Categorical Data Analysis

Syllabus

Deyuan Li
School of Management
Fudan University

Fall 2015

Self-introduction

- Deyuan Li (黎德元), PhD in Statistics
- Education and Work Background
  - 1992-2000: Peking University, China
  - 2001-2004: Erasmus University Rotterdam, the Netherland
  - 2005-2007: University of Bern, Switzerland
  - 2008-2013: Associate Professor, Fudan University, China
  - 2014- : Professor, Fudan University, China
  - Research: Extreme Value Theory (极值统计)
    for more details, visit
    http://my.gl.fudan.edu.cn/teacherhome/lideyuan/

Self-Introduction

- Hobbies
  - Travel
  - Photography
  - Badminton

- Contact information

Office:	Room 736, Siyuan Building
Phone:	021-25011216
E-mail:	deyuanli@fudan.edu.cn
Office hours:	13:30-14:30, Monday

Course Objectives and Requirements

Objectives: Introduce several statistical methods to analyze categorical data. It covers contingency tables, generalized linear models, logistic linear models, logit Models, loglinear models, models for matched pairs and generalized linear mixed models.

Requirements:

- Strong background of probability, mathematical statistics and statistical computing software SAS (or R, C).
- Homework will be issued several times and students are required to independently finish it (NOT copy from others).
- Slides, exercises and exams will be given in English, but answers can be given in English or in Chinese.

Schedule

Week	Contents
1:	Distributions and Inference for Categorical Data
2:	Describing Contingency Tables
3:	Inference for Contingency Tables
5-6:	Introduction to Generalized Linear Models
7:	Logistic Regression
8:	Building and Applying Logistic Regression Models
9:	Midterm Exam (Nov. 5)
10:	Logit Models for Multivariate Responses
11:	Loglinear Models for Contingency Tables
12:	Models for Matched Pairs
13:	Analyzing Repeated Categorical Response Data
14-15:	Random Effects: Generalized Linear Mixed Models
16:	Presentation and Review
17-18:	Final Exam

15次课（13次讲课，1次期中考试，1次报告与总结）

References

Textbook: *Categorical Data Analysis*, second edition, by Alan Agresti, 2002, Wiely.

website for the related material:
http://www.stat.ufl.edu/∼aa/cda/cda.html

Three references are highly recommended:

1. 齐亚强，《分类数据分析》，重庆大学出版社，2012。
2. *An Introduction to Categorical Data Analysis*, second edition, by Alan Agresti, Wiley, 2007.
3. 张淑梅等，《属性数据分析引论》，高等教育出版社，2008。

Grading

The final grade of this course consists of

Attendance:	5%;
(Team) Homework:	15%;
Midterm Exam:	30%;
Final Exam:	50%.

Four comments to all of the students:

1. No calculation, no answer.
2. Need to spend 2 hours for each lecture.
3. On time, no food.
4. Shut down mobile phones.

Any Question?

Example 1. 上市企业特别处理ST

特别处理（special treatment, ST）：上市公司出现财务状况或其他状况异常，交易所对该公司股票交易实行特别处理。包括

- 限制该股票日涨跌幅度（如±5%）；
- 该股票名称前冠以ST;
- 公司中期报告必须审计；
- 若持续亏损，有退市风险。

目的：保护投资人的利益。

问题：如何从财务报表预测公司将来是否被ST?

Example 1. 上市企业特别处理ST

数据:

- $Y = 1$（ST公司）；$Y = 0$（非ST公司）
- 财务指标:
  - $X_1 = ARA$：应收账款与总资产的比例
  - $X_2 = ASSET$：对数变换后的资产规模
  - $X_3 = ATO$：资产周转率
  - $X_4 = ROA$：资产收益率
  - $X_5 = GROWTH$：销售收入增长率
  - $X_6 = LEV$：债务资产比例
  - $X_7 = SHARE$：第一大股东持有比例
- 响应变量（Y），解释变量（$X_1, X_2, ..., X_7$）

模型:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_7 X_7 + \varepsilon \quad ?$$

Example 1. 上市企业特别处理ST

新思路：对$P(Y=1)$建模，如：

$$\log \frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_7 X_7$$

$\Rightarrow$ Logistic regression （逻辑回归）。

$Y=0$或1：名义数据（nominal data）

其他应用：移动通讯客户流失等。

Example 2. 消费者偏好度

手机品牌：苹果、三星、索尼、华为、小米、锤子等

- 市场竞争激烈，产品周期短；
- 手机功能、体积、重量、应用、价格等都很关键，决定了产品的市场；
- 新产品立项初期会做消费者调查研究，以获得对消费者偏好（preference）的准确判断，进而指导新产品的研发，减小市场风险。
- 问卷调查⇒市场偏好：
  5 =非常喜欢；4 =喜欢；3 =无所谓；
  2 =不喜欢；1 =极不喜欢。

模型？

Example 2. 消费者偏好度

模型：

$$\log \frac{P(Y \leq j)}{P(Y > j)} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \quad j = 1, 2, ..., 4.$$

$Y = 1, 2, ..., J$：有序数据（ordinal data）

其他应用：信用卡用户信用评级等。

Example 3. 付费搜索广告

付费搜索广告的优点：价格便宜；广告效果可追踪；门槛低。

付费搜索广告的发展3个阶段：

- 按广告主出价高低排名 — 早期Overture和百度的广告系统
- 结合出价和点击率决定广告排名 — Google的Panama系统
- 全局的进一步优化 — Google最新算法

问题：如何估计一广告（链接）在一段时间内的点击次量。

Example 3.付费搜索广告

数据：

- 因变量：$Y = 0, 1, 2, \ldots$（点击量）
- 解释变量：
  - X_1：关键词长度
  - X_2：展现量
  - X_3：平均点击价格
  - X_4：平均排名

模型：$Y \sim \text{Poisson}(\lambda)$ and

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

$\Rightarrow$ Poisson Regression （泊松回归）或者Loglinear （对数线性模型）

$Y = 0, 1, 2, \ldots$：计数数据（count data）

其他应用：超市消费者到访频数等。