指数的基准年龄, 因而样地的立地指数为:

$$L = k\ \exp(-b/20)\ 。$$

将上式代入(1.1) 式, 得到一致性立地指数方程, 即在基准年龄时, 树高年龄曲线值等于优势高值[4]: $H = L\ \exp[b(-1/A + 1/20)]$ (1.2)
取对数, 上式变为线性方程式: $y = a + bx$ (1.3)
其中 $y = \ln H$, $a = \ln L$; $x = 1/20 - 1/A$ (1.4)

如果测定了 p 个样地, 其中第 i 个样地在 n_i 个时间点上量测了林分年龄 A_{ij} 和优势高 H_{ij} ($j = 1\cdots, n_i$)。把时间和优势高按(1.4) 变换成 y_{ij} 和 X_{ij}, 设第 i 个样地的立地指数是 L_i(未知), 令 $a_i = \ln L_i$, 假定全部样地有统一的 b 值, 把这些值代入(1.3) 式, 得到模型:

$$y_{ij} = a_i + bx_{ij} + \epsilon_{ij} \quad (i = 1, \cdots, p, j = 1, \cdots, n_i) \tag{1.5}$$

其中 ϵ_{ij} 表示观测值对方程的误差, a_i 和 b 是待求参数。

模型(1.5) 是一个用于求斜率 b 和各样地立地指数 $L_i = \exp a_i$ 的线性模型。用最小二乘法解(1.5) 的最常用方法是哑变量方法。

2 哑变量方法简介

在《多元统计分析方法》[5] 的“一元线性模型”一章中, 讲述了线性模型, 例如模型(1.5) 的一般求解方法。(1.5) 也可看成自变量带有定量变量的数量化方法 I[6] 的问题, 即把 a_i 看成一个项目(立地) 的 p 个类目(样地) 的得分值。但是在英文文献中, 多将此方法称为哑变量(dummy variable) 方法[7]。哑变量方法很自然地解释了这个模型, 并把它化成多元回归问题。

引入 p 个哑变量 $Z_1 \cdots Z_p$, 它们是样地号的函数, 定义为:

$$Z_k(i,j) = \begin{cases} 1 & \text{当 } i = k \text{ 时} \\ 0 & \text{当 } i \neq k \text{ 时} \end{cases} \quad k = 1, \cdots, p \tag{2.1}$$

这样, (1.5) 式可以写成多元回归的形式:

$$y_{ij} = a_1 Z_1(i,j) + \cdots + a_p Z_p(i,j) + bx_{ij} + \epsilon_{ij} \quad (i = 1, \cdots, p; j = 1, \cdots, n_i) \tag{2.2}$$

采用通常最小二乘法可同时算出 $a_1, \cdots, a_p$ 和 b 值, 即同时得到各样地的立地指数 $L_i = \exp(a_i)$ 和优势高生长过程曲线(1.2)。

由此看出, 对于线性方程, 哑变量方法就是数量化方法 I[6], 但哑变量方法很容易推广到非线性方程的求解上去。例如: 直接用非线性方程(1.2) 求解各样地的立地指数 L_i 及统一的斜率 b 值, 同样可引入哑变量(2.1), 把观测值及哑变量代入(1.2) 式, 得到非线性的模型

$$H_{ij} = [L_1 Z_1(i,j) + \cdots + L_p Z_p(i,j)]\exp[b(1/20 - 1/A_{ij})] + \epsilon_{ij}$$
$$(i = 1, \cdots, p, j = 1, \cdots, n_i) \tag{2.3}$$

采用任何一种非线性回归方程解法求解(2.3) 式的待估参数 L_i(第 i 样地的立地指数) 和 b 值。

在相关系数很大时, 两种解法相差很小, 本文采用线性方程(2.2) 来估计参数。

3 哑变量方法与导向曲线法比较

为了说明用哑变量方法计算的立地指数曲线族优于导向曲线法, 用一简化的例子, 从下述

两方面比较: 表 1 列出实验地区抽出的 8 个杉木固定样地观测资料, 每样地复测 3～4 次, 表 2 列出这 8 个样地按年龄和立地的分布情况。通过以下分析, 可看出样地的年龄、立地分布对结果的影响。

表 **1**　部分杉木样地资料

样地号	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	5
年　龄	22	24	28	8	12	14	8	10	12	25	26	28	10	11	12	14
优势高	12.6	13.6	14.9	5.5	6.6	7.2	11.7	13.7	14.3	22.2	22.3	22.6	15.1	16.0	16.9	17.0

(续表 **1**)

样地号	6	6	6	7	7	7	8	8	8
年　龄	22	26	27	8	9	12	25	26	29
优势高	13.4	14.4	14.8	5.3	6.7	8.2	11.6	12.2	13.5

表 **2**　按年龄、立地分布的样地号

立　地	小年龄	大年龄
差	2, 6, 7	1, 8
好	3, 5	4

3.1　用刀切法比较 b 值(斜率) 的稳定性

先将 8 个样地数据, 按公式(1.4) 将 A , H 变换成 x , y, 解哑变量回归方程(2.2) 得斜率 b 的估计值为 - 6.301。再逐个删去一个样地, 用其余 7 个样地的 x , y 解(2, 2) 式, 得到 8 个不同的 b_i 值, 例如 b_1 = - 6.243 表示删去 1 号样地后得到的估计值, 全部结果列入表 3, 这就是刀切法[8]。

表 **3**　b 值的刀切法计算结果

方　法	$-b$	$-b_1$	$-b_2$	$-b_3$	$-b_4$	$-b_5$	$-b_6$	$-b_7$	$-b_8$	标准差
哑变量	6.301	6.243	7.153	6.697	6.306	6.560	6.248	5.191	6.214	0.560
导向曲线	7.513	8.103	5.924	9.455	5.754	9.034	7.821	5.192	8.604	1.635

同样, 用通常的导向曲线法, 分别全部样地并依次删除一个样地, 估计斜率 b 值。其结果也列入表 3, 最后一列是 "刀切 "的 8 个 b 值的标准差, 它表示由于抽取样地不同可能对 b 的估计值造成的影响。由表 3 可见, 导向曲线法的标准差约为哑变量方法的 3 倍, 说明哑变量方法的估计值要比导向曲线法的估计值稳定得多。

由立地指数方程(1.2) 可看出, 对于舒马克立地指数方程, 是由 b 值决定方程的形状, 因而, b 的估计值对样地抽取依赖较小的方法, 就是一种较好的方法。

3.2　用相对残差平方和比较

若 $\hat{H}$ 是估计树高, H 是实测树高, 则 $\ln\hat{H} - \ln H = \dfrac{\hat{H} - H}{H}$, 即对数值的残差近似等于相对残差, 因而用对数形式线性化的回归估计残差平方和近似为相对误差的残差平方和。由于树高越大, 估计值的绝对误差越大, 因此用相对误差来度量立地曲线的精度可能更好, 这就是本文用对数线性形式来估计 b 值的主要原因。

用哑变量方法求出 b 值的同时, 也求出了各样地的立地指数 L_i (或其对数 a_i) (见表 4), 因而得到各样地立地指数方程的对数形式: $\hat{y}_{ij} = a_i + b(1/20 - 1/A_j)$　　(3.1)

其中 $\hat{y}_{ij} = \ln(H_i(A_j))$；表示第 i 样地在 A_j 年的估计树高的对数。

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(\hat{y}_{ij} - y_{ij})^2 \qquad (3.2)$$

其中 y_{ij} 为第 i 样地在 A_j 年的实测树高，Q 表示立地曲线族相对误差的残差平方和。用表 1 所列 8 个样地及 $b = -6.301$ 算出 $Q = 0.035$。

用导向曲线法算出了 $b = -7.513$，还需再估计各样地的立地指数，才能得到立地曲线族。由优势高和立地的关系(1.2)式，在求出 b 值后，可由每个观测点 A_{ij} 和 H_{ij}，反求出该点的立地指数 L_{ij}，取某样地全部观测点立地指数 L_{ij} 的平均值为该样地的立地指数 L_i，再取 $a_i = \ln L_i$，按(3.1)及(3.2)式算出立地曲线族的相对误差平方和 $Q = 0.0408$。表 4 列出了上述计算结果，可看出哑变量法与导向曲线法所求立地指数相差无几，一般都相差在 1 m 之内，但导向曲线法的残差平方和约为哑变量法的 1.2 倍，说明哑变量法优于导向曲线法。

表 **4** 两种方法残差平方和比较

样地号	哑变量法		导向曲线法	
	L_i	a_i	L_i	a_i
1	12.46	2.522	12.3	2.511
2	8.48	2.138	8.9	2.192
3	18.39	2.912	19.6	2.977
4	20.75	3.032	20.4	3.018
5	20.42	3.016	21.4	3.061
6	13.35	2.591	13.2	2.580
7	9.46	2.247	10.1	2.317
8	11.47	2.441	11.3	2.427
残差平方和 Q	0.0350		0.0408	

3.3 理论比较

导向曲线法的缺点是将各不相关的样地放在一起作回归，造成不同立地的样地在各年龄阶上所占比例对曲线的斜率产生很大影响。哑变量方法的优点在于哑变量代表了各样地的立地指数。其 b 值本质上是各样地单独 b 值的某种平均，因而由其算出的 b 值较稳定。哑变量方法的缺点是要求各样地有重复观测的资料，这给用临时性样地配立地指数曲线带来了困难。可用树干解析材料来补充，但树干解析的内插点不宜应用过多。建议只应用 2 个内插点，因为解析木内插点的树高是线性内插而得，造成树高生长曲线接近一个直线，这种误差会对哑变量法造成较大影响。

利用哑变量方法还可以进一步检查哪些样地是属于同一立地曲线族的(即检查 b 值是否相同)[9]。利用此法还可进行立地指数曲线族的分类，将另文专述。

4 部分杉木产区 b 值的哑变量法估计

现有南方 4 省 7 个地区杉木固定样地 221 块，详见表 5。这些地区按 b 值可以分成 5 类，即 5 个立地曲线族，每一族有相近的 b 值，用哑变量方法求得的各 b 值见表 6。

表 **5** 杉木固定样地基本情况

地 点	样地数	观测次数	年龄 (a)	立地指数 (m)
广西武宣	31	3～7	7～20	11～18
广西凭祥	35	2～3	6～36	8～16
湖南株洲	48	3～5	5～55	10～22
湖南会同	20	3～5	6～16	11～18
江西分宜	55	4～8	4～18	12～18
福建南平等	15	2～4	6～28	10～22
福建邵武等	19	2～4	6～28	14～22
合 计	221	–	–	–

表 **6** 立地曲线族分类

类	地区名	b 值	b 的误差限
1	江西分宜 湖南株洲	6.3747	0.0651
2	广西凭祥 湖南会同	5.7909	0.4201
3	广西武宣	7.6767	0.9094
4	福建南平等	10.6309	0.7171
5	福建邵武等	4.3148	0.8563

5　结论与讨论

用刀切法比较立地曲线族斜率 b 值的稳定性, 可以看出导向曲线法的标准差约为哑变量方法标准差的 3 倍。从相对残差平方和来比较, 导向曲线法的残差平方和约为哑变量方法的 1. 2 倍, 说明哑变量方法的斜率估计值比导向曲线法稳定。这是因为导向曲线法是将互不相关的样地放在一起作回归, 会产生过分依赖样地年龄和立地分布, 从而使 b 值缺乏稳定性。但哑变量方法要求重复观测资料, 也是难于普遍应用的原因。

从现有 4 省 7 地区固定样地观测资料应用哑变量方法求算 b 值结果看, b 值的范围为 4. 314 8～10. 630 9, 共分成了 5 类, 造成 b 值差异的原因尚需进一步研究。

参　考　文　献

1　克拉特 J L, 弗尔森 J C, 皮纳尔 L V, 等. (范济洲, 关玉秀, 于政中, 等译). 用材林经理学—— 定量方法. 北京: 中国林业出版社, 1990. 44～57.

2　南方十四省(区) 杉木栽培科研协作组. 全国杉木(实生林) 地位指数表的编制与应用. 林业科学, 1982, 18(3): 266～277.

3　胡希 B., 米勒 C. L., 比尔斯 T. W. 测树学. 北京: 农业出版社, 1979. 1～354.

4　骆期邦, 吴志德, 蒋菊生, 等. Richards 函数拟合多形地位指数曲线模型的研究. 林业科学研究, 1989, 2(6): 534～539.

5　唐守正. 多元统计分析方法. 北京: 中国林业出版社, 1986. 191～230.

6　董文泉, 周光亚, 夏立显. 数量化理论及其应用. 长春: 吉林人民出版社, 1979. 1～48.

7　Johnston J. Econometric Methods (2nd Edition). New York: McGraw-Hill Book Company, 1972. 176～192.

8　Efron B. (黄承明译). 计算机与统计理论: 思考不可思议的事. 应用数学与计算数学, 1980, (5): 52～79.

9　唐守正, 李希菲. 同龄林自稀疏方程的验证. 林业科学, 1995, 31(1): 27～33.

Research on the Use of Dummy Variables Method to Calculate the Family of Site Index Curves

Li Xifei　　Hong Lingxia

Abstract　Using dummy variables method to calculate the family of the site index curves is introduced in this paper. The method is able to settle the problem that the slop of guide curve is too dependent on site-age distribution of sampling plots when guide curve method is employed. The variations of jackknife's slops which are got from dummy variables method and guide curve method are compared. The result shows that the standard deviation of guide curve method is as three times as that of dummy variables method. The result proves that dummy variables method is better than guide curve method.

Key words　site index curves　dummy variables method　guide curve method

Li Xifei, Associate Professor, Hong Lingxia (The Research Institute of Forest Resource Information Techniques, CAF Beijing　100091).