



Statistical Models and Applications

统计模型和应用研究

宋连莲 刘帆/著



科学出版社

(C-0901. 31)



Statistical Models and Applications

统计模型和应用研究



科学出版社互联网入口
经管分社: (010)64012800 销售: (010)64031535
E-mail: jingguanfa@mail.sciencep.com

www.sciencep.com

ISBN 978-7-03-052328-0



9 787030 523280 >

定价: 102.00元

Statistical Models and Applications

统计模型和应用研究

宋连连 刘帆/著



Statistical Models and Applications

统计模型和应用研究

宋连莲 刘 帆 著

科 学 出 版 社

北 京

内 容 简 介

本书针对常用的 probit 模型、MNL 模型、survival 等离散选择模型和贝叶斯、多选择等动态模型进行了系统的描述和分析,并在此基础上,构建出新型的拓展模型,解决新的实际问题。本书融合统计学、经济学和行为学三门学科的基础知识,论述统计模型在消费者购买研究、客户关系管理、电视节目观看行为研究、广告收视率预测研究、网页优化研究、客户信用评估等领域的应用价值。

本书适用于统计学、经济学和管理科学等学科教学科研人员及学生,也适合大量参与创新、研发活动的企业决策者、实施者及咨询机构人员等阅读。本书的诸多实证研究与结论,以及所提出的拓展研究模型,可以在企业高层管理者面对复杂多变的市场需求时,提出较为精准的预测方案,帮助其理解和分析并制定相应的策略。

图书在版编目(CIP)数据

统计模型和应用研究 = Statistical Models and Applications/宋连莲,刘帆著.
—北京:科学出版社,2017.8

ISBN 978-7-03-052328-0

I. ①统… II. ①宋… ②刘… III. ①统计模型—研究—英文
IV. ①C815

中国版本图书馆 CIP 数据核字 (2017) 第 052824 号

责任编辑:方小丽 李 莉/责任校对:彭珍珍

责任印制:吴兆东/封面设计:无极书装

科 学 出 版 社 出 版

北京东黄城根北街 16 号

邮政编码:100717

<http://www.sciencep.com>

北京京华虎彩印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2017 年 8 月第 一 版 开本: B5 (720×1000)

2017 年 8 月第一次印刷 印张: 17

字数: 340 000

定价: 102.00 元

(如有印装质量问题,我社负责调换)

This book is financially supported by NSFC (No. 71602089), Natural and Social Science
Foundation of Jiangsu Province (No. BK20160785, 16GLC001).

Contents

1	General Statistical Models	1
1.1	Motivation	1
1.2	Properties of Discrete Choice Models	3
1.3	Multinomial Logit Model	5
1.4	Nested Logit Model	7
1.5	Probit Models	9
1.6	Mixed Logits	11
1.7	Negative Binomial Distribution	12
2	Customer-base Analysis	19
2.1	Introduction	19
2.2	Research Background	22
2.3	Week Effect on Purchasing Pattern	27
2.4	The Mixture Gamma Heterogeneity	47
2.5	Conclusions	66
3	Identifying Future Defaulters: A Hierarchical Bayesian Method	69
3.1	Introduction	69
3.2	Model Specification	72
3.3	Extension	73
3.4	Estimation	74
3.5	Empirical Analysis	78
3.6	Discussions and Conclusions	86
4	Modeling Audience Choice Behavior on Television Programs	89
4.1	Introduction	89
4.2	Literature Related to the Dynamic Multiple Programs Adoption Model	91
4.3	A Dynamic Multiple Programs Adoption Model	94
4.4	Model Construction	98
4.5	Empirical Observations	110
4.6	Estimation Results	111
4.7	Conclusions and Limitations	116

5	Measuring Television Commercial Performance	119
5.1	Television Audience Behavior towards Advertisements	119
5.2	Literature Review	121
5.3	Problem Introduction	124
5.4	Data Analysis for Model Building	128
5.5	Methodology and a Hypothetical Example	132
5.6	Empirical Examples for Model Application	150
5.7	Conclusions and Limitations	157
6	Modeling the Audience's Dynamic Decision Procedure	161
6.1	Introduction	161
6.2	Data	163
6.3	Model Development	166
6.4	Estimation Discussion	171
6.5	Managerial Implications	175
6.6	Conclusions and Future Research Directions	178
7	Click Behavior and Web Design	180
7.1	Introduction	180
7.2	Web Design Problem Illustration	181
7.3	Related Literature	183
7.4	Modeling Methods in the Use of Time Analysis	184
7.5	The Web-log File and Data Preparation	187
7.6	Use of Time Models for Web-log Data	188
7.7	Applications to the Design of the Home Page of a Website	192
7.8	Conclusions	197
8	Multiple Choices of Web Links When Searching for Information	199
8.1	Problem Description	199
8.2	Research Background	201
8.3	Model Construction	203
8.4	The Clickstream Data Preparation	205
8.5	Empirical Study	207
8.6	Practical Implications	209
8.7	Conclusions	212
8.8	Limitations and Future Researches	213
9	Modeling Waiting Time of Online Application on Education Website	215
9.1	Introduction and Literature Review	215

9.2	Empirical Observation	217
9.3	Survival Model Analysis	219
9.4	Basic Model Selection	222
9.5	Proportional Hazard Model	226
9.6	Conclusions	230
10	Measuring the Entrepreneurship Performance via Improved Global Entrepreneurship Index	232
10.1	Introduction	232
10.2	Study Background	233
10.3	Problem Formulation	235
10.4	Stochastic Multicriteria Acceptability Analysis	239
10.5	Empirical Analysis	240
10.6	Conclusions	244
	References	246
	Appendix	259

1 General Statistical Models

1.1 Motivation

In the mid-1980s, Train wrote the book *Qualitative Choice Analysis*. In this book, he introduced the basic models—logit and nested logit models in detail, including the model properties and applications in many areas. However, the first-generation choice models suffered many limitations which inhabited them in resolving real problems. More than 20 years have passed, in 2003, Train wrote the second book, *Discrete Choice Models with Simulation*, for stating the tremendous progress in the approaches and methods of choice analysis. Simulation had been the center of the progress, since it supplied a numerical approximation to integrals. Till today, more than 10 years have passed since 2003, the famous discrete choice models have been used in a wide range of our life, including marketing, housing, transportation and energy. However, a new generation of problems has emerged, which brings forth a new burden on the researchers.

The problems or situations that need modeling and explaining have become more complex and full of high technology. For example, marketing problems are not restricted to the physical shops, the purchase behavior has transferred onto the Internet. Original discrete choice models may not satisfy the requirements of the new issues. We summarize the new generation of problems as two categories. One is that the modeling methods themselves have become more complicated than the traditional discrete choice models. To fit the real situations, new modeling methods are developed when none of the existing models seem to be correct. People's behavior is changeable in a short period of time. Sometimes, we combine two different models into a new one to fit the disparate purchasing behavior and apply a new theory to a particular problem even though that theory has never been used in similar situations. The other is for the estimation method to implement the new models. When developing new models and employing a new theory, the relevant validations are necessary. Researchers need to be able to develop very complex probabilities and the corresponding likelihood function, and program the procedures into computer softwares. That is a very difficult process, since researchers often need to change their traditional computational perspectives,

capture the up-to-date technics of coding and finally achieve the aims of parameter estimation and model managerial implications.

The purpose of this book is to extend the model construction and application of the traditional discrete choice models using the empirical examples. From the improved model development to the innovative parameter estimation procedure, we illustrate all steps in very detail. In particular, the major contribution of this book is to deal with the new problems emerged in today's complex situation through proposing innovative models. Not only do the proposed models have advantages in describing and explaining the choice behavior, but they also have advantages in conducting new managerial implications for managers.

Specifically, we follow the traditional steps of applying statistical models in marketing. Firstly, we discuss the customer-base analysis and answer such questions as which customers are more likely to stop purchasing from the firm? What level of transactions should be expected from the customer base? How do the customers respond to the marketing activities? Secondly, we extend the traditional methods of applying classification models into the area of credit scoring: 1) relaxing the assumption of the independence between the probability and the time of default; 2) treating the missing defaulting labels as latent variables and applying an augmentation technique; 3) introducing a discrete truncated exponential distribution to model the time of default. An empirical analysis is also conducted to show the performance of our proposed model. Thirdly, we extend the model application in the area of television audience viewing behavior, including television program and advertisement (ad) choice behavior: 1) modeling the dynamic multiple choice behavior of television audience to predict and improve program audience ratings; 2) modeling commercial viewing patterns to convert minute information into second information to evaluate advertisement effectiveness. We use three chapters to discuss the topics suggested in those two areas. The common aim of those two topics is to construct statistical models that examine television audience viewing behavior in order to increase audience ratings of target programs or advertisements (ads). The differences between them can be described from two aspects. On the one hand, topic 1 targets audience viewing behavior to programs. Whereas, topic 2 develops a model for the viewing behavior towards television advertisements. On the other hand, topic 2 develops a methodology to measure audience ratings. Whereas, topic 1 predicts audience ratings using a more sophisticated model. Fourthly, the fast development of web technologies has revolutionized the way we spend our time. The number of web users is increasing every year. The statistical

models are absolutely being used in modeling the web browsing behavior. Those topics are described by two chapters in this book. Fifthly, the entrepreneurship area is also discussed in this book by Chapter 10. We develop a new model which combines the statistical theory into the stochastic multicriteria acceptability analysis to build an improved global entrepreneurship index (GEI), which we term it the holistic acceptability index. That method differs from the conventional wisdom that assigns exact values to corresponding weights, but explores the weight space to allow each country to specify a preference. Sixthly, our method is confirmed using an empirical study measuring the entrepreneurship of the top 10 countries in terms of 2016 GEI.

1.2 Properties of Discrete Choice Models

1.2.1 Background

Discrete choice model (DCM) is a modeling method of choice from a set of alternatives. That set of alternatives is called the choice set. The set must be mutually exclusive and collectively exhaustive and the number of alternatives is finite. The number of alternatives is too large for the respondent to consider, McFadden has suggested a sampling method to select a sample of alternative for respondents to make choices. An assumption is suggested that a utility function can be found to describe a decision maker's preference when he is confronted with a set of alternatives. Utility, as a measure of well-being, has no natural level or scale. While, neo-classical economists comment that a utility function can allow the decision maker to rank the alternatives in a consistent and unambiguous manner. Then another assumption of utility maximizing behavior by the decision maker is proposed: the choice is a deterministic process.

However, psychologists present that it is impossible to specify and construct a discrete choice model that will always succeed in predicting the chosen alternative by the decision maker. Accordingly, Thurstone (1927) proposed a utility function that allows the decision maker to rank the alternatives, but there are fluctuations inherent in the process of evaluating alternatives. As noted by Tversky (1972), "when faced with a choice among several alternatives, people often experience uncertainty and inconsistency. That is, people are often not sure which alternative they should select, nor do they always take the same choice under seemingly identical conditions." Thus, the choice is an outcome of a probabilistic process.

1.2.2 Random Utility

Two sources are for the randomness of the utility function. One is the inability of the analysts to formulate individual behavior, since they lack information of individuals' characteristics and information of alternatives' attributes. The other is the stochastic behavior of the decision maker, the circumstances under the decisions can "perturb" the perception and the desirability of alternatives. A random utility function is necessary to model the random utility.

The random utility function is constructed as this:

$$U = \text{deterministic component } (V) + \text{random component } (\varepsilon)$$

where V presents observable attributes of alternatives and characteristics of individuals; ε captures the effects of stochastic behavior of individuals and unobserved attributes and characteristics.

$$U_{ij} = V_{ij}(x_{ij1}, x_{ij2}, \dots, x_{ijk}) + \varepsilon_{ij}$$

That is the utility function for individual i and alternative j , where x_{ijk} denotes the observed characteristics of individual i and the observed k th attributes of alternative j . Let X_{ij} be a vector of x_{ijk} , that is, $X'_{ij} = (1, x_{ij1}, x_{ij2}, \dots, x_{ijk})$.

That X -vector might include simple attributes (e.g., income, price per trip, etc.) or explicit interactions of the attributes of the alternatives and the individuals (e.g., price, income).

If we can assume a linear function for the effects of the factors, the random utility becomes:

$$U_{ij} = X'_{ij}\beta + \varepsilon_{ij} \quad \text{where } \beta' = (\beta_0, \beta_1, \dots, \beta_k) \quad (1.1)$$

β_0 is the intercept term representing the intrinsic preference of individual i towards alternative j ; β_k is the response coefficient of the k th attribute; U_{ij} is the utility of the j th alternative to the i th individual.

Equation (1.1) is under these two assumptions: 1) the vector β reflects the response to the attribute of the alternative, which is assumed to be identical for all the individuals in the population; 2) random variation in the U_{ij} is introduced through the additive disturbance term ε_{ij} which is assumed to be independently and identically distributed across individuals and alternatives for multinomial logit (MNL), but the error terms could be dependent on each other for other models.

1.3 Multinomial Logit Model

1.3.1 General Multinomial Logit Model

The random components of the utilities of different alternatives are IID (independent and identically distributed) according to Gumbel distribution (extreme value distribution).

Derivation of multinomial logit is shown as below:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \quad \text{where } \varepsilon_{ij} \sim \text{Gumbel}(\eta, \mu)$$

Since ε is Gumbel distributed with parameters (η, μ) , then:

$$F(\varepsilon) = \exp[-e^{-\mu(\varepsilon-\eta)}] \text{ and } f(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} \exp[-e^{-\mu(\varepsilon-\eta)}], \quad \mu > 0, \quad -\infty < \varepsilon < \infty$$

Without loss of generality, set $\eta = 0$. Then:

$$U_1 = V_1 + \varepsilon_1 \sim \text{Gumbel}(V_1, \mu)$$

$$U_2 = V_2 + \varepsilon_2 \sim \text{Gumbel}(V_2, \mu)$$

...

$$U_j = V_j + \varepsilon_j \sim \text{Gumbel}(V_j, \mu)$$

The probability that alternative 1 is chosen as below:

$$P_i(1) = \Pr[U_1 \geq U_2, U_1 \geq U_3, \dots, U_1 \geq U_j]$$

$$= \int_{-\infty}^{\infty} f(U_1) \int_{-\infty}^{U_1} f(U_2) \int_{-\infty}^{U_1} f(U_3) \dots \int_{-\infty}^{U_1} f(U_j) dU_1 dU_2 dU_3 \dots dU_j = \frac{e^{\mu V_1}}{e^{\mu V_1} + e^{\mu V_2} + \dots + e^{\mu V_j}}$$

In the general multinomial logit model, the coefficients reflect the responses (tastes) of individuals in the population to the attributes of alternatives.

1.3.2 Random-coefficients Multinomial Logit Model

In the coefficients in the general multinomial logit model, β reflects the responses (tastes) of individuals in the population to the attributes of alternatives. A constant β means that all individuals' tastes are identical with respect to the observed attributes of alternatives embodied in the X_{ij} vectors. Consequently, the logit model formulation implies that all individuals of identical observed characteristics have identical tastes with respect to the observed attributes of alternative. The problems are emerging: a constant β may not be the case in reality; the variation in response to different attributes (unobserved heterogeneity) will influence individual's choice behavior; if such heterogeneity is ignored, the parameter estimates of the logit model will have a

downward bias. In order to resolve the problems, a random-coefficient specification is used to deal with the unobserved heterogeneity. The parameters of the logit model are treated as realisation of random variables representing the individuals' preferences and their responses to the marketing variables.

The random utility can be written as below:

$$U_{ij} = X'_{ij}\beta + \varepsilon_{ij} \text{ where } \beta \text{ is a random vector independent of } \varepsilon_{ij}$$

For a particular individual i , the vector $\Theta_i = \{\beta_{0i1}, \beta_{0i2}, \dots, \beta_{0iJ}, \beta_{i1}, \beta_{i2}, \dots, \beta_{iK}\}$ is assumed to be a realisation from a continuous multivariate distribution $G(\Theta)$. Conditional on the value of the vector Θ from a continuous distribution, the probability of a randomly drawn individual choosing alternative j on occasion t will be given by:

$$P_t(j|\Theta) = \frac{\exp\{\beta_{0j} + \sum_{k=1}^K \beta_k x_{jkt}\}}{\sum_{l=1}^J \exp\{\beta_{0l} + \sum_{k=1}^K \beta_k x_{lkt}\}}$$

$$\text{Therefore, } P_t(j) = \int P_t(j|\Theta) dG(\Theta).$$

It is difficult to specify the underlying distribution of heterogeneity, parameters usually involve evaluating difficult multiple integrals when a continuous multivariate distribution is used. Thus, maximum likelihood estimation (MLE) is selected to estimate the model parameters. The likelihood function for individual i conditional on the value of the vector Θ_s is given by:

$$L_i(\cdot | \Theta_s) = \prod_{t=1}^{T_i} \left\{ \prod_{j=1}^J [P_{it}(j | \Theta_s)]^{\delta_{ijt}} \right\}$$

where δ_{ijt} is 1 if the i th individual chooses alternative j on occasion t and it is 0 otherwise. J is the total number of alternatives and T_i denotes the number of occasions made by the i th individual. Therefore, the overall sample likelihood function is given by:

$$L = \prod_{i=1}^I \left[\sum_{s=1}^S \left[\prod_{t=1}^{T_i} \left[\prod_{j=1}^J [P_{it}(j | \Theta_s)]^{\delta_{ijt}} \right] \right] \alpha(\Theta_s) \right]$$

The number of support point S can be determined by using a stopping rule procedure based on the Bayesian information criterion (BIC). $BIC = -LL + \frac{1}{2} * R * \log(N)$, where LL is log-likelihood; R represents the number of parameters estimates; N denotes the total number of observation in the sample.