
广义线性模型

张振虎 草稿

张振虎

2021 年 02 月 26 日

目录

1 前言	1
2 概率基础	3
2.1 概率模型	3
2.1.1 概率律	3
2.1.2 离散模型	4
2.1.3 连续模型	5
2.2 条件概率	5
2.3 联合概率	7
2.4 全概率与贝叶斯定理	8
2.5 独立性	10
2.6 随机变量	11
2.6.1 离散随机变量	12
2.6.2 连续随机变量	12
2.6.3 累积分布函数	13
2.6.4 随机变量的函数	14
2.6.5 期望与方差	15
2.7 边缘化	17
2.8 常见概率分布	18
2.8.1 伯努利分布	18
2.8.2 二项式分布	19
2.8.3 类别分布	21
2.8.4 多项式分布	21
2.8.5 高斯分布	22
2.8.6 卡方分布	26
2.8.7 t 分布	28
2.8.8 F 分布	29
3 最大似然估计	33
3.1 最大似然估计	33
3.2 伯努利分布	36
3.3 类别分布	37
3.4 高斯分布	38
3.5 总结	39
4 推荐与检验	41
4.1 统计量和充分统计量	41

4.2	抽样分布	43
4.2.1	正态分布	44
4.2.2	t 分布	44
4.2.3	卡方分布	45
4.3	极限理论	46
4.3.1	马尔可夫和切比雪夫不等式	47
4.3.2	弱大数定律	48
4.3.3	依概率收敛	48
4.3.4	中心极限定理	49
4.3.5	强大数定理	50
4.4	似然估计量	51
4.4.1	估计量的偏差与方差	52
4.4.2	信息量	53
4.4.3	最大似然估计的特性	56
4.5	置信区间	59
4.5.1	均值参数的 Z 区间估计	61
4.5.2	均值参数的 T 区间估计	62
4.5.3	方差参数的区间估计	62
4.6	简单假设检验	64
4.6.1	Z 检验	70
4.6.2	T 检验	70
4.6.3	卡方检验	71
5	指数族	73
5.1	指数族的定义	73
5.1.1	伯努利分布	74
5.1.2	类别分布	75
5.1.3	泊松分布	76
5.1.4	高斯分布	76
5.1.5	其它常见指数族	77
5.2	指数族的期望与方差	77
5.3	最大似然估计	78
5.4	最大似然估计与 KL 散度的关系	80
6	线性回归模型	83
6.1	最小二乘	83
6.1.1	最小误差	83
6.1.2	参数估计	85
6.2	线性回归的概率解释	86
6.2.1	高斯分布	86
6.2.2	参数估计	87
7	广义线性模型	89
7.1	指数族分布	90
7.1.1	自然指数族	90
7.1.2	示例: 高斯分布	92
7.1.3	示例: 伯努利分布	93
7.2	广义线性模型	93
7.3	例子	97
8	参数估计	99
8.1	最大似然估计	99
8.2	泰勒级数	104
8.3	梯度下降法	104
8.4	牛顿法	105

8.4.1	算法推导	106
8.4.2	标准连接函数	107
8.4.3	迭代初始值的设定	108
8.5	迭代重加权最小二乘 (IRLS)	109
8.5.1	算法推导	110
8.5.2	算法过程	111
8.6	估计量的标准误差	114
8.7	分散参数的估计	114
9	模型评估	117
9.1	拟合优度	117
9.1.1	嵌套模型	118
9.1.2	对数似然比 (Likelihood ratio)	120
9.1.3	偏差 (deviance)	120
9.1.4	决定系数 R^2	123
9.1.5	广义皮尔逊卡方统计量	126
9.2	残差分析 (Residual analysis)	127
9.2.1	Response residuals	127
9.2.2	Working residuals	127
9.2.3	Partial residuals	127
9.2.4	Pearson residuals	128
9.2.5	Deviance residuals	128
9.2.6	Adjusted deviance residuals	128
9.2.7	Likelihood residuals	128
9.2.8	Score residuals	128
9.2.9	Anscombe residuals	128
9.3	模型选择 (model selection)	129
9.3.1	Criterion measures	130
10	模型检验	133
10.1	GLM 中的抽样分布	133
10.1.1	得分统计量 (score statistic)	134
10.1.2	最大似然估计量	136
10.1.3	偏差 (deviance) 统计量	138
10.1.4	参数估计量	139
10.2	GLM 中的模型检验	139
10.2.1	模型检验	139
10.2.2	参数检验	140
10.2.3	模型比较	140
10.2.4	正态性检验	142
10.3	案例	142
10.3.1	线性回归	142
10.3.2	GLM	142
10.4	笔记	142
11	高斯模型	143
11.1	传统线性回归	143
11.2	高斯分布	144
11.3	高斯回归模型	144
11.4	参数估计	145
11.4.1	似然函数	145
11.4.2	IRLS	146
11.4.3	拟合优度	147
11.5	其它链接函数	149

12 逆高斯模型	151
12.1 逆高斯分布	151
12.2 逆高斯回归模型	153
12.3 参数估计	154
12.3.1 似然函数	154
12.3.2 IRLS	154
12.3.3 拟合优度	155
12.4 其它连接函数	155
13 二项式模型	157
13.1 伯努利分布	157
13.2 逻辑回归模型	158
13.2.1 模型定义	158
13.2.2 参数估计	159
13.2.3 odds 与 logit	160
13.3 二项式分布	160
13.4 二项式回归模型	161
13.4.1 模型定义	161
13.4.2 参数估计	162
13.5 其它连接函数	163
13.5.1 恒等连接函数	164
13.5.2 probit 回归	164
13.5.3 log-log 和 clog-log	166
13.6 分组数据与比例数据	167
14 泊松模型	169
14.1 泊松 (Poisson) 分布	169
14.1.1 推导过程	170
14.1.2 泊松分布的特性	171
14.2 泊松回归模型	172
14.3 参数估计	173
14.4 拟合统计量	174
14.5 频率模型	174
14.6 泊松模型的局限性	175
15 指数模型	177
15.1 指数 (exponential) 分布	177
15.1.1 推导过程	177
15.1.2 分布的特性	178
15.2 指数回归模型	178
15.3 参数估计	180
15.3.1 似然函数	180
15.3.2 拟合优度	181
15.3.3 IRLS	181
16 Gamma 模型	183
16.1 Gamma 函数	183
16.2 Gamma 分布	183
16.3 Gamma 回归模型	186
16.4 参数估计	188
16.4.1 似然函数	188
16.4.2 IRLS	188
16.4.3 拟合优度	188
16.5 其他连接函数	189
16.5.1 对数 Gamma 模型	189

16.5.2 恒等 (identity) Gamma 模型	189
17 过度分散	191
17.1 什么是过度分散	192
17.2 过度分散的检测	193
17.3 过度分散的影响	193
17.4 标准误差的修正	194
18 负二项式模型	195
18.1 负二项式分布	195
18.1.1 从二项式分布推导	196
18.1.2 泊松-伽马混合分布	197
18.1.3 辅助参数 α 的影响	198
18.2 负二项回归模型	201
18.3 参数估计	203
18.3.1 IRLS	204
18.3.2 参数 α 的估计	205
18.4 负二项式模型扩展	206
18.4.1 对数连接函数	206
18.4.2 参数 α 的估计	207
18.4.3 几何模型	207
18.4.4 广义负二项式模型	209
19 零计数问题	211
19.1 零截断模型	211
19.1.1 零截断泊松模型	212
19.1.2 零截断负二项式模型	212
19.2 零膨胀模型	212
19.2.1 Hurdle 模型	213
19.2.2 Zero-inflate 模型	214
20 多项式模型	217
20.1 类别分布	217
20.2 softmax 回归模型	218
20.2.1 模型定义	218
20.2.2 参数估计	222
20.3 多项式分布	223
20.4 多项式回归模型	223
21 有序离散模型	225
21.1 有序逻辑回归	225
21.2 参数估计	227
21.3 连接函数	228
21.3.1 logit	228
21.3.2 probit	228
21.3.3 clog-log	229
21.3.4 log-log	229
21.3.5 cauchit	229
21.4 总结	229
22 附录	231
22.1 标准正态累积分布表	231
22.2 卡方分布临界值表	232
23 参考文献	235

CHAPTER 1

前言

概率基础

广义线性模型的理论大量依赖概率论的知识，因此本章先回顾一下概率论的一些基础知识。为了帮助非数学专业的读者更容易理解和入门，本章乃至本书都是采用大白话的方式进行讲解，并不追求严谨的学术定义，所以一些描述可能并不严谨。

2.1 概率模型

在日常生活中，经常会遇到某些”事情”的结果是不确定的，比如投掷一枚硬币，其结果可能正面朝上，也可能反面朝上，更有可能是立着。一般来说，如果一件”事情”的结果是不确定的，那么就意味着这件”事情”多个可能的结果。反过来，如果一件”事情”只有一种结果，那么这个结果的发生就是必然的，这样的”事情”的结果就是确定性的。通常可以把”事情”的结果具有不确定性的现象，称为**随机现象**。比如投硬币、掷骰子等。

一个具有不确定性的”事情”，其结果的发生具有随机性。那么每种结果发生的”可能性”是多少呢？能否具体的量化出来呢？如果可以把每种结果的可能性量化出来，就可以帮助我们对结果进行预判。最典型的例子就是赌博，投掷一枚骰子的结果是随机的，如果能清楚的知道每个点数的可能性的大小，就可以一直押注最大可能性的点数，这样就稳赚不赔了。

2.1.1 概率律

概率模型就是对不确定现象的数学描述，每种可能结果发生的可能性的量化结果就是**概率律**。比如正常的投掷一枚正常的硬币，其结果是正面向上的概率是0.5，反面向上的概率也是0.5，至于立起来的结果，我们认为其几乎不可能发生，因此立起来的概率是0。概率值越大，意味其发生的可能性越大。

我们可以把每一个概率模型都关联着一个**试验**，试验的所有可能结果和这个概率模型的所有结果一一对应，该试验的所有可能结果就构成**样本空间**，用 Ω 表示样本空间，样本空间的子集称之为**随机事件**，通常用大写的字母表示随机事件。

我们以掷骰子为例，一个六面体的骰子，把投掷骰子的行为定义试验，投掷的结果有六种可能，这六种结果就构成了样本空间 Ω ，空间 Ω 中有六个样本点，点数为1的样本点(结果)就是一个随机事件，同样点数为2的样本点(结果)也是一个随机事件。以此类推，样本空间 Ω 中的每一个样本点都可以看做是一个随机事件，随机事件的结果可以是发生，也可以是不发生。

随机事件是一个随机试验的样本空间的子集, 注意, 这里是子集, 而不是单个样本点。子集是样本点的集合, 可以包含多个样本点。比如掷骰子的试验, 其样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$, 样本空集的一个子集 $A = \{1, 3, 5\}$, 可以描述成“结果为单数的随机事件”, 与之对应的另一个随机事件就是“结果为偶数”。再比如, 可以定义一个“结果为 1 或 2”的随机事件。原则上对于随机事件的定义(子集的划分)没有限制, 但是, **一个样本空间的多个随机事件必须是互斥的**。像“结果是 1 或者 3”与“结果是 1 或者 4”, 这样两个事件是不允许的, 因为它们两个存在交集 1。并且, **一个样本空间中所有事件的并集, 是这个样本空间全部的样本点**。

概率是对一个随机事件发生的可能性的量化结果, 对概率最直观的理解是 **频率**。在重复进行多次互不影响试验的结果中, 事件发生的频率就可以看做是这个事件发生的概率。随机事件 A 的概率记作 $P(A)$ 。

比如投掷骰子的试验, 假设重复进行 N 次, 点数为 1 的事件发生了 n_1 次, 点数为 2 的事件发生了 n_2 次, 以此类推, 点数为 i 的事件发生了 n_i 次。则有, $N = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$ 。其中每个点数发生的频次(数)是 n_i , 发生的 **频率**是 $\frac{n_i}{N}$ 。则点数为 1 的事件发生概率为 $P(1) = \frac{n_1}{N}$, 同理, 点数为 i 的事件发生概率为 $P(i) = \frac{n_i}{N}$ 。

按照 **概率 = 频率** 的定义, 概率自然也符合频率的一些特性。比如频率一定是正数, 并且频率是大于等于 0 小于等于 1 的, 并且, 同一个样本空间中, 所有随机事件发生的频率之和为 1, 这是因为所有事件发生的频次相加就等于试验总次数。

$$\frac{n_1}{N} + \cdots + \frac{n_i}{N} = \frac{N}{N} = 1 \quad (2.1.1)$$

假定我们已经确定了样本空间 Ω 以及与之关联的试验, **概率律**确定了任何结果或者任何结果的集合(称为随机事件)的似然(可能性)程度。更精确一点的说, 它给每一个事件 A , 确定一个数 $P(A)$, 称为事件 A 的概率, 概率律需要满足下面几条公理。

概率公理

- **非负性**。对一切事件 A , 满足 $P(A) \geq 0$
- **可加性**。设 A 和 B 是两个互不相容的事件, 则它们的并满足:

$$P(A \cup B) = P(A) + P(B) \quad (2.1.2)$$

更一般地, 若 A_1, A_2, \dots 是互不相容的事件序列, 则它们的并满足

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots \quad (2.1.3)$$

- **归一化**。整个样本空间 Ω (称为必然事件) 的概率为 1, 即 $P(\Omega) = 1$ 。

2.1.2 离散模型

当样本空间由有限个样本点组成时, 称之为**离散模型**。假设一个离散样本空间 $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$, 其含有 N 个样本点, N 是有限的, 定义在这个样本空间的随机事件集合为 $S = \{s_1, s_2, \dots, s_N\}$ 。假设随机事件集合 S 和样本空间 Ω 是一一对应的, 即事件 s_i 表示实验结果是样本点 ω_i 。则事件 s_i 发生的概率为

$$P(s_i) = \frac{\text{试验结果中 } \omega_i \text{ 的次数}}{\text{试验的总次数}} \quad (2.1.4)$$

并且, 事件 $\{s_1, s_2, \dots, s_n\}$ 的概率是 $P(s_i)$ 之和。

$$P(s_1, s_2, \dots, s_n) = P(s_1) + P(s_2) + \dots + P(s_n) \quad (2.1.5)$$

2.1.3 连续模型

若样本空间是一个连续值集合, 称之为连续模型, 此时样本点的数量是无限的。连续值模型和离散模型有很大的不同, 在连续值模型中, 由于样本点的数量是无限的, 如果单个样本点的概率为正数, 则所有样本点的概率之和将无穷大, 这显然是不行的。因此我们将连续值模型中, 单个样本点的概率定义为 0。那要如何表示连续值模型的概率呢?

连续值模型的样本空间是一段连续值的区间, 我们可以把这个区间分给成一份一份的, 然后定义每一份的概率就是这一份的长度和整个区间长度的比值。比如, 在赌场中有一种幸运大转盘的赌具, 假设这个圆盘被分割成 12 个扇形, 转动圆盘, 当圆盘停止时, 指针指向哪个区域, 就表示这个区域所代表的事件发生了。如果圆盘是被等分成 12 份, 则指针落在每个区域的概率都是 $1/12$ 。

显然, 如果样本空间是一个一维空间, 则可以划分成一个个的线段, 每个线段可以代表一个事件, 事件的发生概率就是线段长度和样本空间总长度的比值。如果样本空间是一个二维平面空间, 则可以划分成一个个子平面, 每个子平面代表一个事件, 事件的发生概率就是子平面的面积和样本空间总面积的比值。以此可以类推更高维的空间。

连续概率模型的计算, 就是把整个样本空间分割成子区间, 每个子区间的概率值就是这个子区间和整个样本空间的比值。显然通过这样的定义得到的概率律, 也是符合概率的三个公理的。本质上就是把连续值区间离散化了。

2.2 条件概率

条件概率是给定部分信息的基础上对实验结果的一种推断。例如在连续两次抛掷骰子的试验中, 已知两次抛掷的点数的总和为 9, 第一次抛掷的点数为 6 的可能性有多大。换句话说, 假设我们已经知道给定的事件 B 发生了, 而希望知道令一个给定事件 A 发生的可能性。因此, 我们需要构建一个新的概率律, 它顾及了事件 B 已经发生的信息, 求出任何事件 A 发生的概率。这个概率就是给定 B 发生之间事件 A 的 **条件概率**, 记作 $P(A|B)$, 读作 B 的条件下 A 的概率。当然, 条件概率也必须符合三条概率公理。

我们用实际的例子来说明条件概率。

例 1: 箱子里取球

假设我们有两个箱子分别为 a_1, a_2 , 箱子中分别装有红色球和白色球。假设 a_1 箱子中有 4 个红色球和 6 个白色球, a_2 箱子中有 8 个红色球和 2 个白色球。另外我们有一个特殊的硬币, 投放后正面向上的概率是 0.6, 反面向上的概率是 0.4。现在我们进行如下实验:

- 步骤 1. 投掷硬币, 然后观察硬币的朝向, 根据硬币的朝向选择一个箱子。如果正面向上就选择 a_1 箱子; 如果反面朝上, 就选择 a_2 箱子。
- 步骤 2. 从选出的箱子中随机(不允许刻意挑选)取出一个球, 并记录球的颜色。

假设投掷硬币的结果组成样本空间 $\Omega_{\text{币}} = \{\text{正}, \text{反}\}$, 正面向上的结果定义为事件 $B_{\text{正}}$, 反面向上的结果定义为事件 $B_{\text{反}}$ 。取出球的颜色组成的样本空间为 $\Omega_{\text{球}} = \{\text{红}, \text{白}\}$, 定义红色的结果为事件 $A_{\text{红}}$, 白色的结果为事件 $A_{\text{白}}$ 。

- 已知事件 $B_{\text{正}}$ 发生的情况下, 事件 $A_{\text{红}}$ 发生的概率, 就是条件概率 $P(A_{\text{红}}|B_{\text{正}}) = 4/10 = 0.4$ 。
- 已知事件 $B_{\text{正}}$ 发生的情况下, 事件 $A_{\text{白}}$ 发生的概率, 就是条件概率 $P(A_{\text{白}}|B_{\text{正}}) = 6/10 = 0.6$ 。
- 已知事件 $B_{\text{反}}$ 发生的情况下, 事件 $A_{\text{红}}$ 发生的概率, 就是条件概率 $P(A_{\text{红}}|B_{\text{反}}) = 8/10 = 0.8$ 。
- 已知事件 $B_{\text{反}}$ 发生的情况下, 事件 $A_{\text{白}}$ 发生的概率, 就是条件概率 $P(A_{\text{白}}|B_{\text{反}}) = 2/10 = 0.2$ 。

例 2: 掷骰子

假设有一个六面的骰子, 投掷结果中每个面的概率相同。如果我们已知试验的结果是偶数, 即 2, 4, 6 这三种结果之一发生, 由于这三个结果发生的可能性是相等的, 这样可以得到

$$P(\text{试验结果是 } 6 | \text{试验结果是偶数}) = \frac{1}{3} \quad (2.2.1)$$

从这个结果的推导可以看出, 对于等概率模型的情况, 下面关于条件概率的定义是合适的, 即

$$P(A|B) = \frac{\text{事件 } A \cap B \text{ 的试验结果数}}{\text{事件 } B \text{ 的试验结果数}} \quad (2.2.2)$$

将这个结果推广, 我们得到下面的条件概率的定义:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.2.3)$$

其中假定 $P(B) > 0$ 。如果 $P(B) = 0$, 相应的条件概率是没有意义的。总而言之, $P(A|B)$ 是事件 $A \cap B$ 的概率与事件 B 的概率的比值。

这个式子可以理解成, 在事件 B 发生的结果中, 事件 A 发生的结果数和事件 B 发生次数的比值。如下图所示, 整个矩形空间 S 是全部结果集, 两个圆圈分别是事件 A 和事件 B 发生的结果集, A 与 B 的交集部分, 就是 A 与 B 同时发生的结果集。

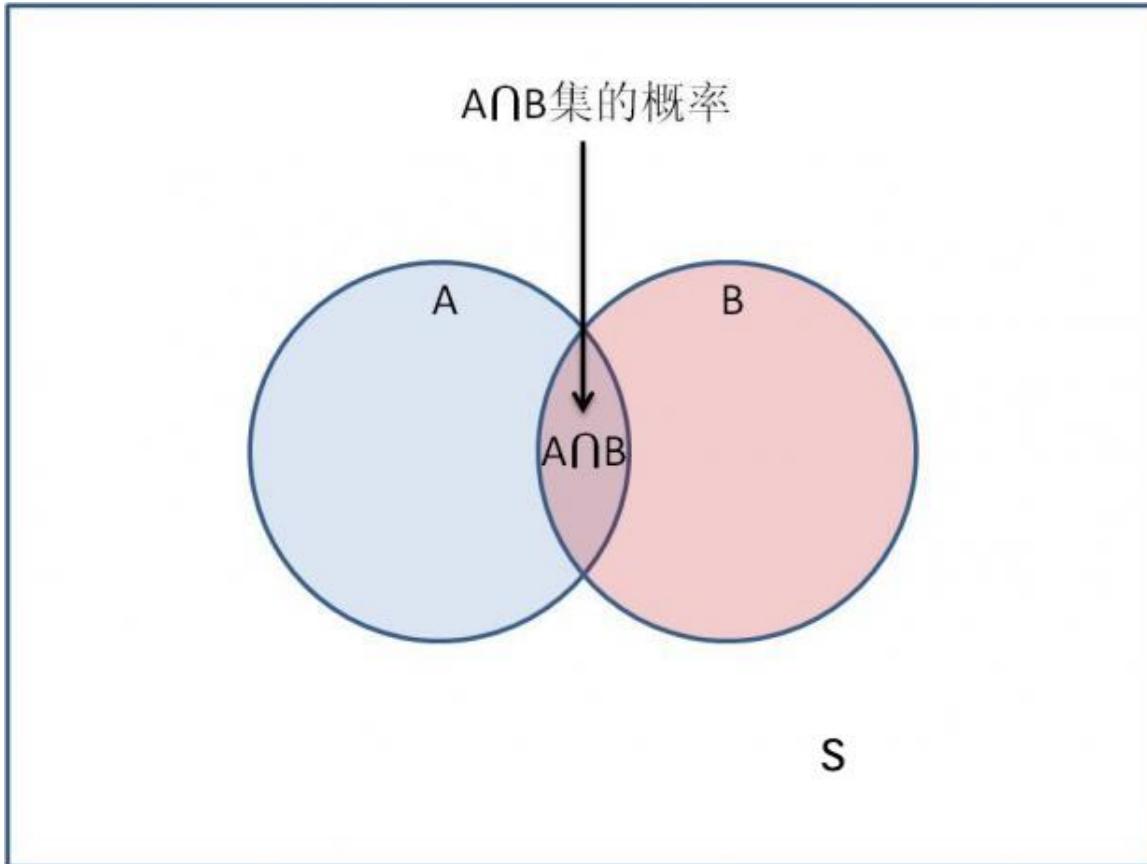


图 2.2.1: $P(A) = \frac{A}{S}$, $P(B) = \frac{B}{S}$, $P(A|B) = \frac{A \cap B}{B} = \frac{P(A \cap B)}{P(B)}$

条件概率 $P(A|B)$ 表示在 B 发生的条件下 A 发生的概率，就是 **限定在 B 的范围内 A 发生的概率**。 B 的范围内就是 B 发生结果内， B 的范围内 A 的结果数是 $A \cap B$ 结果数，因此条件概率 $P(A|B)$ 就等于

$$P(A|B) = \frac{A \cap B}{B} \quad (2.2.4)$$

注意，这里分母是 B 而不是 S ，因为是 B 的前提下。分子分母同时除以 S 后等价于

$$P(A|B) = \frac{(A \cap B)/S}{B/S} = \frac{P(A \cap B)}{P(B)} \quad (2.2.5)$$

条件概率的性质

- 设事件 B 满足 $P(B) > 0$ ，则给定 B 的条件下，事件 A 的条件概率由下式给出

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.2.6)$$

- 由于条件概率所关心的事件都是事件 B 的子事件，可以把条件概率看作是 B 上的概率律，即把事件 B 看作是全空间或者必然事件。
- 当试验的 Ω 为有限集，并且所有试验结果为等可能的情况下，条件概率可以由下式给出。

$$P(A|B) = \frac{\text{事件 } A \cap B \text{ 的试验结果数}}{\text{事件 } B \text{ 的试验结果数}} \quad (2.2.7)$$

总结起来就一句话，条件概率就是把试验结果空间缩小到一个更小的空间，其它照旧。

2.3 联合概率

假设两个随机事件 A 和 B ，在已知 B 发生的条件下 A 发生的概率是条件概率 $P(A|B)$ 。那如果不知道 B 是否发生了呢？在没有任何已知条件下， A 和 B 同时发生的概率是什么呢？

我们定义，**属于不同样本空间的多个随机事件同时发生的概率为联合概率**，记作 $P(A, B)$ 。

观察 图 2.2.1， A 和 B 同时发生的集合就是 $A \cap B$ ，因此 A 和 B 的联合概率为

$$P(A, B) = \frac{A \cap B}{S} \quad (2.3.1)$$

实际上，联合概率和条件概率之间是存在关系的，二者可以互相转换。

$$P(A, B) = \frac{A \cap B}{S} = \frac{B}{S} \times \frac{A \cap B}{B} = P(B)P(A|B) \quad (2.3.2)$$

我们继续以箱子里取球为例，在上面的例子中，

- 事件 $B_{\text{正}}$ 与 $A_{\text{红}}$ 同时发生概率就是 $P(B_{\text{正}}, A_{\text{红}}) = 0.6 \times 0.4 = 0.24$ 。
- 事件 $B_{\text{正}}$ 与 $A_{\text{白}}$ 同时发生概率就是 $P(B_{\text{正}}, A_{\text{白}}) = 0.6 \times 0.6 = 0.36$ 。
- 事件 $B_{\text{反}}$ 与 $A_{\text{红}}$ 同时发生概率就是 $P(B_{\text{反}}, A_{\text{红}}) = 0.4 \times 0.8 = 0.32$ 。
- 事件 $B_{\text{反}}$ 与 $A_{\text{白}}$ 同时发生概率就是 $P(B_{\text{反}}, A_{\text{白}}) = 0.4 \times 0.2 = 0.08$ 。

也可以用一个 2×2 的表格来表示。

	$B_{\text{正}}$	$B_{\text{反}}$	
$A_{\text{红}}$	$0.6 \times 0.4 = 0.24$	$0.4 \times 0.8 = 0.32$	
$A_{\text{白}}$	$0.6 \times 0.6 = 0.36$	$0.4 \times 0.2 = 0.08$	

(2.3.3)

当随机事件更多的时候, 上面的表格就无法表示了, 此时可以用如下的表格

B	A	$P(A, B)$	
正	红	$0.6 \times 0.4 = 0.24$	
正	白	$0.6 \times 0.6 = 0.36$	
反	红	$0.4 \times 0.8 = 0.32$	
反	白	$0.4 \times 0.2 = 0.08$	

(2.3.4)

可以看到联合概率可以分解成条件概率的乘积, 我们可以扩展到更多事件的联合概率。假设有 A, B, C, D 四个随机事件, 它们组成的联合概率可以分解为

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C) \quad (2.3.5)$$

更一般地, 假设有 A_1, A_2, \dots, A_N 共 N 个随机事件, 它们的联合概率可以写为

$$P(A_1, A_2, \dots, A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_N|A_1, A_2, \dots, A_{N-1}) \quad (2.3.6)$$

公式 (2.3.6) 被称为联合概率的 **链式法则**。

联合概率就是多个随机事件同时发生的概率, 它的计算方法就是按照事件发生的先后顺序拆解成一系列条件概率的乘积。在公式 (2.3.6) 中, 事件 A_1 是第一个发生的事件, 它的前面没有其它事件了, 因此 A_1 的发生概率不是条件概率, 而是 $P(A_1)$ 。在 A_1 之后发生的事件就都是以 A_1 为前置条件了, 依次类推, 最后一个事件 A_N 是在其它 $N-1$ 个事件之后发生的。

提示: 先发生的事件是后发生的事件的前置条件, 你可以把先发生的事件理解成“因”, 后发生的事件看作是“果”, 那么条件概率就是一种因果关系。

最后, 联合概率也是一个合格的概率律, 也符合概率三公理。

2.4 全概率与贝叶斯定理

上一节我们讲到, 联合概率可以按照事件发生的顺序拆解成条件概率的乘积。如果不按照事件发生的顺序呢, 可不可以把顺序反过来呢? 答案是可以的, 但是反过来后将会面临一个问题, 最后一个事件 A_N 的概率 $P(A_N)$ 是什么?

$$P(A_1, A_2, \dots, A_N) = P(A_N)P(A_{N-1}|A_N)P(A_{N-2}|A_N, A_{N-1}) \cdots P(A_1|A_N, A_{N-1}, \dots, A_2) \quad (2.4.1)$$

我们继续以箱子取球为例, 为简化说明, 我们重新定义事件 B 是事件 $B_{\text{正}}, B_{\text{反}}$ 的集合, 记作 $B = \{B_{\text{正}}, B_{\text{反}}\}$ 。同理, 事件 A 是事件 $A_{\text{红}}, A_{\text{白}}$ 的集合, 记作 $A = \{A_{\text{红}}, A_{\text{白}}\}$ 。

在这个例子中, 先投掷硬币, 然后根据硬币的朝向决定从哪个箱子里取球。因此事件 B 先发生, 事件 A 后发生, 可以把事件 B 看做“因”, 把事件 A 看做“果”。根据链式法则, 二者的联合概率 $P(A, B)$ 可以写成

$$P(A, B) = P(B)P(A|B) \quad (2.4.2)$$

如果我们把公式 (2.4.2) 中的事件顺序反过来, 就是

$$P(A, B) = P(A)P(B|A) \quad (2.4.3)$$

这时就产生了一个问题, $P(A)$ 是多少? $P(B|A)$ 又是多少? 形象一点就是, $P(A)$ 代表 $P(\text{果})$, $P(B|A)$ 代表 $P(\text{因}| \text{果})$, 从“果”到“因”就是**推断 (inference)** 问题。本节我们讨论的 **全概率公式**和**贝叶斯定理**就是分别来求得 $P(A)$ 和 $P(B|A)$ 的方法, 我们先从 $P(A)$ 说起。

事件 $A = \{A_{\text{红}}, A_{\text{白}}\}$ 表示球的颜色事件(集合), 球的颜色有红白两种, 取到红球白球的概率会受到硬币事件 $B = \{B_{\text{正}}, B_{\text{反}}\}$ 的影响, 因此可以把事件 B 看做是“因”, 事件 A 看作是“果”。现在我们回顾一下 B 与 A 的联合概率表, 如下表所示。

B	A	$P(A, B)$	
正	红	$P(\text{正})P(\text{红} \text{正}) = 0.6 \times 0.4 = 0.24$	
正	白	$P(\text{正})P(\text{白} \text{正}) = 0.6 \times 0.6 = 0.36$	
反	红	$P(\text{反})P(\text{红} \text{反}) = 0.4 \times 0.8 = 0.32$	
反	白	$P(\text{反})P(\text{白} \text{反}) = 0.4 \times 0.2 = 0.08$	

(2.4.4)

事件 $A = A_{\text{红}}$ 的状态被分割成了两部分, 一部分是 $P(A = \text{红}, B = \text{正})$, 另一部分是 $P(A = \text{红}, B = \text{反})$ 。显然 $P(A = \text{红})$ 的概率需要把两部分加起来。

$$P(A = \text{红}) = P(A = \text{红}, B = \text{正}) + P(A = \text{红}, B = \text{反}) = 0.24 + 0.32 = 0.56 \quad (2.4.5)$$

同理, $P(A = \text{白})$ 为

$$P(A = \text{白}) = P(A = \text{白}, B = \text{正}) + P(A = \text{白}, B = \text{反}) = 0.36 + 0.08 = 0.44 \quad (2.4.6)$$

可以看出, 事件 A 的每个状态都被事件 B 分割了, 分割的份数就是事件 B 的状态数量, 要想求得事件 A 每个状态的概率, 就需要把被 B 分割的各个部分加起来才行。总结起来就是这样

$$\begin{aligned} P(A = a) &= P(B = b_1)P(a|B = b_1) + P(B = b_2)P(a|B = b_2) + \cdots + P(B = b_n)P(a|B = b_n) \\ &= \sum_{i=1}^n P(B = b_i)P(a|B = b_i) \\ &= \sum_B P(B)P(a|B) \\ &= \sum_B P(A = a, B) \end{aligned} \quad (2.4.7)$$

以上就是全概率公式, 简单来说 全概率公式就是联合概率中消除掉一个事件得到剩余事件的概率。消除掉的方法就是对这个事件的各个状态进行求和, 如果被消除事件是一个连续值概率模型, 就把求和符号换成积分。

全概率定理

设 B_1, B_2, \dots, B_n 是一组互不相容的事件, 形成样本空间的一个分割 (每一个试验结果必定使得其中一个事件发生), 又假定对每一个 i , $P(B_i) > 0$, 则对于任何事件 A , 下列公式成立。

$$\begin{aligned} P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n) \\ &= P(A, B_1) + P(A, B_2) + \cdots + P(A, B_n) \\ &= \sum_{i=1}^n P(A, B_i) \end{aligned} \quad (2.4.8)$$

回到最初的问题, 我们已经可以通过全概率公式得到“果”的概率 $P(A)$, 现在看下如何从“果”推断出“因”, 即 $P(B|A)$ 。我们可以把公式 (2.4.2) 和公式 (2.4.3) 放在一起。

$$P(A, B) = P(A)P(B|A) = P(B)P(A|B) \quad (2.4.9)$$

通过移项可得

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \quad (2.4.10)$$

其中分子部分 $P(B)P(A|B)$ 是正向的”因果”关系, 我们是已知的, 分母 $P(A)$ 可以通过全概率公式得到。二者的比值就得到了 $P(B|A)$ 。

公式 (2.8.9) 就是 **贝叶斯定理**, 又叫做贝叶斯公式。贝叶斯定理就是贝叶斯推断的核心, 经常被用来做 **因果推断**。有许多”原因”可以造成某一种”结果”, 当已知结果要推断成因时, 就是”因果推断”。所谓推断成因, 就是推断出造成这一结果的每种原因的概率是多少。

现在设事件 B_1, B_2, \dots, B_n 是原因, 而 A 代表由原因引起的结果。 $P(A|B_i)$ 表示在因果模型中由”原因” B_i 造成结果 A 出现的概率。当观察到结果 A 的时候, 我们希望反推结果 A 是由原因 B_i 造成的结果 $P(B_i|A)$ 。 $P(B_i|A)$ 代表新进得到的信息 A 之后 B_i 出现的概率, 因此称之为 **后验概率**。 $P(B_i)$ 是在没有信息之前就知道的 B_i 出现的概率, 因此称之为 **先验概率**。 $P(A|B_i)$ 是有了原因 B_i 之后 A 出现的可能性, 因此称之为 **似然**。最后, 贝叶斯公式可以用如下方式描述。

$$\text{后验概率} = \frac{\text{先验概率} \times \text{似然}}{\text{全概率}} \quad (2.4.11)$$

此外, 可以注意到, 分母全概率公式 $P(A) = \sum_B P(B)P(A|B)$ 其实就是分子 $P(B)P(A|B)$ 的累加。

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} = \frac{P(B)P(A|B)}{\sum_B P(B)P(A|B)} \quad (2.4.12)$$

因此作为分母的全概率可以看做是分子的 **归一化项**, 归一化的结果是把等式右侧的数值转换到区间 $[0, 1]$, 使其符合概率的定义。相对于分子来说, 分母的值是固定不变的, 贝叶斯公式公式 (2.4.12) 可以简写成一个正比关系。

$$\begin{aligned} P(B|A) &\propto P(B)P(A|B) \\ \text{后验} &\propto \text{先验} \times \text{似然} \end{aligned} \quad (2.4.13)$$

2.5 独立性

前面的内容中, 我们探讨了多个随机事件的关系, 条件概率、联合概率以及因果推断。那如果两个随机事件没有任何关系呢? 如果随机事件之间没有任何关系, 我们称它们是 **相互独立事件**。

如果两个事件 A 和 B 是相互独立的, 则事件 B 的发生不会改变事件 A 的概率, 反之亦然。

$$\begin{aligned} P(A) &= P(A|B) \\ P(B) &= P(B|A) \end{aligned} \quad (2.5.1)$$

在上述等式成立的情况下, 我们称事件 A 独立于事件 B , 记作 $A \perp\!\!\!\perp B$ 。如果两个事件是独立的, 则它们的联合概率将变得简单。

$$P(A, B) = P(A)P(B) \quad (2.5.2)$$

有些时候, 单独看两个事件可能不是独立的, 但是在给定另外一个条件下是独立的。例如在给定事件 C 发生的情况下, 事件 A 与 B 是相互独立的, 记作 $A \perp\!\!\!\perp B|C$, 这种情况称之为 **条件独立**。

$$P(A, B|C) = P(A|C)P(B|C) \quad (2.5.3)$$

$$P(A, B, C) = P(C)P(A, B|C) = P(C)P(A|C)P(B|C) \quad (2.5.4)$$

2.6 随机变量

试验的所有可能结果形成了样本空间 $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ ，样本空间中的每个样本点 ω_i 就表示试验的一种可能结果，一个试验的结果必定是样本空间 Ω 中的一个。在许多概率模型中试验结果是数值化的，例如股价或者骰子的点数等等，也有一些试验结果不是数值化的，例如投硬币的结果。显然这样既有非数字又有数值的情况，不利于我们的研究和处理，因此我们希望能全部转换成数值进行处理，而这可以通过 **随机变量** 来实现。

我们把样本空间中的每一个可能的试验结果，关联一个特定的数，这种试验结果与数的对应关系形成 **随机变量**。更直白的说就是，我们用一个变量符号来表示实验结果，变量的取值就是试验结果所对应的数。从数学上将，随机变量是试验结果的实值函数，随机变量通常用大写的字母表示。我们用两个例子来说明。

首先以投硬币的试验为例，投硬币的结果形成样本空间 $\Omega = \{\text{正, 反}\}$ ，首先我们需要把非数值的样本映射成数值，比如我们把正面向上的结果映射成数字 1，把反面向上的结果映射成数字 0，样本空间就变成了一个数值空间 $\Omega = \{1, 0\}$ 。然后定义一个随机变量 X 来表示试验的结果，那么变量 X 的取值空间就是数值空间 $\Omega = \{1, 0\}$ 。 $X = 1$ 表示试验结果是正面向上， $X = 0$ 表示试验结果是反面向上。

再比如掷骰子的试验中，试验结果的样本空间就是六种点数，记作 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。在这个试验中，样本空间已经是数值化了，因此就不需要进行数值转化了，我用变量 X 表示试验的结果，则变量 X 的取值空间就是 $\{1, 2, 3, 4, 5, 6\}$ ，通常用小写的字母表示变量的取值，此时变量 $X = x$ 就表示试验结果是 x 。

现在再举几个随机变量的例子。

1. 连续抛掷一枚硬币共 5 次，在这个试验中正面出现的次数是一个随机变量。
2. 在两次抛掷一个骰子的试验中，下面的例子是随机变量。
 - a. 两次抛掷骰子所得到的点数之和。
 - b. 两次抛掷得到 6 点的次数。
 - c. 第二次抛掷所得到的点数的 5 次方。
3. 在传输信号的试验中，传输信号所需的时间、接收到的信号中发生错误的次数、传输信号过程中的时间延迟等都是随机变量。

与随机变量相关的主要概念

在一个试验的概率模型之下：

- **随机变量** 是试验结果的实值函数；
- **随机变量的函数** 定义了另一个随机变量；
- 对于一个随机变量，我们可以定义一些平均量，例如 **均值** 和 **方差**；
- 可以在某事件或某随机变量的 **条件** 之下定义一个随机变量；
- 存在一个随机变量与某事件或者某随机变量相互 **独立** 的概率。

随机变量的这些特性当中，比较重要的一点是，**随机变量的函数仍然是一个随机变量**。这一点在本书之后的内容中会使用，比如统计量、参数估计量就是建立在这一点之上。

提示：如果你难以理解随机变量的概念，没关系，可以暂时先把随机变量就理解成随机事件。虽然这样不是很准确，但不妨碍对本书之后内容的理解。随机事件拥有的特性随机变量也有，比如条件概率、联合概率、独立性等等，二者的差别就是，随机事件只有发生、不发生两种结果，而随机变量可以多种结果值，并且随机变量的值是数值（数字）。

样本空间的大小可以是有限的, 也可以是无限的, 有限的样本空间是离散概率模型, 无限的样本空间是连续值概率模型。随机变量是样本空间的实值函数, 因此随机变量也分为 **离散随机变量** 和 **连续随机变量**。

2.6.1 离散随机变量

若一个随机变量的值域 (随机变量的取值范围) 为一个有限集合或最多为可数无限集合, 则称这个随机变量为 **离散的**。由于它只能取有限多个值, 所以是离散的随机变量。

离散随机变量, 既然称为 **随机变量**, 意味着它的取值并不是确定性, 而是具有 **随机性**, 有可能是值域中的任何一个值, 值域中每个值都有一定的概率。假设随机变量 X 表示一次投掷硬币的试验结果, $X = 1$ 表示正面向上结果, $X = 0$ 表示反面向上的结果, 对于一枚正常的硬币两种结果应该是等概率的, 随机变量 X 每种取值的概率情况为

$$\begin{aligned} P(X = 1) &= 0.5 \\ P(X = 0) &= 0.5 \end{aligned} \tag{2.6.1}$$

对于离散随机变量, 其值域是有限个, 因此有时也可以用一个表格的形式表达其各个值的概率情况

X	$P(X)$
1	0.5
0	0.5

(2.6.2)

我们把一个随机变量各个可能取值的概率分布情况, 称为随机变量的 **概率分布**。虽然我们可以用表格的形式表达离散随机变量的概率分布, 但是如果随机变量的值域规模较大, 表格将变得异常庞大, 使用起来也是不方便的。此时, 可以用一个 **数学函数** 来表达随机变量的概率分布, 用来表达随机变量的概率分布的函数就称为 **概率分布函数**。比如, 对于仅有 $\{0, 1\}$ 两个值的随机变量 X , 假设其为 1 的概率是 π , 那么它的概率分布函数可以为

$$P(X) = \pi^x(1 - \pi)^{(1-x)} \tag{2.6.3}$$

把随机变量的某个可能取值代入到概率分布函数, 得到的就是变量为这个值的概率。我们把 1 代入到公式 (2.6.3) 得到就是随机变量 X 为 1 的概率 $P(X = 1)$ 。

$$P(X = 1) = \pi^1(1 - \pi)^{(1-1)} = \pi \tag{2.6.4}$$

同理, 把 0 带入到公式 (2.6.3) 得到就是 $P(X = 0)$ 。

$$P(X = 0) = \pi^0(1 - \pi)^{(1-0)} = 1 - \pi \tag{2.6.5}$$

前文已经讲过随机变量是随机事件的一个扩展, 随机变量的概率分布也是满足概率三公理的, 包括非负性、可加性和归一化。假设有一个离散随机变量 Y , 其值域空间为 $\{y_1, y_2, \dots, y_n\}$, 任意一个值 y_i 的概率记为 $P(y_i)$, 则有如下归一化等式成立。

$$\sum_{i=1}^n P(y_i) = 1 \tag{2.6.6}$$

2.6.2 连续随机变量

在前文讲过, 当样本空间有无限个样本点时就是连续概率模型, 同理, 若一个随机变量可以取到无限多个数, 那么这个随机变量就是 **连续值随机变量**。在连续概率模型中, 单个样本点的概率是 0, 我们需要把整个连续的样本区间划分成子区间, 让后定义样本点落在每个子区间的概率。同理, 连续值随机变量也是一样的。

假设一个连续值随机变量 X ，其值域空间是整个实数域 $X \in \mathcal{R}$ ，它的概率分布函数是 $f(x)$ 。随机变量 X 的取值落在区间 $[a, b]$ 的概率为

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (2.6.7)$$

一定要注意，对于连续值随机变量的概率分布函数 $f(x)$ ， $f(x)$ 的值并不是概率值，而是这一点的 **密度值**，在一个区间上的积分结果随机变量落在这个区间的概率值。

如果我们把这个区间 $[a, b]$ 扩大到 X 的整个值域空间 \mathcal{R} ，此时相当于 $a = -\infty, b = \infty$ ，不管试验结果是什么， X 的值肯定会落在自己的值域空间，落在整个值域空间的概率必然是 1。因此有如下约束必须满足

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1 \quad (2.6.8)$$

也就是说，我们约束连续值随机变量概率分布函数在整个值域的积分必须是 1，这就和离散随机变量值域内所有值的概率之和是 1 一样。虽然在这个例子中，连续值随机变量 X 的值域是整个实数域，但这不是必须的，连续值随机变量的值域空间可以是任意的区间，一般是根据试验的样本空间确定的，对值域空间的范围和大小并没有额外的限制。但不管值域空间是如何的，都是可以通过调整 $f(x)$ 以使得它在整个空间的积分是 1。

离散随机变量的概率分布函数可以直接为每个点计算出概率值（类比于每个点的质量），因此通常称为 **概率质量函数 (probability mass function, pmf)**，而连续值随机变量的概率分布函数，需要积分才能得到概率值（质量），函数本身相当于每个点的 **密度值**，因此连续值随机变量的概率分布函数一般称为 **概率密度函数 (probability density function, pdf)**。

2.6.3 累积分布函数

对于离散随机变量和连续随机变量分别用概率质量函数和概率密度函数刻画他们的概率分布情况，本节我们介绍另一种刻画概率分布的方法，累积分布函数（Cumulative Distribution Function, CDF）。

累积分布函数是概率质量（密度）函数的积分函数，通常使用小写字母 f 代表概率质量（密度）函数和函数，而用大写字母 F 表示累积分布函数。

累积分布函数

假设随机变量 X 的 CDF 是 x 的函数 F_X ，对于每一个 x ， $F_X(x)$ 定义为 $P(X \leq x)$ 。特别地，当 X 为离散或连续的情况下，

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{x \leq x} P_X(k), & \text{若 } X \text{ 是离散的} \\ \int_{-\infty}^x f_X(t)dt, & \text{若 } X \text{ 是连续的} \end{cases} \quad (2.6.9)$$

累积分布函数表示 $F_X(x)$ 将 X 取值的概率由 $-\infty$ 累计到 x 。

在一个概率模型中，随机变量可以有不同的类型，可以是离散的，也可以是连续的，甚至可以是既非离散也非连续的。但不管什么类型的随机变量，它们都会有一个相对应的累积分布函数。这是因为 $\{X \leq x\}$ 是一个随机事件，这些事件的概率形成概率分布。

累积分布函数的性质

随机变量 X 的累积分布函数 F_X 由下式定义，

对每一个 x , $F_X(x) = P(X \leq x)$,

并且 F_X 具有下列形式

- F_X 是单调非减函数。

若 $x_1 \leq x_2$, 则 $F_X(x_1) \leq F_X(x_2)$

- 当 $x \rightarrow -\infty$ 时的时候, $F_X(x)$ 趋近于 0, 当 $x \rightarrow \infty$ 时的时候, $F_X(x)$ 趋近于 1。
- 当 X 是离散随机变量的时候, $F_X(x)$ 为 x 的阶梯函数。
- 当 X 是连续随机变量的时候, $F_X(x)$ 为 x 的连续函数。
- 当 X 是离散随机变量并取整数值的时候, 累积分布函数 $F_X(x)$ 和概率质量函数 $f_X(x)$ 可以利用求和或差分互求:

$$F_X(k) = \sum_{i=-\infty}^k f_X(i) \quad (2.6.10)$$

$$f_X(k) = P(X \leq k) - P(X \leq k-1) = F_X(k) - F_X(k-1)$$

其中 k 可以是任意整数。

- 当 X 是连续随机变量的时候, 累积分布函数 $F_X(x)$ 和概率密度函数 $f_X(x)$ 可以利用积分或微分函数互求:

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \quad (2.6.11)$$

$$f_X(x) = \frac{dF_X}{dx}(x)$$

2.6.4 随机变量的函数

之前已经提到过, 你可以把随机变量看做随机事件的扩展, 随机事件只有发生、不发生两个状态, 而随机变量是试验结果样本空间到数值的映射, 它可以有更多的状态, 样本空间中每个样本点对应着随机变量的一个取值。随机事件拥有的特性随机变量也有, 比如条件概率、联合概率、贝叶斯定理等等, 对随机变量都是成立的, 只需要把那些大写字母的符号看做是随机变量即可, 这里就不再赘述了。本节我们讨论之前没有讨论过的内容, 随机变量的函数。

假设 X 是一个随机变量, $g(X)$ 是随机变量 X 的一个函数, $g(X)$ 就相当于对 X 施行了一个变换, 这种变换可以是线性的也可以是非线性的。假设 $g(X)$ 是一个线性变换

$$Y = g(X) = aX + b \quad (2.6.12)$$

其中 a, b 是数值。我们也可以考虑非线性的函数, 比如

$$Y = g(X) = X^2 \quad (2.6.13)$$

设 $Y = g(X)$ 是随机变量 X 的函数, 由于对于 X 的每一个可能取值, 也对应的 Y 的一个数值, 故 Y 也是一个随机变量, Y 的概率分布可以通过 X 的概率分布计算得到。

设离散随机变量 X 的值域为 $\{-1, -2, 0, 1, 2\}$, 并且每个值的概率都是 $1/5$ 。 Y 是 X 的平方, 即 $Y = g(X) = X^2$, 则 Y 的取值空间为 $\{0, 1, 4\}$ 。

$$Y = \begin{cases} 0, & x = 0 \\ 1, & x \in \{-1, 1\} \\ 4, & x \in \{-2, 2\} \end{cases} \quad (2.6.14)$$

Y 的概率分布为

$$\begin{aligned} P(Y = 0) &= P(X = 0) = \frac{1}{5} \\ P(Y = 1) &= P(X = -1) + P(X = 1) = \frac{2}{5} \\ P(Y = 4) &= P(X = -2) + P(X = 2) = \frac{2}{5} \end{aligned} \quad (2.6.15)$$

多个随机变量的函数也是一样的。假设 X 和 Y 是两个随机变量, 那么 $Z = g(X, Y)$ 也是一个随机变量, 即使更多的随机变量的函数也是如此。

2.6.5 期望与方差

随机变量的取值不是确定性的, 有一定的随机性, 它的概率分布给出了其所有可能取值的概率。随机变量的概率分布不是很方便进行比较和评价, 通常, 我们希望将这些信息综合成一个能代表这个随机变量的 **数值**。在数学和统计学中, **矩 (moment)** 是对变量分布和形态特点的一组度量。**n 阶矩**被定义为变量的 n 次方与其概率分布函数之积的积分, 变量的一阶矩就是变量的期望值, 也叫平均值。

随机变量的期望

设离散随机变量 X 的概率质量函数为 $f(x)$, X 的 **期望值** (也称为 **期望**或者 **均值**) 由下式给出

$$\mathbb{E}[X] = \sum_x x f(x) \quad (2.6.16)$$

设连续值随机变量 Y 的概率密度函数为 $g(x)$, Y 的 **期望值** (也称为 **期望**或者 **均值**) 由下式给出

$$\mathbb{E}[Y] = \int y g(y) dy \quad (2.6.17)$$

随机变量的期望值是一个数值, 而不再是随机变量。随机变量的期望值可以看做是这个变量的 **中心**, 大量重复试验结果的数学平均值就渐近等于变量的期望值, 在之后讲最大似然估计时会详细介绍。

我们已经知道随机变量的函数也是一个随机变量, 随机变量的函数的期望可以用如下方式得到。

随机变量函数的期望

设随机变量 X 的概率分布函数为 $f(x)$, 又设 $h(x)$ 是变量 X 的一个函数, 则 $h(x)$ 的期望由下式得到

$$\mathbb{E}[h(x)] = \int h(x) f(x) dx \quad (2.6.18)$$

如果 X 是离散随机变量, 只需要把积分换成求和。

直接使用变量计算的矩被称为原始矩 (raw moment), 比如期望就是原始矩。移除均值后计算的矩被称为中心矩 (central moment), 变量的一阶原始矩等价于数学期望 (expectation)、二至四阶中心矩被定义为方差 (variance)、偏度 (skewness) 和峰度 (kurtosis)。

随机变量另一个常见的独立方法就是 **方差 (variance)**, 方差是二阶中心矩。所谓中心矩就是去除中心(期望值), 所谓的二阶就是二次方, 因此方差的计算方法为

$$V(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (2.6.19)$$

注意上式中最外层又求了一次期望, 这是因为 $(X - \mathbb{E}[X])^2$ 本身是随机变量 X 的一个函数, 它也是一个随机变量, 因此再对它求一次期望以便得到一个数值。根据随机变量函数期望的求法, 方差的计算方法为

$$V(X) = \int (X - \mathbb{E}[X])^2 f(x) dx \quad (2.6.20)$$

其中 $f(x)$ 是变量 X 的概率分布函数。

方差的值是原始变量的平方, 量纲发生了变化, 其量纲是原始值的平方, 不利于和原始值进行比较, 因此定义方差的非负平方根为 **标准差**, 标准差和原始变量的量纲是一致的, 可以直接进行比较。

现在我们来看一下均值和方差的一些性质, 首先考虑随机变量 X 的线性函数

$$Y = aX + b \quad (2.6.21)$$

其中 a 和 b 是已知常数, $f(x)$ 为随机变量 X 的概率分布函数, 关于线性函数 Y 的均值和方差, 我们有

$$\begin{aligned} \mathbb{E}[Y] &= \sum_x (ax + b) f(x) \\ &= a \sum_x x f(x) + b \sum_x f(x) \\ &= a\mathbb{E}[X] + b \end{aligned} \quad (2.6.22)$$

进一步地

$$\begin{aligned} V(Y) &= \sum_x (Y - \mathbb{E}[Y])^2 f(x) \\ &= \sum_x [ax + b - (a\mathbb{E}[X] + b)]^2 f(x) \\ &= \sum_x (ax - a\mathbb{E}[X])^2 f(x) \\ &= a^2 \sum_x (x - \mathbb{E}[X])^2 f(x) \\ &= a^2 V(X) \end{aligned} \quad (2.6.23)$$

随机变量的线性函数的均值和方差

设 X 为随机变量, 令

$$Y = aX + b \quad (2.6.24)$$

其中 a 和 b 为给定的常数, 则

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b \quad V(Y) = a^2 V(X) \quad (2.6.25)$$

特别注意, 这只对线性函数成立, 非线性函数不成立。

此外, 还有一个用矩表达方差的重要公式。

用矩表达的方差公式

$$V(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (2.6.26)$$

证明如下：

$$\begin{aligned}
 V(X) &= \sum_x (x - \mathbb{E}[X])^2 f(x) \\
 &= \sum_x (x^2 - 2x\mathbb{E}[X] + \mathbb{E}[X]^2) f(x) \\
 &= \sum_x x^2 f(x) + 2\mathbb{E}[X] \sum_x x f(x) + \mathbb{E}[X]^2 \sum_x f(x) \\
 &= \mathbb{E}[X^2] - 2\mathbb{E}[X] \times \mathbb{E}[X] + (\mathbb{E}[X])^2 \\
 &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
 &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
 \end{aligned} \quad (2.6.27)$$

2.7 边缘化

边缘化 (marginalization)

边缘化 (marginalization)，又叫边际化，它是指从多个随机变量的联合概率分布中求解出部分随机变量联合概率分布的过程。字面意思就是：在一个随机变量集合中，把其中部分随机变量“边缘化 (marginalization)”，“边缘化”这个词的目标变量是剩下的变量子集，而不是被“去掉”的那些。

假设已知随机变量 A 和 B 的联合概率分布为 $P(A, B)$ ，以及变量 A 的概率分布 $P(A)$ ，现在想求出变量 B 的概率分布 $P(B)$ 。这个过程就是把 B 进行边缘化，求出边缘概率分布 $P(B)$ 。

边缘概率分布

边缘分布 (Marginal Distribution) 是指，在多个随机变量的联合概率分布中，只包含其中 **部分变量**（可以是多个联合）的概率分布。边缘概率分布可以通过对联合概率分布在除目标变量以外的其他变量（边缘化）求和 (积分) 得到。

它的计算过程其实就是利用全概率公式把变量 A 从联合概率分布中 **消除掉**，因此可以称为 **消元法**。

$$P(B) = \sum_A P(A, B) = \sum_A P(A)P(B|A) \quad (2.7.1)$$

当已知 $P(A)$ 和 $P(A, B)$ ，可以利用边缘化的方法求得 $P(B)$ 的概率分布。即使更多的随机变量也是如此，比如，已知 4 个变量的联合概率分布 $P(A, B, C, D)$ 以及 $P(C), P(D|C)$ ，想要求 $P(A, B)$ 的概率分布，这时就需要“边缘化”随机变量 A 和 B ，也就是从 $P(A, B, C, D)$ “消除掉”随机变量 C 和 D ，从而得到 $P(A, B)$ 。

$$P(A, B) = \sum_C \sum_D P(A, B, C, D) = \sum_C \sum_D P(C)P(D|C)P(A, B|C, D) \quad (2.7.2)$$

如果是连续随机变量，就把求和换成积分。

2.8 常见概率分布

本节我们介绍一些已知的并且常用的概率分布，这些概率分布会在本书之后的章节中频繁使用，需要读者对这些分布的特性十分熟悉。

2.8.1 伯努利分布

伯努利分布是最简单的离散概率分布，伯努利分布是单次伯努利试验结果的分布。**伯努利试验 (Bernoulli experiment)** 是在同样的条件下重复地、相互独立地进行的一种随机试验，其特点是该随机试验只有两种可能结果：发生或者不发生。最简单的例子就是投掷硬币的试验，投掷硬币的结果只有正面（正面发生）和反面（正面不发生）两种结果，投硬币试验就是一种伯努利实验。

单次伯努利试验结果的概率分布就称为伯努利概率分布，服从伯努利概率分布的随机变量可以称为**伯努利变量**。假设随机变量 X 是伯努利变量，设正面向上的概率为 π ，则反面向上的概率为 $1 - \pi$ ，正面向上的结果用数字 1 表示，反面向上的结果用数字 0 表示，即

$$X = \begin{cases} 1, & \text{若正面向上} \\ 0, & \text{若反面向上} \end{cases} \quad (2.8.1)$$

它的概率分布为

$$P(X) = \begin{cases} \pi, & \text{若} X = 1 \\ 1 - \pi, & \text{若} X = 0 \end{cases} \quad (2.8.2)$$

分段函数不利于参与计算，通常伯努利变量的概率质量函数可以写成如下简单形式

$$P(X) = \pi^x (1 - \pi)^{1-x} \quad (2.8.3)$$

根据随机变量期望的计算公式，伯努利变量的期望为

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \{0,1\}} x P(X) \\ &= \sum_{x \in \{0,1\}} x \pi^x (1 - \pi)^{1-x} \\ &= \pi \end{aligned} \quad (2.8.4)$$

可以看到，对于伯努利变量来说， π 既是 $X = 1$ 的概率，也是它的期望值。我们再来看下伯努利变量的方差。

$$\begin{aligned} V(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \sum_{x \in \{0,1\}} (x - \mathbb{E}[X])^2 P(X) \\ &= \sum_{x \in \{0,1\}} (x - \pi)^2 \pi^x (1 - \pi)^{1-x} \\ &= (-\pi)^2 \pi^0 (1 - \pi)^{1-0} + (1 - \pi)^2 \pi^1 (1 - \pi)^{1-1} \\ &= \pi^2 (1 - \pi) + (1 - \pi)^2 \pi \\ &= \pi(1 - \pi) \end{aligned} \quad (2.8.5)$$

2.8.2 二项式分布

单次伯努利试验的结果分布是伯努利分布, 如果进行多次伯努利试验, 试验结果中证明向上的次数定义为随机变量则这个随机变量的概率分布是二项式分布, 注意这多次伯努利试验要求是同样的条件下重复地、相互独立地进行的。

假设我们在同样的条件下重复地、相互独立进行了 n 次伯努利实验, 单次试验成功的概率为 π , n 试验后一共成功的次数为 X , 则 X 的概率分布称为二项式分布, 一般记作 $X \sim B(n, \pi)$, 其概率质量函数为

$$P(X) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad (2.8.6)$$

变量 X 表示在 n 次试验中成功的次数, 只是一个次数累计, 对成功的位置并没有限制, 因此在 n 次试验中任意位置成功 X 次都可以, 所以上式中需要一个组合数项 $\binom{n}{x}$ 。由于是在同样的条件下重复地、相互独立的进行试验, 因此每一次试验成功的概率都是 π 。

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x=0}^n x P(X) \\ &= \sum_{x=0}^n x \binom{n}{x} \pi^x (1 - \pi)^{n-x} \\ &= \sum_{x=1}^n \frac{x n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x} \\ &= \sum_{x=1}^n \frac{n\pi(n-1)!}{(x-1)![n(n-1)-(x-1)]!} \pi^{x-1} (1 - \pi)^{(n-1)-(x-1)} \\ &= n\pi \underbrace{\sum_{x=1}^n \frac{(n-1)!}{(x-1)![n(n-1)-(x-1)]!} \pi^{x-1} (1 - \pi)^{(n-1)-(x-1)}}_{\text{相当于 } n-1 \text{ 次二项分布的累加, 等于 } 1} \\ &= n\pi \end{aligned} \quad (2.8.7)$$

二项式变量 X 表示 n 次伯努利试验中成功的次数, 它的期望值 $n\pi$ 。现在我们把它的概率分布用柱状图的形式呈现出来, 以便直观的感受其中的变化。

图 2.8.1 所示是进行 $n = 50$ 次伯努利试验时, 不同 π 值的情况下, 二项式变量 X 的概率分布图。可以看到概率最大 (最高的柱子) 的点就是期望值 $n\pi$ 的点, 随着从期望值的点向着两侧延伸概率逐渐变小。有两个特殊的情况, 当 $\pi = 0$ 时意味着成功的概率是 0, 因此图形上 $x = 0$ 的点概率是 1, 其它点 $x > 0$ 的概率是 0。反过来, 当 $\pi = 1$ 时意味着成功的概率是 1, 因此图形上 $x = n$ 的点的概率是 1, 其它点 $x < n$ 的概率是 0。

最后我们看下二项式分布的方差, 其计算过程如公式 (2.8.8) 所示, 可以看到二项式分布的方差就是伯努利分

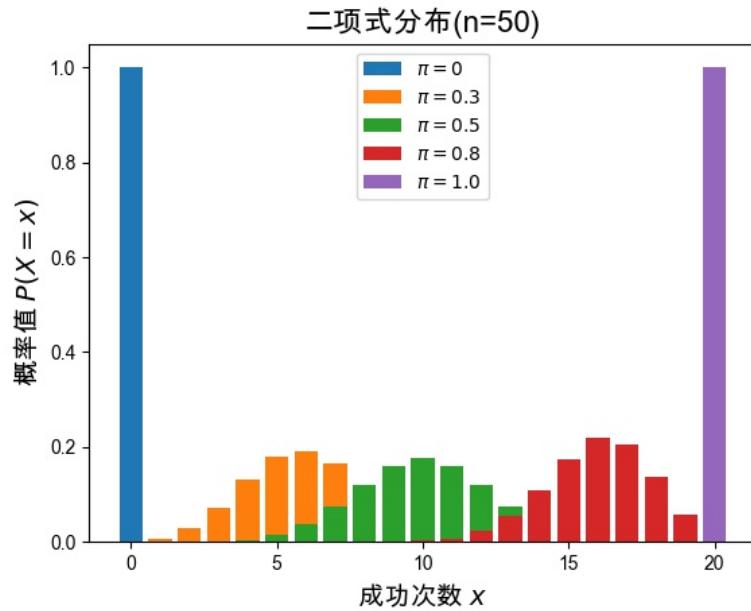


图 2.8.1: 当 $\pi = 0$ 时, $P(X = 0) = 1, P(X > 0) = 0$; 当 $\pi = 1$ 时, $P(X = n) = 1, P(X < n) = 0$ 。

布方差的 n 倍。

$$\begin{aligned}
 V(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \mathbb{E}[X(X-1) + X] - (\mathbb{E}[X])^2 \\
 &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 \\
 &= \sum_{x=0}^n \left[x(x-1) \binom{n}{x} \pi^x (1-\pi)^{n-x} \right] + n\pi + (n\pi)^2 \\
 &= \sum_{x=0}^n \underbrace{\left[x(x-1) \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x} \right]}_{x=0, x=1 \text{ 时此项为 0, 求和可以省去}} + n\pi + (n\pi)^2 \\
 &= \sum_{x=2}^n \left[n(n-1)\pi^2 \frac{(n-2)!}{(x-2)![n-2-(x-2)]!} \pi^{x-2} (1-\pi)^{(n-2)-(x-2)} \right] + n\pi + (n\pi)^2 \tag{2.8.8} \\
 &= n(n-1)\pi^2 \underbrace{\sum_{x=2}^n \left[\frac{(n-2)!}{(x-2)![n-2-(x-2)]!} \pi^{x-2} (1-\pi)^{(n-2)-(x-2)} \right]}_{\text{相当于概率分布 } B(n-2, \pi) \text{ 的累积, 其结果为 } 1} + n\pi + (n\pi)^2 \\
 &= n(n-1)\pi^2 + n\pi + (n\pi)^2 \\
 &= n^2\pi^2 - n\pi^2 + n\pi + (n\pi)^2 \\
 &= n\pi(1-\pi)
 \end{aligned}$$

2.8.3 类别分布

伯努利随机变量只有两个离散状态, 如果一个离散随机变量拥有更多的离散状态, 就称这个变量为类别变量 (categorical variable), 它的概率分布称为类别分布 (categorical distribution)。显然类别随机变量也是一个离散随机变量, 它比伯努利变量拥有更多的可能取值。

假设随机变量 X 是一个拥有 K 个可能取值的离散随机变量其取值空间记作 $\mathcal{X} = \{x_1, \dots, x_K\}$ 。变量 X 取值为 x_i 的概率记作 π_i , 类似于伯努利变量, 类别变量的概率质量函数可以用一个分段函数表示。

$$P(X) = \begin{cases} \pi_1, & \text{若 } X = x_1 \\ \pi_2, & \text{若 } X = x_2 \\ \dots, & \dots \\ \pi_K, & \text{若 } X = x_K \end{cases} \quad (2.8.9)$$

同样也需要满足概率和为 1 的约束, $\sum_{k=1}^K \pi_k = 1$ 。

分段函数的形式不够简洁, 通常会借用一个指示函数改写一下类别分布的概率质量函数, 使它的形式更利于参与到各类复杂计算中。

指示函数

定义如下函数为指示函数。

$$\mathbb{I}(x, a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{otherwise} \end{cases} \quad (2.8.10)$$

当满足 $x = a$ 时函数输出值为 1, 反之函数输出值为 0。

利用指示函数可以把公式 (2.8.9) 改写成如下更简洁的形式。

$$P(X) = \prod_{k=1}^K \pi_i^{\mathbb{I}(x, x_k)}, \quad \sum_{k=1}^K \pi_k = 1 \quad (2.8.11)$$

虽然通常会用连续的整数 $1, 2, 3, \dots, K$ 表示对应的类别 $\{x_1, x_2, x_3, \dots, x_K\}$, 但是要注意, 对于类别变量的各个类别 x_k 之间是没有任何顺序、大小关系的, 各个类别之间是独立的。因此它的期望和方差也是每个类别单独计算。

$$\begin{aligned} \mathbb{E}[X = x_k] &= \pi_k \\ V(X = x_k) &= \pi_k(1 - \pi_k) \end{aligned} \quad (2.8.12)$$

由于类别变量类别数多于 2 个, 所以还需要给出类别之间的协方差。

$$\text{Cov}(X_i, X_j) = -\pi_i \pi_j, \quad i \neq j \quad (2.8.13)$$

2.8.4 多项式分布

我们知道, 二值离散变量称为伯努利变量 (Bernoulli variable), 其概率分布称为伯努利分布 (Bernoulli distribution), 多次伯努利采样称为二项式分布 (binomial distribution), 伯努利分布是二项式分布特例, 即仅进行单次试验的情况。相对应的, 多值离散变量称为类别变量 (categorical variable), 其概率分布称为类别分布 (categorical distribution), 多次类别分布采样称为多项式分布 (multinomial distribution), 类别分布是多项式分布的特例。

我们用 M 表示变量的取值个数, 比如对于伯努利变量 $K = 2$, 用 n 表示试验次数 (采样次数):

1. 当 $K = 2, n = 1$ 时, 是伯努利分布 (Bernoulli distribution)。
2. 当 $K = 2, n > 1$ 时, 是二项式分布 (binomial distribution)。
3. 当 $K > 2, n = 1$ 时, 是类别分布 (categorical distribution)。
4. 当 $K > 2, n > 1$ 时, 是多项式分布 (multinomial distribution)。

假设随机变量 X 服从多项式分布, x_k 表示 n 次试验结果中类别 k 出现的次数。对照着二项式分布的概率质量函数, 可以直接给出多项式分布的概率质量函数。

$$P(X) = \frac{n!}{x_1!x_2!\cdots x_K!} \prod_{k=1}^K \pi_k^{x_k} \quad (2.8.14)$$

同理多项式分布的期望和方差也是每个类别单独计算的。

$$\begin{aligned} \mathbb{E}[X = x_k] &= n\pi_k \\ V(X = x_k) &= n\pi_k(1 - \pi_k) \\ Cov(X_i, X_j) &= -n\pi_i\pi_j, \quad i \neq j \end{aligned} \quad (2.8.15)$$

2.8.5 高斯分布

高斯分布 (Gaussian distribution), 以德国数学家卡尔·弗里德里希·高斯的姓冠名, 因其是日常生活中最常见的连续值概率分布, 常在自然和社会科学领域中代表一个不明的随机变量, 在统计学上十分重要, 经常又被称为正态分布 (Normal distribution)、常态分布、正规分布。

一个连续随机变量 X 若它的概率密度函数具有如下形式, 则称它为高斯随机变量或者正态随机变量, 记作 $X \sim N(\mu, \sigma^2)$ 。

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.8.16)$$

高斯随机变量的期望和方差由下式给出

$$\mathbb{E}[X] = \mu, \quad V(X) = \sigma^2 \quad (2.8.17)$$

高斯变量的概率密度函数公式 (2.8.16) 中的 μ 和 σ^2 分别对应着变量的期望和方差, σ 为标准差 (standard deviation, SD)。期望决定分布的“中心”, 方差影响着分布的“宽度”, 我们通过图形来直观的感受下。

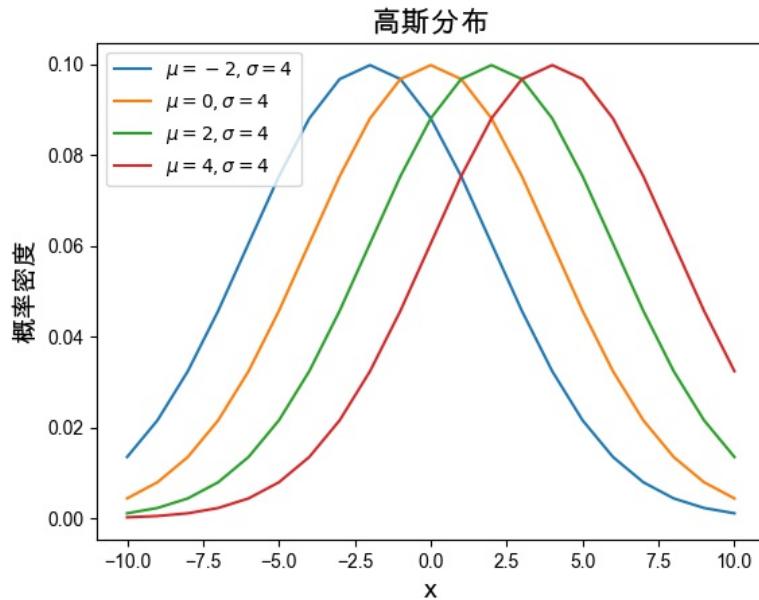
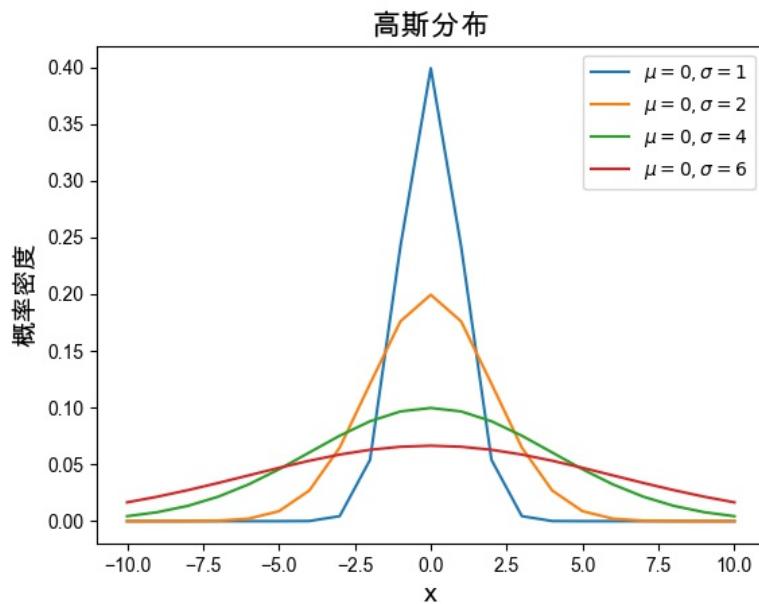
图 2.8.2 是期望参数不同取值的情况下, 高斯分布概率密度函数图形的变化, 为了凸显期望参数的影响, 我们固定方差参数的值。高斯分布的概率密度函数的曲线呈现一个重型曲线的形状, 曲线的最高点就是期望值所在的点, 显然期望值是概率最大的点。随着期望值从 -2 到 4 的变化, 钟形曲线向右发生平移, 因为曲线的中心在右移, 期望值的变化会导致概率密度函数的曲线发生平移。

接下来, 我们固定期望值参数为 0 , 观看不同的方差参数对曲线的影响, 如 图 2.8.3 所示。可以看出方差 σ^2 越小图形越“尖锐”, 反之, 方差 σ^2 越大图形越“宽胖”。

正态分布的累积分布函数 (CDF) 为

$$F_X(x; \mu, \sigma) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt \quad (2.8.18)$$

图 2.8.4 是正态分布的累积分布函数的图形, 由于均值参数 μ 只影响图形的左右平移, 不影响图形的形状, 因此我们固定了均值参数 $\mu = 0$, 观看不同标准差参数 σ 对图形曲线的影响。可以看到标准差 σ 影响着曲线的斜率, σ 越小曲线斜率越大, 斜率越大说明累积概率上升的越快, 对照着 图 2.8.3 的概率密度曲线, 不难理解这一点。

图 2.8.2: 固定标准差参数 $\sigma = 4$, 不同的期望参数 μ 下图形的变化。图 2.8.3: 固定期望参数 $\mu = 0$, 不同的标准差参数 σ 下图形的变化。

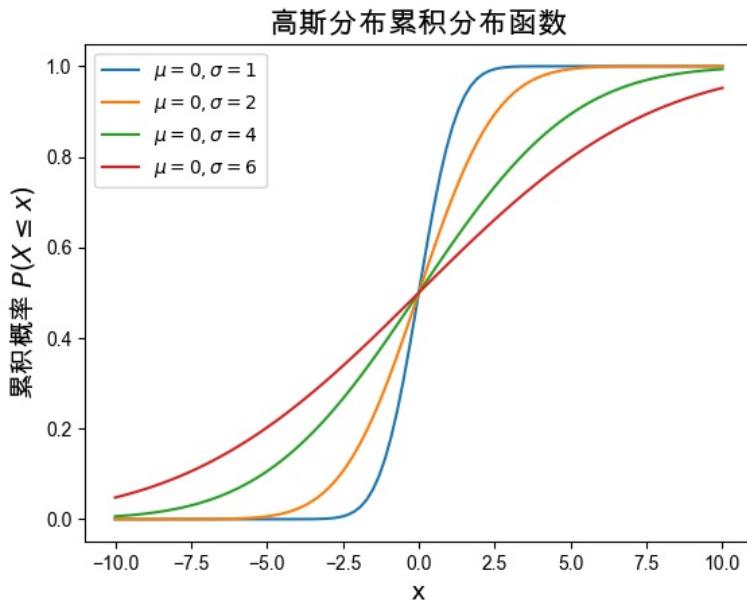


图 2.8.4: 固定均值 $\mu = 0$ 时, 不同标准差 σ 下正态累计分布函数的变化。

线性变换之下正态性不变

设 X 是正态随机变量, 其均值为 μ , 方差为 σ^2 。若 $a \neq 0$ 和 b 是两个常数, 则随机变量

$$Y = aX + b \quad (2.8.19)$$

仍然是正态随机变量, 并且其均值和方差分别为

$$\mathbb{E}[Y] = a\mu, \quad V(Y) = a^2\sigma^2 \quad (2.8.20)$$

记作 $Y \sim N(a\mu, a^2\sigma^2)$

多个独立正态随机变量之和仍然是正态变量

设 $X \sim N(\mu_X, \sigma_X^2)$ 与 $Y \sim N(\mu_Y, \sigma_Y^2)$ 是互相独立的正态随机变量, 那么

- 它们的 和 $U = X + Y$ 也满足正态分布 $U \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- 它们的 差 $V = X - Y$ 也满足正态分布 $V \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

标准正态分布

满足 $\mu = 0, \sigma^2 = 1$ 的正态分布称为 **标准正态分布**, 此时其概率密度函数变得简单

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} \quad (2.8.21)$$

标准正态分布的累积分布函数习惯上记作 Φ 。

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt \quad (2.8.22)$$

任意一个非标准的正态随机变量都可以转换成标准正态随机变量。假设一个随机变量 $X \sim N(\mu, \sigma^2)$ ，则有如下随机变量是标准正态随机变量。

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \quad (2.8.23)$$

这个转换关系一定要牢记预选，在本书之后的内容中会多次应用。

我们知道正态分布的概率密度函数曲线是一个左右对称的钟形曲线，对称的中心线就是期望值 μ 的直线，期望值附近的概率是最大的。根据连续值随机变量概率密度的定义，概率密度函数曲线和 x 轴所围成的整个区域面积为 1，我们可以把整个区域分成一些固定的子区域。如图 2.8.5 所示，以 0 点为中心，向两边延伸， $[0, 1]$ 的区间面积占整个曲线面积的 34.13%，由于曲线是对称的， $[-1, 0]$ 的区间面积也是 34.13%。同理， $[1, 2]$ 和 $[-2, -1]$ 都是 13.6%， $[2, 3]$ 和 $[-3, -2]$ 都是 2.14%。

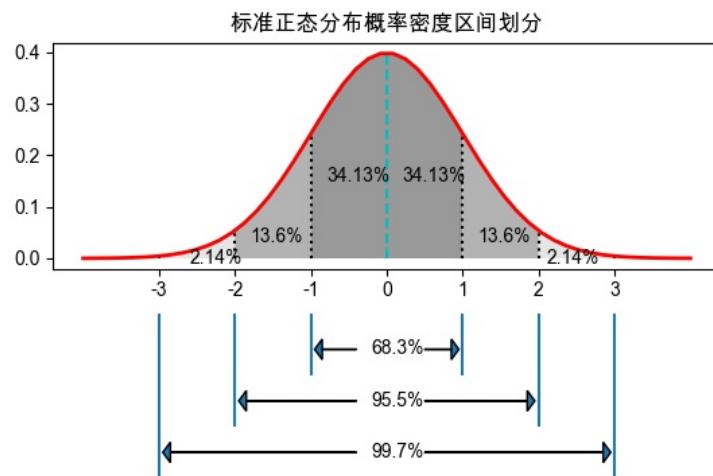


图 2.8.5: 标准正态分布 $N(0, 1)$ 概率密度函数的区间划分。

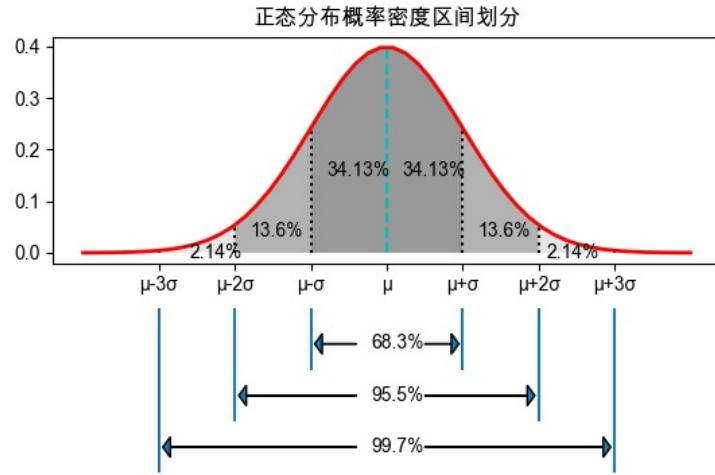
我们把中心两边合并起来， $[-1, 1]$ 的区间面积占比就是 68.3%， $[-2, 2]$ 的区间面积占比就是 95.5%， $[-3, 3]$ 的区间面积占比就是 99.7%，剩下的两端 $(-\infty, -3)$ 以及 $(3, \infty)$ 加起来不足 0.3%。这意味着，如果我们从一个服从标准正态分布 $N(0, 1)$ 的随机变量进行取样（试验），样本值（试验结果）落在

- 区间 $[-1, 1]$ 的概率是 68.3%。
- 区间 $[-2, 2]$ 的概率是 95.5%。
- 区间 $[-3, 3]$ 的概率是 99.7%。

上面是对标准正态分布概率密度函数的区间划分，那非标准正态分布是什么样的呢？其实是类似的，变化的只是子区间的范围而已。图 2.8.6 是正态分布 $N(\mu, \sigma^2)$ 的概率密度函数曲线的区间划分，和标准正态分布相比只是横轴发生了平移和缩放而已，中心的不再是 0 而是 μ ，子区间扩大了一个标准差 σ 。

- 区间 $[\mu - 1\sigma, \mu + 1\sigma]$ 的概率是 68.3%。
- 区间 $[\mu - 2\sigma, \mu + 2\sigma]$ 的概率是 95.5%。
- 区间 $[\mu - 3\sigma, \mu + 3\sigma]$ 的概率是 99.7%。

正态分布概率密度函数曲线的这个区间划分一定要理解，在之后讨论假设检验时会使用到。

图 2.8.6: 正态分布 $N(\mu, \sigma^2)$ 概率密度函数的区间划分。

2.8.6 卡方分布

高斯变量的线性变换后仍然是服从高斯分布的, 本节我们看下非线性的结果。设随机变量 Z_i 是均值为 0 方差为 1 的标准正态分布变量, 随机变量 X 是由 k 个独立的 **标准正态分布变量** 的平方和组成。

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2 = \sum_{i=1}^k Z_i^2 \quad (2.8.24)$$

则随机变量 X 被称为服从自由度 (degree of freedom, df) 为 k 的卡方分布, 记作 $X \sim \chi^2(k)$ 。卡方分布 (chi-square distribution), 也可以写作 χ^2 分布, 是一种特殊的伽玛分布, 是统计推断中应用最为广泛的概率分布之一, 例如假设检验和置信区间的计算。卡方分布的概率密度函数为:

$$f(x; k) = \frac{\frac{1}{2}^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0 \quad (2.8.25)$$

卡方分布是一个平方和, 因此变量值一定是大于等于 0 的, 对于 $x < 0$, 我们认为 $f(x) = 0$ 。公式中的符号 Γ 表示 Gamma 函数, Gamma 函数相当于阶乘函数在实数域的扩展, 有关 Gamma 函数更多的细节我们以后再讨论。卡方分布的概率密度函数看上去十分复杂, 没关系, 我们不需要记住, 需要的时候翻书就可以了。

自由度为 k 的卡方变量的平均值是和方差是分别为

$$\mathbb{E}[X] = k, \quad V(x) = 2k \quad (2.8.26)$$

现在我们来看下不同自由度下, 卡方分布概率密度函数曲线的变化, 如图 2.8.7 所示, 可以看到随着自由度的增加, 卡方分布的曲线逐步变成钟形曲线, 越来越进行正态分布。

卡方分布的累积分布函数为

$$F(x; k) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})} \quad (2.8.27)$$

其中 γ 为不完全 Γ 函数。

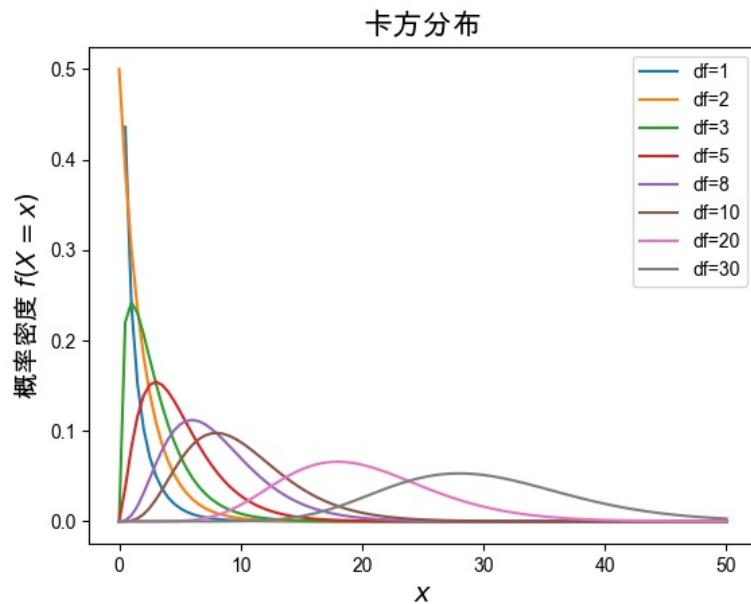


图 2.8.7: 随着自由度的增加, 卡方分布的曲线逐步变成钟形曲线。

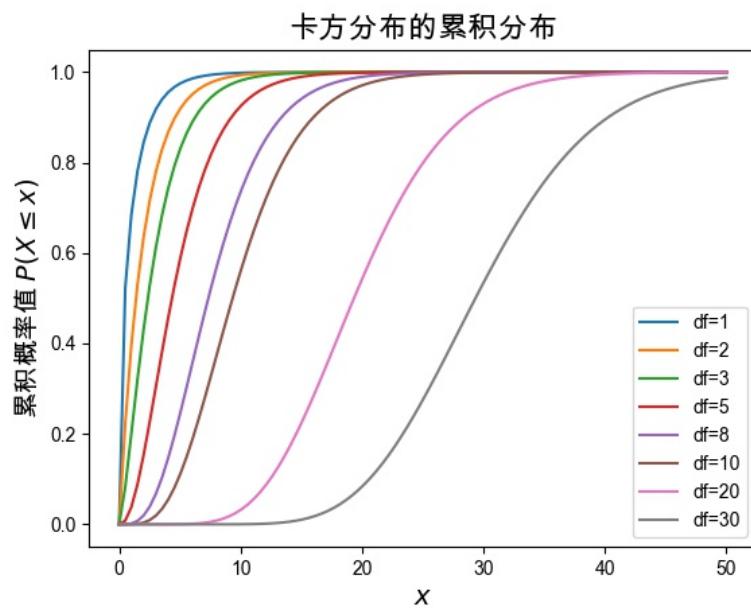


图 2.8.8: 卡方分布的累积分布函数。

可以明显看出卡方分布的自由度参数影响着其累积分布函数的斜率, 自由度越小斜率越大, 自由度越大斜率越小。

卡方分布的可加性

由卡方变量的定义可得, 独立卡方变量之和同样服从卡方分布。特别地, 若 X_1, X_2, \dots, X_n 分别独立服从自由度为 k_1, k_2, \dots, k_n 的卡方分布, 那么它们的和 $\sum_{i=1}^n X_i$ 服从自由度为 $\sum_{i=1}^n k_i$ 的卡方分布。

非中心化卡方分布

卡方分布的定义中要求是 **标准正态分布 (期望为 0, 方差为 1)** 的平方和, 那如何不是标准正态分布呢?

设随机变量 Z_i 是均值为 μ_i 方差为 1 的正态分布变量, 随机变量 X 是由 k 个独立的 **单位方差正态分布变量** 的平方和组成。

$$X = Z_1^2 + Z_2^2 + \dots + Z_k^2 = \sum_{i=1}^k Z_i^2 \quad (2.8.28)$$

则随机变量 X 被称为服从自由度为 k 的 **非中心化卡方分布 (Noncentral chi-squared distribution)**, 为了区分二者, 由标准正态分布得到的卡方分布可以称为 **中心化卡方分布**。非中心化卡方分布的概率密度函数中多了一个非中心化参数 λ 。

$$f(x; \lambda, k) = e^{-\lambda/2} \sum_{i=0}^{\infty} \frac{(\lambda/2)^k}{k!} f_{Y_{k+2i}}(x) \quad (2.8.29)$$

其中 $f_{Y_{k+2i}}(x)$ 是自由度为 $k + 2i$ 的中心卡方分布的概率密度函数。 λ 称为非中心化参数 (non-centrality parameter), 由下式给出

$$\lambda = \sum_{i=1}^k \mu_i^2 \quad (2.8.30)$$

非中心化卡方分布的概率密度函数变得异常复杂, 我们无需关注它的细节, 只需要清楚非中心化卡方分布和中心化卡方分布的区别即可。非中心化卡方分布的期望和方差为

$$\mathbb{E}[X] = k + \lambda, \quad V(x) = 2(k + 2\lambda) \quad (2.8.31)$$

2.8.7 t 分布

t 分布的推导最早由大地测量学家 Friedrich Robert Helmert 于 1876 年提出, 并由数学家 Lüroth 证明。英国人威廉·戈塞 (William S. Gosset) 于 1908 年再次发现并发表了 t 分布, 当时他还在爱尔兰都柏林的吉尼斯 (Guinness) 啤酒酿酒厂工作。酒厂虽然禁止员工发表一切与酿酒研究有关的成果, 但允许他在不提到酿酒的前提下, 以笔名发表 t 分布的发现, 所以论文使用了“学生” (Student) 这一笔名。之后 t 检定以及相关理论经由罗纳德·费希尔 (Sir Ronald Aylmer Fisher) 发扬光大, 为了感谢戈塞的功劳, 费希尔将此分布命名为学生 t 分布 (Student's t)。t 分布是标准正态分布的一个近似分布, 当不知道标准正态分布的方差时, 经常用 t 分布做标准正态分布的一个替代 (近似)。

假设 X 是呈正态分布的独立的随机变量, 它的期望值为 μ , 方差为 σ^2 , 方差未知。令 X_1, X_2, \dots, X_N 为随机变量 X 的一个独立同分布的观测样本序列, 则样本的均值为

$$\bar{X}_N = \frac{X_1 + X_2 + \dots + X_N}{N} \quad (2.8.32)$$

样本的方差为

$$S_N^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \quad (2.8.33)$$

定义如下变量

$$T = \frac{\bar{X}_N - \mu}{\frac{S_N}{\sqrt{N}}} \quad (2.8.34)$$

则变量 T 是一个随机变量, 它的分布称为 t 分布, 它的概率密度函数是

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad (2.8.35)$$

其中 $\nu = N - 1$, 参数 ν 一般称为自由度, Γ 是伽马函数。当 $\nu > 1$ 时, 它的期望是 0, $\nu = 1$ 时未定义。当 $\nu > 2$ 时, 它的方差是 $\nu/(\nu - 2)$, 否则无穷大。

t 分布作为标准正态分布的近似分布, 它的概率密度函数曲线和标准正态分布是非常接近的, 并且随着自由度的增加, 二者越来越接近, 当自由度足够大时, t 分布就等价于标准正态分布。图 2.8.9 展示了自由度分别为 1, 2, 3, 5, 10, 30 时, t 分布和标准正态分布的差异, 可以看到当自由度是 30 时, 二者已经基本重合。

从图形可以看出, 二者期望是一样的, 都是 0。当自由度小于 30 时, t 分布的方差是略大于标准正态分布的。当我们不知道标准正态分布的方差时, 就可以通过标准正态分布的采样 (观测) 样本计算一个样本方差, 利用样本方差确定一个 t 分布, 然后用这个 t 分布近似模拟原来的标准正态分布进行后续的分析使用, 当然如果你的采样样本数量超过 30 个, 直接使用样本方差作为标准正态分布的方差估计值, 然后直接使用标准正态分布 $N(0, S_N)$ 进行分析使用也是可以的, 因为此时 t 分布和标准正态分布已经没有区别了。

2.8.8 F 分布

卡方分布、t 分布和 F 分布是统计学中正态总体的三大抽样分布, 有关什么是总体分布与抽样分布, 我们在节 4.2 会详细讨论。

F 分布也是一个连续值分布, 它概率密度函数十分复杂, 本书不需要过多关注, 这里就不再给出概率密度函数函数的具体形式。我们重点关注卡方分布和 F 分布的关系。

假设 U_1 和 U_2 分别自由度为 d_1 和 d_2 的两个 独立卡方随机变量, 则如下随机变量服从 F 分布。

$$F = \frac{U_1/d_1}{U_2/d_2} \quad (2.8.36)$$

F 分布的概率密度函数有两个参数 d_1 和 d_2 , 记作 $F(d_1, d_2)$ 。F 分布的期望值为

$$\frac{d_2}{d_2 - 2}, \quad d_2 > 2 \quad (2.8.37)$$

它的方差是

$$\frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}, \quad d_2 > 4 \quad (2.8.38)$$

图 2.8.10 是 F 分布的概率密度函数曲线, 可以看出随着 d_1, d_2 的增加, F 分布会变成一个类似正态分布的曲线, 并且越来越尖锐。

到这里我们发现无论是卡方分布、t 分布、F 分布都和正态分布有关系, 实际上, 大部分概率分布都是和正态分布有关系的, 本章最后, 我们给出一张图 (图 2.8.11) 来直观感受一下。

t分布与标准正态分布对比

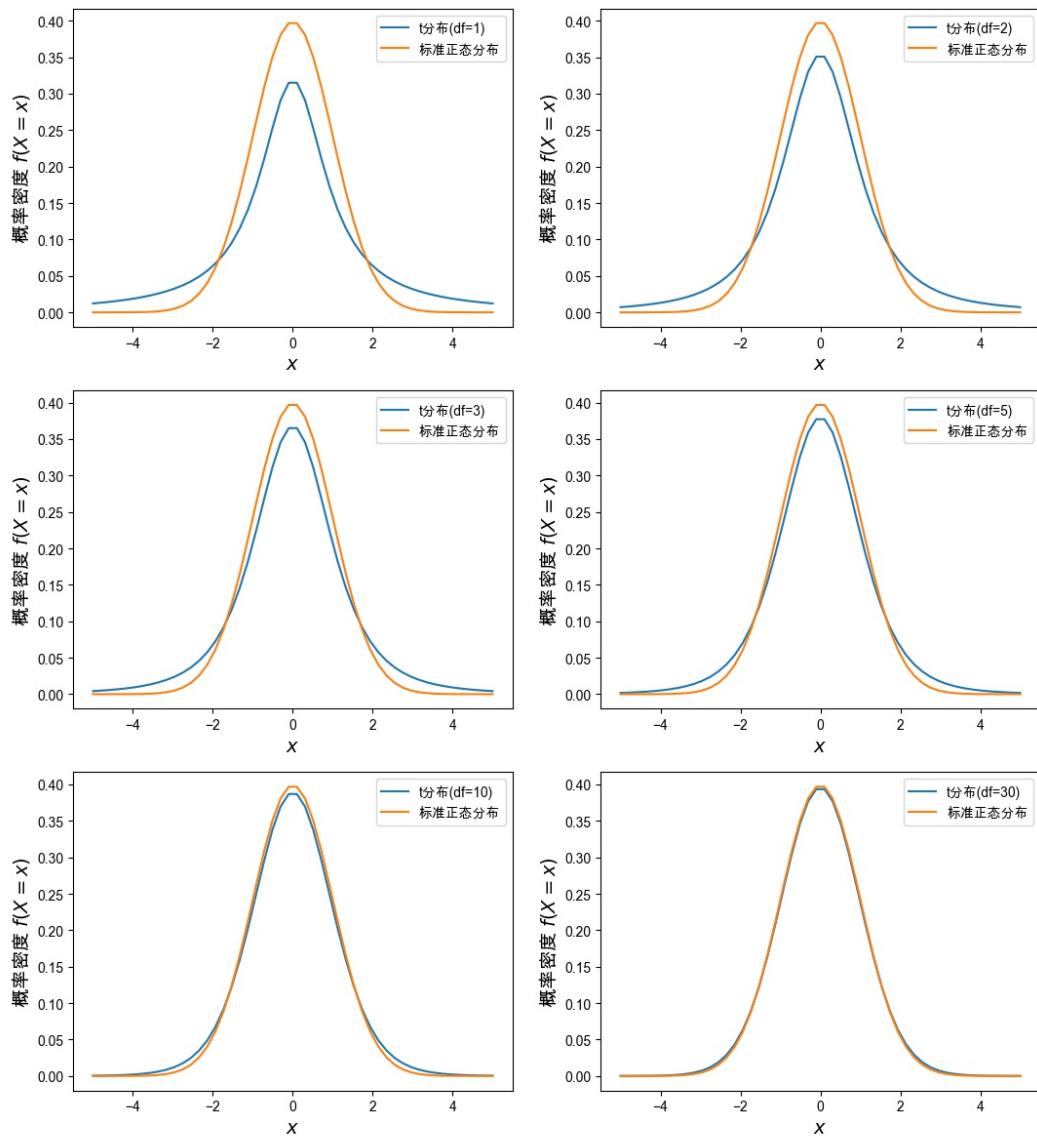


图 2.8.9: t 分布的概率密度函数和标准正态分布概率密度函数的对比。当自由度为 30 时, 二者基本重合。

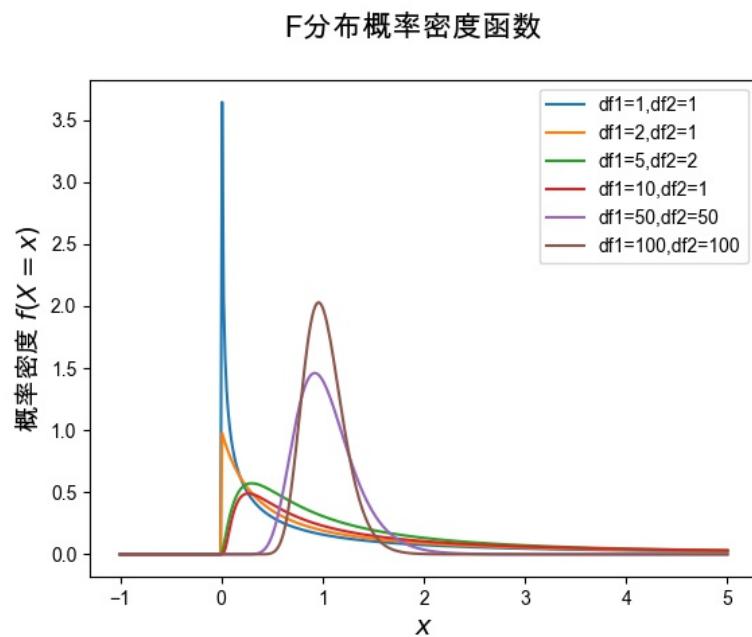
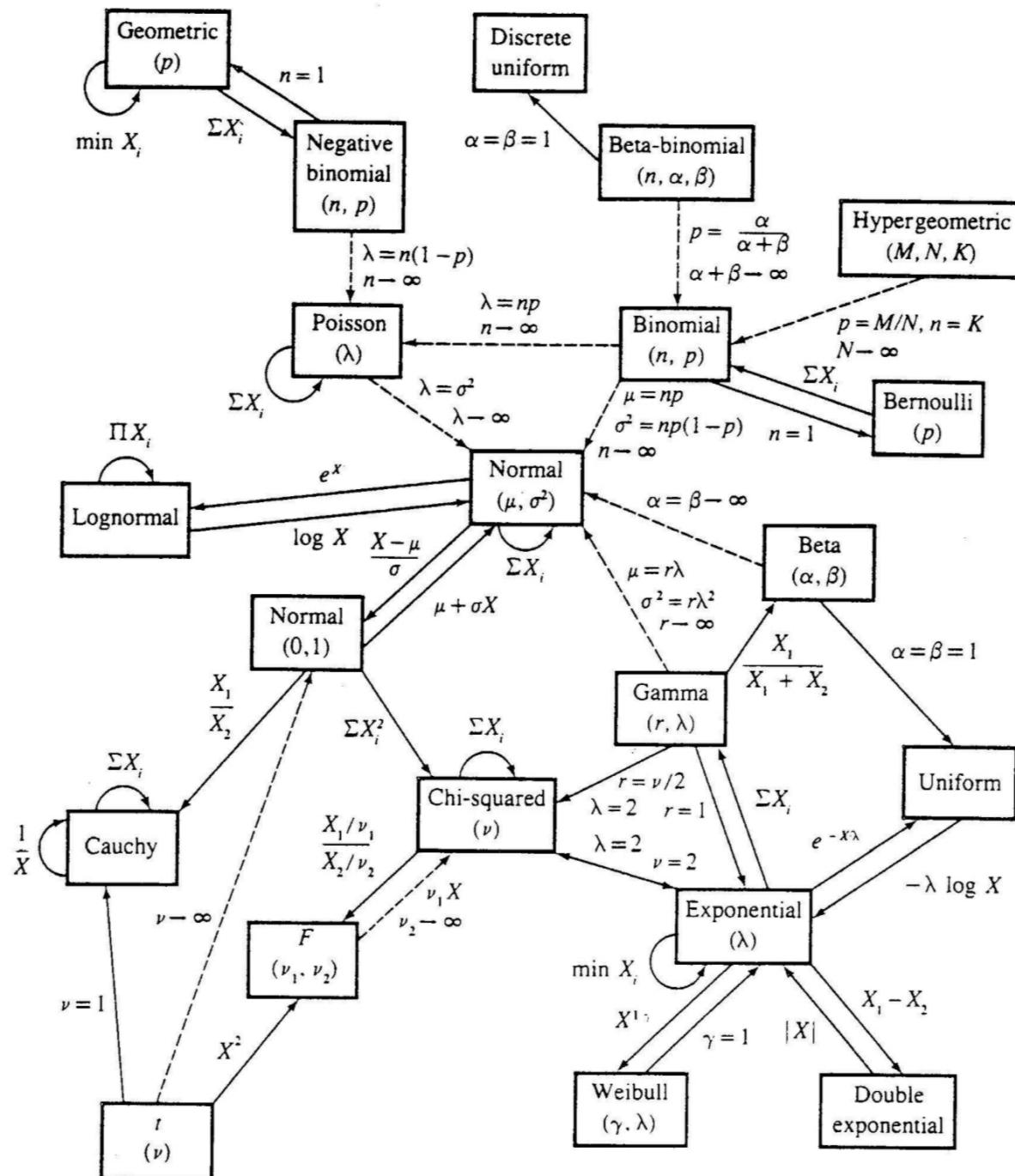


图 2.8.10: F 分布的概率密度



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

图 2.8.11: 概率分布之间的关系图

最大似然估计

在统计学中，把需要调查或者研究的某一现象或者事物的全部数据称为统计总体，或简称 **总体 (population)**。比如，我们要研究中国人的身高分布，那么全国 14 亿人的身高数据就是总体 (population)，这 14 亿身高数据所属的数据分布称为 **总体分布 (population distribution)**，其中每一个人的身高数据，即单个数据称为个体 (individual)。然而在实际中，我们不可能得到 14 亿的全部数据，也就是 **总体数据通常是无法得知的**。这时，可以选择抽样 (sampling)，即从总体当中随机抽取出部分个体，然后得到这部分抽样个体的数据，一次抽样的结果称为一份样本 (sample)。比如，从 14 亿的人群中随机抽取出 1 万的个体，然后去测量这 1 万人的身高数据，这样就得到了一份包含 1 万个数据的样本，样本的容量 (sample size)，或者说样本的大小，是 1 万。注意样本 (sample) 和个体 (individual) 的区别，样本 (sample) 是一次抽样的结果，包含多个个体 (individual) 数据，一份样本中包含的个体数据的数量称为本容量 (sample size)。通常我们会假设总体分布服从某种已知的概率分布，但是分布的某些参数是不确定的，比如全国身高数据服从正态分布，但是期望和方差不知道，这时我们期望能通过样本推断 (估计) 出总体正态分布的期望和方差参数。

推断统计学 (或称统计推断，英语：statistical inference)，指统计学中，研究如何根据样本 (sample) 数据去推断总体 (population) 特征 (或者参数) 的方法，比如根据样本的平均值去估计总体的均值参数。它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断。更概括地说，是在一段有限的时间内，通过对一个随机过程的观察来进行推断的。在统计学中，利用样本推断 (估计) 总体分布参数方法有很多，比如矩估计、最大似然估计、贝叶斯估计等等，本章我们讨论其中应用最为广泛的最大似然估计算法。

3.1 最大似然估计

最大似然估计 (Maximum Likelihood Estimation, MLE)，又叫极大似然估计，是统计学中应用最广泛的一种未知参数估计方法。它可以在已知随机变量属于哪种概率分布的前提下，利用随机变量的一些观测值估计出分布的一些参数值。所谓观测值，就是随机变量的采样值，也就是这个随机变量试验的真实结果值，因为是我们能“看到”的值，所以称为观测值。

假设有一个离散随机变量 X ，其概率质量函数是 $P(X; \theta)$ ，其中 θ 是这个概率分布的参数，其值是未知的。函数 $P(X; \theta)$ 本身是已知的，也就是我们知道 X 所属何种概率分布，比如是高斯分布等等。

现在假设我们有一些变量 X 的观测值，这些观测值集合用符号 $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^N\}$ 表示。这些观测值都是从同一个概率分布 $P(X; \theta)$ 得到的，并且这些样本是独立获取的，即每条样本值不依赖其它样本值，我们可以称这些样本是 **独立同分布的**。

独立同分布

在概率论与统计学中, 独立同分布 (英语: Independent and identically distributed, 或称独立同分配, 缩写为 iid、i.i.d.、IID) 是指一组随机变量中每个变量的概率分布都相同, 且这些随机变量互相独立。

关于样本集的理解

一个随机变量的观测样本集 $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^N\}$ 可以看做是对同一个随机变量独立的采样 (试验) N 次得到的。也可以看做是有 N 个一模一样 (相同的概率分布) 的随机变量 X , 每个独立取样一次得到总共 N 条观测样本。在统计推断的讨论中, 多数情况下会解释成第二种, 这两种理解方法是等价, 不管哪种理解方法, 这个样本集都是满足 **独立同分布** 的。

我们知道其中任意一条样本 x_i 的发生概率是 $P(x_i; \theta)$, 那么所有样本发生的联合概率是 $P(\mathcal{D}; \theta) = P(x^{(1)}, \dots, x^{(N)}; \theta)$, 又由于所有样本是满足独立同分布的 (i.i.d) 的, 根据联合概率分布的分解法则有

$$P(\mathcal{D}; \theta) = P(x^{(1)}, \dots, x^{(N)}; \theta) = \prod_{i=1}^N P(x_i; \theta) \quad (3.1.1)$$

假设 θ 的可能取值空间为 Θ , 记作 $\theta \in \Theta$ 。不论 θ 取何值, 都有一定的可能 (概率) 产生出这个样本集 \mathcal{D} , 但显然 θ 的值会影响着这个样本的产生概率 $P(\mathcal{D}; \theta)$ 。换句话说就是, 不同的 θ 值会得到不同的样本联合概率 $P(\mathcal{D}; \theta)$ 。

现在我们思考 θ 真实值是什么。事实上, 我们根本无从得知参数 θ 的真实值。但我们可以换个思路, 我们可以从 θ 的取值空间 Θ 中挑一个最好的的出来。那么什么是最好的, 这个最好的标准是什么?

最大可能性

常识告诉我们, 概率越大的事情越容易发生, 概率越小的事情越不容易发生。观测样本集的发生概率 $P(\mathcal{D}; \theta)$ 越大, 我们就越容易见到我们现在看到的样本。既然现在这个样本集 \mathcal{D} 已经真实的发生了 (我们观测到了), 是不是可以认为这个样本集的 $P(\mathcal{D}; \theta)$ 概率是最大的, 使者 $P(\mathcal{D}; \theta)$ 最大的 θ 是最优的选择呢?

在概率统计中, 把观测样本的联合概率称为 **似然 (likelihood)**, 一般用符号 $L(\theta; \mathcal{D}) = P(\mathcal{D}; \theta)$ 表示, 有时也称为似然函数 (likelihood function)。

最大似然估计非标准定义

观测样本集的似然 (联合概率) 取得最大值时参数的值作为参数估计值的方法称为最大似然估计。

观测样本集的似然函数就是样本集的联合概率

$$L(\theta; \mathcal{D}) = P(\mathcal{D}; \theta) = \prod_{i=1}^N P(x_i; \theta) \quad (3.1.2)$$

最优的 θ 值是令观测样本发生概率最大的值, 也就是令似然函数取得最大。参数 θ 的最大似然估计值可以写为

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta; \mathcal{D}) = \arg \max_{\theta} \prod_{i=1}^N P(x_i; \theta) \quad (3.1.3)$$

仔细观察后发现, 似然函数是每条样本概率 $P(x_i; \theta)$ 的连乘, 而概率值都是在 $[0, 1]$ 之间的, 一系列小于 1 的数字连乘会趋近于 0。而计算机在处理浮点数时存在精度问题, 太小的值是无法表示的。所以一般我们会为似然函数加上一个对数操作来解决计算机的精度问题, 我们把加了对数的似然函数称为 **对数似然函数 (log-likelihood function)**, 一般用符号 ℓ 表示。

$$\ell(\theta; \mathcal{D}) = \log L(\theta; \mathcal{D}) \quad (3.1.4)$$

通过极大化对数似然函数 $\ell(\theta; \mathcal{D})$ 得到 $\hat{\theta}$ 和极大化似然函数 $L(\theta; \mathcal{D})$ 是等价的, 这里不再证明, 有兴趣的读者可以参考其他资料。

$$\begin{aligned} \hat{\theta}_{ML} &= \arg \max_{\theta} \ell(\theta; \mathcal{D}) \\ &= \arg \max_{\theta} \log \prod_{i=1}^N P(x_i; \theta) \\ &= \arg \max_{\theta} \sum_{i=1}^N \log P(x_i; \theta) \end{aligned} \quad (3.1.5)$$

虽然这里我们是以离散随机变量为例, 但最大似然估计同样可以应用于连续值随机变量的参数估计。连续值随机变量用的是概率密度函数表示其每个状态的概率大小情况, 概率密度函数表示是每一个点的“密度”, 而不是概率值, 但每个点的密度是和它的概率呈正比的。假设连续值随机变量 X 的概率密度函数是 $f(x; \theta)$, 则有

$$P(X = x; \theta) \propto f(X = x; \theta) \quad (3.1.6)$$

最大似然估计是通过极大化对数似然函数求解, 对于连续值随机变量用概率密度函数 $f(X = x; \theta)$ 替换 $P(X = x; \theta)$, 对极大化求解没有任何影响。因此在使用最大似然估计概率模型的分布时, 如果是离散随机变量就用概率质量函数, 如果是连续值随机变量就是概率密度函数。

那么如何进行极大化求解呢? 通常有如下三种方法:

1. **解析法 (Analytic)**, 又叫直接求解法。我们知道一个函数在取得极值时其一阶导数是为 0 的, 因此可以通过令对数似然函数的一阶导数为 0 得到一个方程等式, 然后解这个方程得到 $\hat{\theta}_{ML}$ 。这种方法得到的解称为解析解。

$$\frac{\partial \ell}{\partial \theta} = 0 \quad (3.1.7)$$

函数的一阶导数为 0 的点称为“驻点” (stationary point), 可能为 (局部) 极大或者极小值点, 也可能为鞍点 (saddle point), 可以通过极值点的二阶导数判断是极大值点还是极小值点。并不是所有情况都能得到解析解的, 很多时候是无法直接求得的, 在后面的章节中我们会详细讨论。

2. **网格搜索法 (Grid Search)**。如果我们知道 $\hat{\theta}$ 的值在空间 Θ 中, 可以对这个空间进行搜索来得到似然函数最大的参数值。换句话说, 就是尝试这个空间中的每个值, 找到令似然函数取得最大的参数值。网格搜索方法是一种很好的方法, 它表明可以通过重复逼近和迭代来找到似然函数的最大值。但是, 它在大多数情况下不切实际, 并且当参数数量变多时变得更加困难。
3. **数值法 (Numerical)**。这是现在最常用的算法。本质上就是先为 θ 赋予一个初始值, 然后利用爬山法找到最优解。梯度下降 (上升) 法 (Gradient descent), 牛顿法 (Newton-Raphson), BHHH, DFP 等等都属于这类。

本章我们只使用解析法求解, 在正式讲广义线性模型时再介绍最大似然估计的数值求解法, 下面介绍几个应用最大似然估计的具体例子。

3.2 伯努利分布

假设一个随机变量 X 服从伯努利分布 (Bernoulli distribution), 即只有两种可能的取值 $X \in \{0, 1\}$, 设其取值为 1 的概率为 $P(X = 1) = \theta$, 其概率质量函数为

$$P(X; \theta) = \theta^x (1 - \theta)^{1-x}, x \in \{0, 1\} \quad (3.2.1)$$

其中 θ 是未知的参数, 需要使用最大似然估计得到。假设变量 X 的独立同分布的观测样本集为 $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$, 样本集的规模为 $|\mathcal{D}| = N$ 。

现在我们利用最大似然估计法估计出参数 θ , 首先写出观测样本的对数似然函数。

$$\begin{aligned} \ell(\theta; \mathcal{D}) &= \sum_{i=1}^N \log P(x_i; \theta) \\ &= \sum_{i=1}^N \log[\theta^{x_i} (1 - \theta)^{1-x_i}] \\ &= \sum_{i=1}^N \log \theta^{x_i} + \sum_{i=1}^N \log(1 - \theta)^{1-x_i} \end{aligned} \quad (3.2.2)$$

为方便表述, 我们定义几个统计值, 在样本集 \mathcal{D} 中, n_0 表示随机变量 $X = 0$ 在观测样本中的次数, $\hat{p}_{\mathcal{D}}(0) = \frac{n_0}{N}$ 表示 $X = 0$ 在观测集中出现的相对频率 (经验分数); n_1 表示 $X = 1$ 在样本中的次数, $\hat{p}_{\mathcal{D}}(1) = \frac{n_1}{N}$ 表示 $X = 1$ 在样本中出现的相对频率 (经验分数)。把 n_0, n_1 代入到对数似然函数中可得

$$\begin{aligned} \ell(\theta; \mathcal{D}) &= \sum_{i=1}^N \log \theta^{x_i} + \sum_{i=1}^N \log(1 - \theta)^{1-x_i} \\ &= \sum_{i=1}^N x_i \log \theta + \sum_{i=1}^N (1 - x_i) \log(1 - \theta) \\ &= n_1 \log \theta + n_0 \log(1 - \theta) \end{aligned} \quad (3.2.3)$$

我们知道当对数似然函数的导数为 0 时, 函数取得极值。现在对对数似然函数求导, 并令导数为 0。

$$\begin{aligned} 0 &= \frac{\partial \ell}{\partial \theta} \\ &= \frac{n_1}{\theta} - \frac{n_0}{1 - \theta} \end{aligned} \quad (3.2.4)$$

又由于 $n_0 = N - n_1$, 代入上式可得:

$$\frac{n_1}{N - n_1} = \frac{\theta}{1 - \theta} \quad (3.2.5)$$

化简可得 θ 的估计值

$$\hat{\theta}_{ML} = \frac{n_1}{N} \quad (3.2.6)$$

3.3 类别分布

现在假设随机变量 X 是类别变量, 其取值空间是 $\mathcal{X} = \{x_1, \dots, x_K\}$, 类别分布的概率质量函数可以写成下面的形式

$$P(X; \theta) = \prod_{k=1}^K \theta_k^{\mathbb{I}(x, x_k)} \quad (3.3.1)$$

其中 $\theta = [\theta_1, \dots, \theta_K]$ 为分布的参数。公式 (3.3.1) 的含义是, 变量 X 取值为类别 x_k 的概率是 θ_k , 即 $P(X = x_k; \theta) = \theta_k$ 。其中参数向量 θ 需要满足约束:

$$\sum_{k=1}^K \theta_k = 1, \theta_k \in [0, 1] \quad (3.3.2)$$

现在利用最大似然估计估计出参数向量 θ 的值, 我们继续用符号 \mathcal{D} 表示观测样本集, $|\mathcal{D}| = N$, 则似然函数为

$$\begin{aligned} L(\theta; \mathcal{D}) &= \prod_{i=1}^N P(x_i; \theta) \\ &= \prod_{i=1}^N \prod_{k=1}^K \theta_k^{\mathbb{I}(x_i, x_k)} \\ &= \prod_{k=1}^M \theta_k^{n_k} \end{aligned} \quad (3.3.3)$$

其中 n_k 表示类别 x_k 在样本中出现的次数, 那么有 $\sum_{k=1}^K n_k = N$ 。我们为似然函数加上对数操作, 以便把连乘符号转换成加法。

$$\begin{aligned} \ell(\theta; \mathcal{D}) &= \log L(\theta; \mathcal{D}) \\ &= \sum_{k=1}^K n_k \log \theta_k \end{aligned} \quad (3.3.4)$$

为了找到 θ_k 的最大似然解, 我们需要最大化对数似然函数 $\ell(\theta; \mathcal{D})$, 并且限制 θ_k 的和必须等于 1。这样带有约束的优化问题需要借用拉格朗日乘数 λ 实现, 即需要最大化如下方程。

$$h(\theta) = \sum_{k=1}^K n_k \log \theta_k + \lambda \left(\sum_{k=1}^K \theta_k - 1 \right) \quad (3.3.5)$$

对上述公式求偏导, 并令偏导等于 0。

$$\begin{aligned} 0 &= \frac{\partial h(\theta)}{\partial \theta_k} \\ &= \frac{n_k}{\theta_k} + \lambda \\ \theta_k &= -\frac{n_k}{\lambda} \end{aligned} \quad (3.3.6)$$

把 $\theta_k = -\frac{n_k}{\lambda}$ 代入到约束条件 $\sum_{k=1}^K \theta_k = 1$ 中, 解得 $\lambda = -N$, 最终可求得 θ_k 的值:

$$\hat{\theta}_k = \frac{n_k}{N} \quad (3.3.7)$$

我们发现, 伯努利分布与类别分布的参数最大似然估计值具有相同的形式, 并且最大似然估计值是可以通过样本统计的得到, 这是最大似然估计的一个特性。样本的统计值被称为统计量 (statistic), 统计量 (statistic) 是样本的一个函数, 其代表着从样本中提取的一些“信息”, 比如样本的均值 (mean), 样本的总和 (sum) 等等。很多时候这些信息可以用于确定这个分布的未知参数, 如果仅需要一个统计量就能确定这个分布的未知参数, 而不再需要其它的额外“信息”, 那么这个统计量就称为这个分布 (或者分布族) 的 充分统计量 (sufficient statistic), 在后面的章节中我们会详细讨论充分统计量。

3.4 高斯分布

现在我们假设有一个高斯随机变量 $X \sim N(\mu, \sigma^2)$ ，其参数有两个，均值 μ 和方差 σ^2 。观测样本集为 $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^N\}$ ，我们利用最大似然估计出参数 μ, σ^2 ，首先写出高斯分布的概率密度函数。

$$f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (3.4.1)$$

似然函数为

$$L(\mu, \sigma^2; \mathcal{D}) = \prod_{i=1}^N f(x_i; \mu, \sigma^2) \quad (3.4.2)$$

高斯分布的概率密度函数中有自然常数 e 的指数，因此其对数似然函数，我们选择以常数 e 为底数的对数。

$$\begin{aligned} \ell(\mu, \sigma^2; \mathcal{D}) &= \ln \prod_{i=1}^N f(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^N \ln f(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^N \left[-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\ &= \sum_{i=1}^N \left[-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\ &= -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 \end{aligned} \quad (3.4.3)$$

然后对参数求偏导数，并令偏导数为 0。

$$\begin{cases} \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu) = 0 \\ \frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N (x_i - \mu)^2 = 0 \end{cases} \quad (3.4.4)$$

由第一个等式可以解的：

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x} \quad (3.4.5)$$

其中 \bar{x} 表示样本的平均值，然后代入第二个等式解得：

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (3.4.6)$$

至此，我们就得到了高斯分布的均值参数和方差参数的最大似然估计值。仔细观察可以发现，参数的估计值只依赖两个观测样本的统计量。

$$\sum_{i=1}^N x_i, \sum_{i=1}^N (x_i - \bar{x}) \quad (3.4.7)$$

这两个量被称为高斯分布的充分统计量 (sufficient statistics)，如果你对统计量和充分统计量不理解，没关系，下一章我们会详细讨论。

3.5 总结

最后我们来总结下用最大似然方法估计参数的一般模式。

令 X 表示随机变量, 令 $f(X; \theta)$ 为随机变量 X 的概率质量 (密度) 函数的参数化表示, 其中 θ 是未知参数向量 (注意 θ 表示是多个参数的集合, 不一定只有一个未知参数)。我们把 θ 看做是一个 **非随机变量**, 是一系列数值变量, 但是其具体取值却是未知的。

我们给出这个变量的独立同分布观测样本集, $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$, 上标代表样本编号。目标是学习出参数 θ 。一个在给定观测时学习未知参数的方法是使用最大似然估计, 也叫极大似然估计。

- $f(\cdot; \theta)$ 是一个参数为 θ 的概率分布的概率质量函数 (pdf) 或者概率密度函数 (pmf)。
- $L(\cdot; \mathcal{D}) \triangleq P(\mathcal{D}; \cdot)$ 是随机变量 X 的观测数据集 \mathcal{D} 的似然函数 (likelihood function)。

最大似然估计 (maximum likelihood estimation, MLE) 最直接的理解是: 当使得给定观测数据的似然函数取得最大值时, 此时似然函数中未知参数的取值是最优的。

似然函数的概率解释是: 这些观测值 (观测值是可观测到的随机变量的取值)“同时 (不是时间上的同时, 是联合发生)”发生的概率。最大似然就是: 既然这些样本事件已经发生了, 那么我就假设他们的发生的概率是最大的, 就认为使得这些样本具有最大发生概率时参数的值是最优取值。注意, 最大似然估计的解不一定存在, 也不一定唯一, 这取决于似然函数是否有极值点, 以及有几个极值点。

观测样本集的对数似然函数是

$$\begin{aligned}
 \ell(\theta; \mathcal{D}) &= \log L(\theta; \mathcal{D}) \\
 &= \log P(x^{(1)}, \dots, x^{(N)}; \theta) \\
 &= \log \prod_{i=1}^N f(x_i; \theta) \\
 &= \sum_{i=1}^N \log f(x_i; \theta)
 \end{aligned} \tag{3.5.1}$$

然后通过极大化对数似然函数的方式求解参数的值。当然有些时候可以通过令偏导数为 0 直接求得解析解, 然而有时候却无法得到解析解, 需要用梯度迭代的方法求解。

$$\hat{\theta}_{ML} = \arg \max_{\theta} \ell(\theta; \mathcal{D}) \tag{3.5.2}$$

虽然最大似然估计使用十分广泛, 但是它不是完美的, 在样本较少时, 或者样本有偏时, 得到的估计值偏差较大。例如投掷一个普通的硬币 3 次, 每次都是正面朝上, 这时最大似然估计正面朝上的概率时结论会是 1, 表示所有未来的投掷结果都是正面向上。这明显是有问题的, 当数据集较少时非常容易出现错误的结果, 当然也不是没有解决办法, 比如可以使用贝叶斯估计, 贝叶斯估计可以算是最大似然估计的升级版, 通过增加先验的方式解决这种极端的场景。有兴趣的读者可以参考其他资料了解贝叶斯估计。通过这个小例子让我们意识到, 使用最大似然估计的得到的参数估计值未必是符合我们心意的, 这个估计值到底行不行, 我们需要一种手段来评价参数估计值的好坏, 下一章我们讨论如何评价最大似然估计值的好坏。

推荐与检验

上一章我们介绍了统计推断中应用最广泛的最大似然估计，然而当我们得到一个参数估计值后，我们期望知道这个估计值靠不靠谱，它与参数的真实值又相差多少，本章我们讨论如何评价一个参数估计值的好坏。在正式讨论估计值评价方法之前，需要先熟悉统计推断中一些基本知识，比如充分统计量、费歇尔信息、抽样分布等等，要理解后面的内容需要对这些基础概念十分熟悉才行，希望读者能花些精力理解这些基础知识，如果仅靠本书的内容还不能理解，请辅助参考其它概率与统计学的资料。

4.1 统计量和充分统计量

我们首先讨论统计量以及充分统计量的概念。

假设有一个独立同分布的观测样本集 $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ ，样本集中的样本都是从同一个概率分布 $P(X; \theta)$ 采样得到，其中 θ 是这个分布的未知参数，参数空间为 Θ 。上一章已经讲过，我们可以使用最大似然估计参数 θ ，并且参数的估计值是一个关于样本的函数 $\hat{\theta} = g(\mathcal{D})$ 。在统计学中，把观测样本的函数称为 **统计量**。

统计量 正式地，任意观测样本的实值函数 $T = g(\mathcal{D})$ 都称为一个 **统计量 (statistic)**。一个统计量就是一个关于样本集的函数（允许是向量形式的函数），**在这个函数中不能有任何未知参数**。比如，样本的均值 $\bar{x} = \frac{1}{N} \sum_i^N x_i$ ，最大值 $\max(\mathcal{D})$ ，中位数 $\text{median}(\mathcal{D})$ 以及 $f(\mathcal{D}) = 4$ 都是统计量。但是 $x_1 + \mu$ （ μ 是未知参数）就不是统计量。

参数估计值是统计量

在进行参数估计时，我们能利用的只有观测样本集，因此观测样本是我们进行参数估计的唯一信息源。也就是说，我们能利用的有关参数的所有可用信息都包含在观察样本中。因此，我们获得的参数估计量始终是观测值的函数，即参数估计量是统计量。从某种意义上讲，该过程可以被认为是“压缩”原始观察数据：最初我们有 N 个数字，但是经过这个“压缩”之后，我们只有 1 个数字。这种“压缩”总是使我们失去有关该参数的信息，决不能使我们获得更多的信息。最好的情况是，该“压缩”结果包含的信息量与 N 个观测值中包含的信息量相同，也就是该“压缩”结果包含的信息量已经是关于参数的信息的全部。

统计量是随机变量

观测样本集 \mathcal{D} 可以看成是 N 个服从相同概率分布的随机变量的独立采样，每重新进行一次采样，都会得到不同的样本集，也就是说样本集 \mathcal{D} 本身也是随机（不同采样得到不一样的值）的，因此样本集 \mathcal{D} 可以看做

是由 N 个随机变量组成, 记作 $\mathcal{D} = \{X_1, X_2, \dots, X_N\}$, 大写表示随机变量。统计量作为样本集 \mathcal{D} 的函数, 也就相当于是 N 个随机变量的函数。

$$T = g(\mathcal{D}) = g(X_1, X_2, \dots, X_N) \quad (4.1.1)$$

在第一章我们就讲过, 随机变量的函数仍然是一个随机量, 那么作为样本函数的统计量自然就是随机变量, 而参数估计值是统计量, 因此参数估计值是一个随机变量, 所以有时可以称参数估计值为参数估计量。

充分统计量

假设有一个统计量 $T(\mathcal{D})$, 并且 t 是 T 的一个特定值, 如果在给定 $T = t$ 的条件下, 我们就能计算出样本的联合概率 $P(X_1, X_2, \dots, X_N | T = t)$, 而不再依赖参数 θ , 这个统计量就是 **充分统计量 (sufficient statistic)**。

换种说法, 在给定充分统计量 $T = t$ 条件下, 就能确定参数 θ 的值, 而不再需要额外的信息, 我们可以设想只保留 T 并丢弃所有 X_i , 而不会丢失参数的任何信息! 从上面的直观分析中, 我们可以看到充分统计量“吸收”了样本中包含的有关 θ 的所有可用信息。这个概念是 R.A. Fisher 在 1922 年提出的。

充分性的概念是为了回答以下问题而提出的: 是否存在一个统计量, 即函数 $T(X_1, \dots, X_N)$, 其中包含样本中有关 θ 的所有信息? 如果这样, 则可以将原始数据减少或压缩到该统计信息而不会丢失信息。例如, 考虑一系列成功概率未知的独立伯努利试验。我们可能有一种直觉的感觉, 成功次数包含样本中有关 θ 的所有信息, 而成功发生的顺序没有提供有关 θ 的任何其他信息。对于高斯分布, (样本) 期望和 (样本) 协方差矩阵就是它的充分统计量, 因为如果这两个参数已知, 就可以唯一确定一个高斯分布, 而对于高斯分布的其他统计量, 例如振幅、高阶矩等在这种时候都是多余的。

示例:

假设 X_1, X_2, \dots, X_N 是 N 个独立同分布伯努利变量的采样样本, 其中 $P(X_i = 1) = \theta$ 。我们将验证 $T = \sum_{i=1}^N X_i$ 是 θ 的一个充分统计量。

证明:

由于 X_i 只能取值为 0 或者 1, 所以 $T = t$ 可以看作是在 N 条样本中 $X_i = 1$ 的次数。根据贝叶斯定理有:

$$\begin{aligned} P(X_1, X_2, \dots, X_N | T = t) &= \frac{P(X_1, \dots, X_N)}{P(T = t)} \\ &= \frac{\prod_i \theta^{X_i} (1 - \theta)^{(1 - X_i)}}{P(T = t)} \\ &= \frac{\theta^t (1 - \theta)^{N-t}}{P(T = t)} \end{aligned} \quad (4.1.2)$$

现在看分母部分, T 的含义是在 N 次试验中 1 的数量, 很明显这是二项式分布, 有 N 次试验, 单次成功(为 1) 的概率为 θ , 一共成功 t 次(1 的数量为 t) 的概率分布为 $T = \binom{N}{t} \theta^t (1 - \theta)^{N-t}$, 其中 $\binom{N}{t}$ 是组合数, 从 N 个结果中任意选出 t 个的方法数。把分母代入上式:

$$P(X_1, X_2, \dots, X_N | T = t) = \frac{\theta^t (1 - \theta)^{N-t}}{\binom{N}{t} \theta^t (1 - \theta)^{N-t}} = \frac{1}{\binom{N}{t}} \quad (4.1.3)$$

最终发现, 样本在给定 $T = t$ 的条件下的联合概率与参数 θ 无关, 也就是说在确定了 T 之后, 就可以直接得到样本的联合概率, 而不再依赖参数 θ 。

在很多问题中, **参数的最大似然估计量就是一个充分统计量**, 比如, 伯努利实验的参数估计量就是一个充分统计量 $\hat{\theta}_{ML} = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X}$ 。同样, 贝叶斯参数估计量也是一个充分统计量。最大似然估计量和贝叶斯估计量都是充分统计量的一个函数, 它们“吸收”了观测样本中关于参数的所有有用信息。

注解: 根据统计量的定义: 样本的一个函数可以称为统计量, 样本的求和 $\sum_{i=1}^N X_i$, 样本的均值 $\frac{1}{N} \sum_{i=1}^N X_i$ 都可以称为统计量。所以, 似然估计量 $\hat{\theta}_{ML} = \frac{1}{N} \sum_{i=1}^N X_i$ 可以整体看做一个充分统计量 (样本均值统计量), 也可以看做是充分统计量 (求和统计量) $\sum_{i=1}^N X_i$ 的一个函数。

4.2 抽样分布

在统计学中, 把需要调查或者研究的某一现象或者事物的全部数据称为统计总体, 或简称 **总体 (population)**。比如, 我们要研究中国人的身高分布, 那么全国 14 亿人的身高数据就是总体 (population), 这 14 亿身高数据所属的数据分布称为 **总体分布 (population distribution)**, 其中每一个人的身高数据, 即单个数据称为个体 (individual)。然而在实际中, 我们不可能得到 14 亿的全部数据, 也就是 **总体数据通常是无法得知的**。这时, 可以选择抽样 (sampling), 即从总体当中随机抽取出部分个体, 然后得到这部分抽样个体的数据, 一次抽样的结果称为一份样本 (sample)。比如, 从 14 亿的人群中随机抽取出 1 万的个体, 然后去测量这 1 万人的身高数据, 这样就得到了一份包含 1 万个数据的样本, 样本的容量 (sample size), 或者说样本的大小, 是 1 万。注意样本 (sample) 和个体 (individual) 的区别, 样本 (sample) 是一次抽样的结果, 包含多个个体 (individual) 数据, 一份样本中包含的个体数据的数量称为本容量 (sample size)。

随机变量抽样样本的函数称为统计量, 统计量也是随机变量。既然是随机变量, 那统计量也定然会服从某种概率分布, 在统计学中, 把统计量的概率分布统称为抽样分布 (sample distribution)。抽样分布也称统计量分布、随机变量函数分布, 是指样本统计量的分布。注意, 抽样分布并不是某个具体的概率分布, 而是一个统称, 其实就是“抽样样本的函数 (统计量) 的概率分布”的简称, 比较常见的抽样分布是正态分布、学生 t 分布、卡方分布、F 分布等, 一个统计量具体是哪种抽样分布, 这要取决于总体分布 (抽样样本所属的概率分布) 是什么以及统计量的函数是什么。

假设需要调查国人的身高情况, 想要知道全国人民身高的均值和方差。但是显然不可能测量得到全国人民的身高数据, 然后计算得到均值和方差。在统计学上, 通常通过抽样解决这类问题。根据经验, 我们假设全国人民的身高数据服从正态分布, 记为 $X \sim N(\mu, \sigma^2)$, 变量 X 就是总体正态变量, μ 和 σ^2 分别表示总体的期望和方差。然后从总体中, 随机抽取一份包含 1 万个个体的样本, 并且依次测量出这 1 万个个体的身高数据, 记为 $\mathcal{D} = \{X_1, X_2, \dots, X_N\}, N = 100000$, 这就相当于从总体正态分布中取得一个独立同分布的采样。现在我们要利用这个样本估计出总体的期望参数 μ 和方差参数 σ^2 。

显然我们可以应用最大似然估计得到参数的估计值, 这里我们直接使用 节 3.4 的结论, 期望参数 μ 和方差参数 σ^2 的最大似然估计量分别是

$$\begin{aligned}\hat{\mu} &= \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \\ \hat{\sigma}^2 &= \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2\end{aligned}\tag{4.2.1}$$

显然参数的最大似然估计值是一个样本的函数 (统计量), 因此它是一个随机变量。现在我们看下参数估计量 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 的抽样分布分别是什么。

4.2.1 正态分布

我们先看期望参数估计量的抽样分布, 总体正态分布的期望 (均值) 参数的似然估计量 $\hat{\mu}$ 就等于样本的均值统计量。

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (4.2.2)$$

抽样样本集中每一条样本 X_i 都是正态分布随机变量, 根据第一章讲的正态分布的性质: 多个正态随机变量的线性组合结果仍然是一个正态随机变量, 显然估计量 $\hat{\mu}$ 是服从正态分布的。并且根据期望的计算性质, 可知估计量的均值为

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i] = \frac{1}{N} \sum_{i=1}^N \mu = \mu \quad (4.2.3)$$

参数估计量 $\hat{\mu}$ 的期望就等于总体的期望 μ , 这和我们的直观认知是一致的。现在我们看下参数估计量 $\hat{\mu}$ 的方差计算

$$\begin{aligned} V(\hat{\mu}) &= V\left(\frac{1}{N} \sum_{i=1}^N X_i\right) \\ &= \frac{1}{N^2} \sum_{i=1}^N V(X_i) \\ &= \frac{N\sigma^2}{N^2} \\ &= \frac{\sigma^2}{N} \end{aligned} \quad (4.2.4)$$

最终估计量 $\hat{\mu}$ 的抽样分布是均值为 μ , 方差为 σ^2/N 的正态分布, 记作

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right) \quad (4.2.5)$$

也可以记作

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1) \quad (4.2.6)$$

通常会把这个统计量称为 Z 统计量, Z 统计量是服从标准正态分布的, 但是注意, 要想得到这个统计量需要总体的标准差 σ 是已知的才行。

4.2.2 t 分布

上一节我们讲到总体正态分布的期望估计量 (样本均值统计量) 的抽样分布是正态分布 $N(\mu, \frac{\sigma^2}{N})$, 抽样分布的方差是 $\frac{\sigma^2}{N}$ 。其中含有总体的方差, 然而很多时候总体方差 σ^2 是未知的, 此时需要找一个总体方差 σ^2 的替代值。

显然可以使用方差估计值 (样本的方差统计量) 作为总体方差 σ^2 的近似替代, 样本的方差为

$$S_N = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = \hat{\sigma}^2 \quad (4.2.7)$$

根据 [节 2.8.7](#) 讲的 t-分布的定义, 如下统计量 T 是服从自由度为 $N - 1$ 的 t-分布。

$$T = \frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim T(N - 1) \quad (4.2.8)$$

可以对比下公式 [\(4.2.6\)](#) 和公式 [\(4.2.8\)](#) 的区别, 当总体方差未知的时候, 服从标准正态分布的 Z 统计量无法得到, 此时可以使用 T 统计量作为替代。

但是如果样本数量超过 30, 就可以不使用 T 统计量, 而是直接使用 Z 统计量。在 [节 2.8.7](#) 节讲过, 当样本数量超过 30 的时候, t 分布和标准正态分布基本是重合的, 两者没啥区别, 也就是说此时使用方差估计值 (样本方差) 作为 Z 统计量中总体方差的替代也是可以。即当样本数量 $N > 30$ 时, 如下 Z 统计量的抽样分布近似成立。

$$Z = \frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim N(0, 1), \quad N > 30 \quad (4.2.9)$$

4.2.3 卡方分布

现在看下方差估计量 $\hat{\sigma}^2$ 的抽样分布, 总体正态分布 $\mathcal{N}(\mu, \sigma^2)$ 方差参数的无偏计量为

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2 \quad (4.2.10)$$

方差估计量 (样本方差) 的抽样分布是和卡方分布的相关的, 有如下 (渐近) 分布成立。

$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N-1) \quad (4.2.11)$$

证明过程如下:

根据卡方分布的定义: 多个标准正态分布的平方和服从卡方分布, 可知如下变量 Z 是自由度为 N 的卡方随机变量

$$W = \sum_{i=1}^N \left(\frac{X_i - \mu}{\sigma} \right)^2 \quad (4.2.12)$$

分子部分加上同时减去一个 \bar{X} , Z 保持不变。

$$W = \sum_{i=1}^N \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \quad (4.2.13)$$

然后把平方展开

$$W = \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^N \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + 2 \left(\frac{\bar{X} - \mu}{\sigma^2} \right) \sum_{i=1}^N (X_i - \bar{X}) \quad (4.2.14)$$

上述等式右边的最后一项是 0。

$$\sum_{i=1}^N (X_i - \bar{X}) = N\bar{X} - N\bar{X} = 0 \quad (4.2.15)$$

因此 Z 简化成

$$W = \sum_{i=1}^N \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + N \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \quad (4.2.16)$$

然后可以把方差的估计量公式 (4.2.10) 代入到等式中。

$$\begin{aligned} W &= \frac{N-1}{(N-1)\sigma^2} \sum_{i=1}^N (X_i - \bar{X})^2 + \sum_{i=1}^N \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &= \frac{(N-1)\hat{\sigma}^2}{\sigma^2} + \frac{N(\bar{X} - \mu)^2}{\sigma^2} \end{aligned} \quad (4.2.17)$$

移项可得

$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} = W - \underbrace{\frac{N(\bar{X} - \mu)^2}{\sigma^2}}_{\text{标准正态分布的平方}} \quad (4.2.18)$$

又因为有 $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$, 等式右侧的第二项是一个标准正态分布的平方。显然等式右侧变成一个自由度为 $N-1$ 的卡方分布。

$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(N-1) \quad (4.2.19)$$

注意, 如果方差估计量用的似然估计量 (有偏估计)

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \quad \text{有偏估计} \quad (4.2.20)$$

卡方统计量就变成

$$\frac{N\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(N-1) \quad (4.2.21)$$

4.3 极限理论

我们已经理解了总体、样本、统计量、抽样分布的概念, 并且知道正态分布的样本均值统计量的抽样分布是正态分布。抽样样本集 $\mathcal{D} = \{X_1, X_2, \dots, X_N\}$ 的另一个相关因子是样本的数量 N , 正所谓量变引起质变, 本节我们讨论当 N 极大时样本统计量会呈现出什么性质。

设 X_1, X_2, \dots, X_N 为一个独立同分布的随机变量序列, 其公共分布的均值为 μ , 方差为 σ^2 。定义

$$S_N = X_1 + X_2 + \dots + X_N \quad (4.3.1)$$

为这个随机变量序列之和, 本节的极限理论研究 S_N 以及与 S_N 相关的变量在 $N \rightarrow \infty$ 时的极限性质。

由随机变量序列的各项之间的相互独立性可知

$$V(S_N) = V(X_1) + V(X_2) + \dots + V(X_N) = N\sigma^2 \quad (4.3.2)$$

显然当 $N \rightarrow \infty$ 时, S_N 是发散的, 不可能有极限。但是 样本均值统计量

$$M_N = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{S_N}{N} \quad (4.3.3)$$

却不同, 经过简单计算可知

$$\mathbb{E}[M_N] = \mu, \quad V(M_N) = \frac{\sigma^2}{N} \quad (4.3.4)$$

当 $N \rightarrow \infty$ 时, 样本均值统计量 M_N 的方差趋近于 0。方差趋近于 0 意味着 M_N 就与 μ 特别接近。这种现象就是大数定律的内容。按通常的解释, 当样本量 N 很大的时候, 从 X 抽取的样本平均值 M_N 就是变量 X 的平均值 $\mathbb{E}[X]$ 。这里对 X 属于哪种概率分布并没有限制, 非正态分布也符合这个定律。

下面考虑另一个随机变量, 用 S_N 减去 $N\mu$, 可以得到零均值随机变量序列 $S_N - N\mu$, 然后再除以 $\sigma\sqrt{N}$, 就得到随机变量序列

$$Z_N = \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \quad (4.3.5)$$

易证明

$$\mathbb{E}[Z_N] = 0, \quad V(Z_N) = 1 \quad (4.3.6)$$

因为 Z_N 的均值和方差不依赖于样本容量 N , 所以它的分布既不发散, 也不收敛于一点。**中心极限定理**就研究 Z_N 的分布的渐近性质, 并得出结论: 当 N 充分大的时候, Z_N 的分布就接近标准正态分布。

4.3.1 马尔可夫和切比雪夫不等式

我们首先介绍一些重要的不等式, 这些不等式是大数定律和中心极限定理的基础。这些不等式使用随机变量的均值和方差去分析事件的概率, 在随机变量 X 的均值和方差易于计算, 但分布不知道或不易计算时, 这些不等式就非常有用。

首先介绍 **马尔可夫不等式**。粗略的讲, 该不等式是指, 一个 **非负**随机变量如果均值很小, 则该随机变量取大值的概率也非常小。仔细想一想, 这句话其实很好理解。

马尔可夫不等式

设随机变量 X 只取非负值, 则对任意 $a > 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad (4.3.7)$$

下面介绍 **切比雪夫不等式**, 粗略的讲, 切比雪夫不等式是指如果一个随机变量的方差非常小的话, 那么该随机变量取远离均值 μ 的概率也非常小。注意的是: **切比雪夫不等式并不要求所涉及的随机变量非负**。

切比雪夫不等式

设随机变量 X 的均值为 μ , 方差为 σ^2 , 则对任意 $c > 0$,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} \quad (4.3.8)$$

切比雪夫不等式和马尔可夫不等式都是描述的随机变量 X 的某部分概率的上界, 切比雪夫不等式比马尔可夫不等式更准确, 即由切比雪夫不等式提供的概率的上界离概率的真值更近, 这是因为它利用了 X 的方差的信息。当然一个随机变量的均值和方差也仅仅是粗略地描述了随机变量的性质, 所以由切比雪夫不等式提供的上界与精确概率也可能不是非常接近。

4.3.2 弱大数定律

弱大数定律是指独立同分布的随机变量序列的样本均值，在大样本的情况下，以很大的概率与随机变量的均值非常接近。

下面考虑独立同分布的随机变量序列 X_1, X_2, \dots, X_N ，它们的公共分布（总体分布）的均值为 μ ，方差为 σ^2 。定义样本均值

$$M_N = \frac{1}{N} \sum_{i=1}^N X_i \quad (4.3.9)$$

则

$$\mathbb{E}[M_N] = \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \dots + \mathbb{E}[X_N]}{N} = \frac{N\mu}{N} = \mu \quad (4.3.10)$$

再运用独立性可得

$$\begin{aligned} V(M_N) &= \frac{V(X_1 + X_2 + \dots + X_N)}{N^2} \\ &= \frac{V(X_1) + V(X_2) + \dots + V(X_N)}{N^2} \\ &= \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \end{aligned} \quad (4.3.11)$$

利用切比雪夫不等式可得

$$P(|M_N - \mu| \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2} \quad \text{对任意的 } \epsilon > 0 \text{ 成立} \quad (4.3.12)$$

注意，对任意固定的 $\epsilon > 0$ ，上面不等式的右边在 $N \rightarrow \infty$ 时趋近于 0，于是就得到如下的弱大数定律。这里要提到的是：当 X_i 的方差无界时，弱大数定律仍然成立，但是需要更严格而精巧的证明，在此省略。因此，在下面陈述的弱大数定律中，只需要一个假设，即 $\mathbb{E}[X_i]$ 是有限的。

弱大数定律

设 X_1, X_2, \dots, X_N 独立同分布，其公共分布的均值为 μ ，则对任意的 $\epsilon > 0$ ，当 $N \rightarrow \infty$ 时，

$$P(|M_N - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + X_2 + \dots + X_N}{N} - \mu\right| \geq \epsilon\right) \rightarrow 0 \quad (4.3.13)$$

弱大数定律是指对于充分大的 N ， M_N 的分布的大部分都集中在 μ 的附近。设包含 μ 的一个区间为 $[\mu - \epsilon, \mu + \epsilon]$ ，则 M_N 位于该区间的概率非常大。当 $N \rightarrow \infty$ 时，该概率为 1。当然当 ϵ 非常小时，则需要更大的 N ，使得 M_N 以很大的概率落在这个区间。弱大数定律的另一个理解就是在 N 充分大时， M_N 依概率收敛于 μ 。

4.3.3 依概率收敛

弱大数定律可以表述为“ M_N 收敛于 μ ”。但是，既然 M_1, M_2, \dots 是随机变量序列，而不是数列，所以这里“收敛”的含义不同于数列的收敛，应该给予更明确的定义。

依概率收敛

设 Y_1, Y_2, \dots 是随机变量序列 (不必相互独立), a 为一个实数, 如果对任意的 $\epsilon > 0$ 都有

$$\lim_{N \rightarrow \infty} P(|Y_N - a| \geq \epsilon) = 0 \quad (4.3.14)$$

则称 Y_N 依概率收敛于 a 。

根据这个定义, 弱大数定律就是说样本均值统计量依概率收敛于总体分布的真值 μ 。更一般地, 利用切比雪夫不等式可以证明: 如果所有的 Y_N 具有相同的期望, 而方差 $V(Y_N)$ 趋近于 0, 则 Y_N 依概率收敛于 μ 。

如果随机变量序列 Y_1, Y_2, \dots 有概率质量函数或者概率密度函数, 且依概率收敛于 a 。则根据依概率收敛的定义, 对充分大的 N , Y_N 的概率质量或者密度函数的大部分“质量”集中在 a 的 ϵ 邻域 $[a - \epsilon, a + \epsilon]$ 内。所以依概率收敛的定义也可以这样描述: 对任意的 $\epsilon > 0$ 和 $\delta > 0$, 存在 N_0 , 使得对所有的 $N \geq N_0$ 都有

$$P(|Y_N - a| \geq \epsilon) \leq \delta \quad (4.3.15)$$

其中 ϵ 称为 精度, δ 称为 置信水平。依概率收敛的定义有如下形式: 任意给定精度和置信水平, 在 N 充分大时 Y_N 等于 a 。

4.3.4 中心极限定理

根据弱大数定律, 样本均值 $M_N = (x_1 + x_2 + \dots + x_N)/N$ 的分布随着 N 的增大, 越来越集中在真值 μ 的邻域内。特别地, 在我们的论证中, 假定 X_i 的方差为有限的时候, 可以证明 M_N 的方差趋近于 0。另一方面, 前 N 项的和

$$S_N = X_1 + \dots + X_N = NM_N \quad (4.3.16)$$

的方差趋近于 ∞ , 所以 S_N 的分布不可能收敛。换一个角度, 我们考虑 S_N 与其均值 $N\mu$ 的偏差 $S_N - N\mu$, 然后乘以正比于 $1/\sqrt{N}$ 的刻度系数。乘以刻度系数的目的就是使新的随机变量具有固定的方差。中心极限定理指出这个新的随机变量的分布趋于标准正态分布。

具体地说, 设 X_1, X_2, \dots 是独立同分布的随机变量序列, 均值为 μ , 方差为 σ^2 。定义

$$Z_N = \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{X_1 + \dots + X_N - N\mu}{\sigma\sqrt{N}} \quad (4.3.17)$$

经过简单计算可以得到

$$\mathbb{E}[Z_N] = \frac{\mathbb{E}[X_1 + \dots + X_N] - N\mu}{\sigma\sqrt{N}} = 0 \quad (4.3.18)$$

$$V(Z_N) = \frac{V(X_1 + \dots + X_N)}{N\sigma^2} = \frac{V(X_1) + \dots + V(X_N)}{N\sigma^2} = \frac{N\sigma^2}{N\sigma^2} = 1 \quad (4.3.19)$$

中心极限定理

设 X_1, X_2, \dots 是独立同分布的随机变量序列, 序列的每一项的均值为 μ , 方差为 σ^2 。记

$$Z_N = \frac{X_1 + \dots + X_N - N\mu}{\sigma\sqrt{N}} \quad (4.3.20)$$

则 Z_N 的 (累积) 分布函数的极限分布为标准正态 (累积) 分布函数。即

$$\lim_{N \rightarrow \infty} P(Z_N \leq x) = \Phi(x), \quad \text{对任意的 } x \text{ 成立} \quad (4.3.21)$$

可以记作

$$Z_N \sim N(0, 1) \quad (4.3.22)$$

或者

$$\frac{S_N - N\mu}{\sigma\sqrt{N}} \sim N(0, 1) \quad (4.3.23)$$

中心极限定理允许人们可以将 Z_N 的分布看成正态分布, 从而可以计算与 Z_N 相关的随机变量的概率问题, 因为正态分布在线性变换之下仍然是正态分布。如果把 Z_N 的分子分母同时除以 N , 就可以用均值统计量 M_N 表示。

$$Z_N = \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{M_N - \mu}{\sigma/\sqrt{N}} \sim N(0, 1) \quad (4.3.24)$$

再经过一些简单的变换, 可以认为均值统计量的极限分布是均值为 μ 方差为 σ^2/N 的正态分布。

$$M_N \sim N(\mu, \frac{\sigma^2}{N}) \quad (4.3.25)$$

中心极限定理对 X_i 的分布并没有任何要求, 但是 X_i 的分布多少还是有一点不一样的地方。

- 当总体分布 X_i 是正态分布 $N(\mu, \sigma^2)$ 时, 无论样本 N 是多少, 均值统计量 M_N 都服从正态分布 $N(\mu, \sigma^2/N)$ 。
- 当总体分布 X_i 不是正态分布时, 均值统计量 \bar{X} 渐近服从 (极限分布) 正态分布 $N(\mu, \sigma^2/N)$, N 越大越接近正态分布。至于 N 是多少才行, 并没有一个准确的判断方法, 这和 X_i 的分布有关。 X_i 的分布与正态分布相差越大, 需要的 N 就越大; 反之, X_i 的分布与正态分布越相似, 需要的 N 越小。

中心极限定理是一个非常具有一般性的定理。对于定理的条件, 除了序列为独立同分布的序列之外, 还假设各项的均值和方差的有限性。此外, 对 X_i 的分布再也没有其它的要求。 X_i 的分布可以是离散的、连续的或是混合的。

这个定理不仅在理论上非常重要, 而且在实践中也是如此。从理论上看, 该定理表明大样本的独立随机变量序列和大致是正态的。所以当人们遇到的随机变量是由许多影响小但是独立的随机因素的总和的情况, 此时根据中心极限定理就可以判定这个随机量的分布是正态的。例如在许多自然或工程系统中的白噪声就是这种情况。

从应用角度看, 中心极限定理可以不必考虑随机变量具体服从什么概率分布, 避免了概率质量函数和概率密度函数的繁琐计算。而且, 在具体计算的时候, 人们只需均值和方差的信息以及简单查阅标准正态分布表即可。

4.3.5 强大数定理

强大数定律与弱大数定律一样, 都是指样本均值统计量收敛于真值 μ 。但是它们强调的是不同的收敛类别, 下面是强大数定律的一般陈述。

强大数定律

设 X_1, X_2, \dots 是均值为 μ 的独立同分布随机序列, 则样本均值 $M_N = (X_1 + X_2 + \dots + X_N)/N$ 以概率 1 收敛于 μ , 即

$$P \left(\lim_{N \rightarrow \infty} \frac{X_1 + X_2 + \dots + X_N}{N} = \mu \right) = 1 \quad (4.3.26)$$

强大数定律与弱大数定律的区别是细微的, 需要仔细说明。弱大数定律是指 M_N 显著性偏离 μ 的事件的概率 $P(|M_N - \mu|) \geq \epsilon$ 在 $N \rightarrow \infty$ 时趋近于 0。但是对任意有限的 N , 这个概率可以是正的 (大于零)。所以可以想象的是, 在 M_N 这个无穷的序列中, 常常有 M_N 显著偏离 μ 。弱大数定律不能提供到底有多少会显著性偏离 μ , 但是强大数定律却可以。根据强大数定律, M_N 以概率 1 收敛于 μ 。这意味着, 对任意的 $\epsilon > 0$, 偏离 $|M_N - \mu|$ 超过 ϵ 的只能发生有限次。

强大数定律中的收敛与弱大数定律中的收敛是两个不同的概念, 现在给出以概率 1 收敛的定义。

以概率1 收敛

设 Y_1, Y_2, \dots 是某种概率模型下的随机变量序列 (不必独立), c 是某个实数, 如果

$$P\left(\lim_{N \rightarrow \infty} Y_N = c\right) = 1 \quad (4.3.27)$$

则称 Y_N 以概率 1 (或几乎处处) 收敛于 c 。

类似于前面的讨论, 我们应该正确理解以概率 1 这种收敛类型, 这种收敛也是在由无穷数列组成的样本空间中建立的: 若某随机变量序列以概率 1 收敛于常数 c , 则在样本空间中, 全部的概率集中在满足极限等于 c 的无穷数列的子集上。但这并不意味其他的无穷序列是不可能的, 只是他们是非常不可能的, 即他们的概率是 0。

4.4 似然估计量

前几节我们已经把评价一个参数估计量所需的基础知识讨论的差不多了的, 参数估计量一定是一个关于样本的函数, 而样本的函数定义为统计量, 因此参数估计量是统计量。统计量也是一个随机变量, 统计量的分布统称为抽样分布。大数定律给出了均值统计量的极限收敛性质, 中心极限定理进一步强化, 给出了均值统计量的极限分布。概率分布的均值参数的最大似然估计量就等于样本的均值估计量, 因此我们可以运用中心极限定理对均值参数的似然估计量进行分析。

设 X_1, X_2, \dots 是独立同分布的随机变量序列, 亦可以看做是某个总体变量 X 的独立同分布的观测样本。 θ 是变量 X 所属分布的一个参数, 它的最大似然估计量记作 $\hat{\theta}$, 假设参数的真实值是 θ_{true} 。

我们知道估计量 $\hat{\theta}$ 是一个随机量, 它不能精确等于参数真实值 θ_{true} 。但是如果当样本数量 N 足够大时, 估计量 $\hat{\theta}$ 可以依概率收敛于参数的真实值 θ , 那么我们就说这个估计量是一致性估计量。

一致性估计量 (Consistent Estimator)

当样本数量趋近于无穷大时, 估计量 $\hat{\theta}$ 以概率收敛于参数的真实值 θ_{true} ,

$$\lim_{N \rightarrow \infty} P(|\hat{\theta}_N - \theta_{true}| \geq \epsilon) = 0, \quad \text{对任意 } \epsilon > 0 \text{ 成立} \quad (4.4.1)$$

则称这个估计量 $\hat{\theta}$ 就是一致性估计量 (Consistent Estimator)。

在统计学中, 一致估计量 (Consistent Estimator)、渐进一致估计量, 亦称相合估计量、相容估计量。其所表征的一致性或 (相合性) 同渐进正态性是大样本估计中两大最重要的性质。随着样本量无限增加, 估计误差在一定意义上可以任意地小。也即估计量的分布越来越集中在所估计的参数的真实值附近, 使得估计量依概率收敛于参数真值。这里定义的一致性称弱相合性。如果将概率收敛的方式改为以概率 1 收敛就称为强相合性。

为什么是依概率收敛, 而不是确定性收敛? 因为参数估计量本身是一个随机变量, 服从某种概率分布, 只能是以某种概率得到某个确定性的值, 所以这里是依概率收敛到真实值。一致性是对参数估计的基本要求, 一个参数估计要是不满足一致性基本无用。

一致性估计量是依概率收敛到真实值的，并不是一定收敛到真实值，所示我们实际上得到的参数估计量和真实值之间还是会存在一定误差的。我们需要对这个误差进行量化评估，以便能评估一个估计量的好坏。

最直接的误差就是估计量和真实值之间的差值， $d = \hat{\theta} - \theta$ ，但是差值 d 有正有负，不易使用，因此我们采用它的平方，定义参数估计量和参数真实值之间的平方误差 (Squared Error, SE) 为

$$SE = (\hat{\theta} - \theta_{true})^2 \quad (4.4.2)$$

其中 $\hat{\theta}$ 是一个随机量，导致 SE 也是一个随机量，我们用它的期望值作为最终的评价误差，平方误差的期望称之为均方误差 (mean square error, MSE)。

$$\begin{aligned} MSE &= \mathbb{E}[(\hat{\theta} - \theta_{true})^2] \\ &= \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta_{true} + \theta_{true}^2] \\ &= (\mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2) + (\mathbb{E}[\hat{\theta}]^2 - 2\mathbb{E}[\hat{\theta}]\theta_{true} + \theta_{true}^2) \\ &= \underbrace{V(\hat{\theta})}_{\text{估计量的方差部分}} + \underbrace{(\mathbb{E}[\hat{\theta}] - \theta_{true})^2}_{\text{偏差部分}} \end{aligned} \quad (4.4.3)$$

显然一个参数估计量和参数真实值之间的误差由两部分组成：**估计量的方差**和**偏差**，其中方差部分是估计量的方差，不是观测变量 X 的方差。两部分都是非负的，因此一个好的估计量要求两部分都必须小。

4.4.1 估计量的偏差与方差

一个估计量的偏差 (bias) 被定义成估计量的期望和参数真实值之间的差值，

$$b(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta_{true} \quad (4.4.4)$$

当偏差为 0 时，就称这个估计量是**无偏估计量**。

无偏估计量 (unbiased estimator)

当一个估计量满足 $b(\hat{\theta}) = 0$ 时，也就是满足**估计量的期望值等于参数的真实值**，就称这个估计量为**无偏估计**。

- 若 $\mathbb{E}[\hat{\theta}] = \theta_{true}$ 对 θ 所有可能取值都成立，则称 $\hat{\theta}$ 为**无偏估计**。
- 若 $\lim_{N \rightarrow \infty} \mathbb{E}[\hat{\theta}] = \theta_{true}$ 对 θ 所有可能取值都成立，则称 $\hat{\theta}$ 为**渐近无偏估计**。

我们不可能指望作为随机量的估计量正好和未知的参数真值相等，因此估计误差一般非零。另一方面，对于 θ 所有可能的取值，如果平均估计误差是零，则得到一个无偏的估计量。渐进无偏只需要随着观测样本数量 N 的增加，估计量变得无偏即可。

除了偏差，我们还对误差中方差部分的大小感兴趣，现在我们看下估计量的方差部分，估计量的方差也是存在下界的，这可以通过一个定理给出。

Cramer-Rao Lower Bound (CRLB) 定理

Cram'er-Rao Lower Bound (CRLB) 定理描述了一个确定性参数 (deterministic parameter) θ 的估计量的方差的下界

$$V(\hat{\theta}) \geq \frac{\left(\frac{\partial}{\partial \theta} \mathbb{E}[\hat{\theta}]\right)^2}{I(\theta)} \quad (4.4.5)$$

其中分子部分是估计量的期望对参数真实值的一阶导的平方, 如果一个估计量是无偏估计, 那么有 $\mathbb{E}[\hat{\theta}] = \theta_{\text{true}}$, 这时分子就等于 1。

$$(\frac{\partial}{\partial \theta} \mathbb{E}[\hat{\theta}])^2 = (\frac{\partial}{\partial \theta} \theta)^2 = 1 \quad (4.4.6)$$

因此对于无偏估计量, 公式 (4.4.5) 可以简化为:

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)} \quad (4.4.7)$$

$I(\theta)$ 是费歇尔信息 (Fisher-Information) 矩阵。根据 CRLB 定理, 可以看出一个估计量的方差是存在下界的, 并且对于无偏估计量, 估计量的方差的最小值是费歇尔信息的倒数, 显然当一个估计量的方差为下界时, 这个估计量是最稳定的。

通常会用如下方式衡量一个无偏估计量的 **有效性** (efficiency),

$$\mathcal{E}(\hat{\theta}) = \frac{1/I(\theta)}{V(\hat{\theta})} \quad (4.4.8)$$

当 $\mathcal{E}(\hat{\theta}) = 1$ 时, 称此估计量为有效估计 (efficient estimator)。

有效估计量 (efficient estimator)

任意一个估计量, 如果其方差为 CRLB 的下限, 那么这个估计量是有效估计量。

估计量的均方误差由方差和偏差组成, 最好的估计量应该是偏差和方差都尽可能的小, 偏差最小为无偏估计, 所以我们定义出最小方差无偏估计。

最小方差无偏估计

当参数 θ 存在多个无偏估计时, 其中方差最小的估计量就称为最小方差无偏估计 (Minimum Variance Unbiased Estimator, MVUE)。显然, MVUE 是使得 MSE 最小的估计量。然而, 最小方差无偏估计量并不总是存在的, 即使存在, 我们也可能找不到, 没有任何一种方法会始终产生 MVUE。查找 MVUE 的一种有用方法是为参数找到充分统计量。

最后我们总结下,

- 如果估计量依概率收敛于参数真值, 则称这个估计量具有相合性, 或者说一致性。
- 如果估计量的期望等于参数真值, 则这个估计量是无偏估计。
- 对于无偏估计量, 估计量的方差的最小值是费歇尔信息的倒数。

4.4.2 信息量

在参数估计问题中, 我们从目标概率分布的观测样本中获取有关参数的信息。这里有一个很自然的问题是: 数据样本可以提供多少关于未知参数信息? 本节我们介绍这种信息量的度量方法。我们还可以看到, 该信息量度可用于查找估计量方差的界限, 并可用于近似估计从大样本中获得的估计量的抽样分布, 并且如果样本较大, 则进一步用于获得近似置信区间。

假设有一个随机变量 X , 其概率质量 (密度) 函数为 $P(X; \theta)$, θ 是模型未知参数, 并且其值未知。概率质量 (密度) 函数描述了在给定 θ 时, 获取一个 X 的观测值的概率。这里我们先看只有一条观测样本的情况, 稍后再说有多条观测样本的情况。

随机变量 X 单条观测样本对数似然函数为

$$\ell(\theta; X) = \log P(X; \theta) \quad (4.4.9)$$

当利用最大似然估计进行参数估计时, 我们需要求对数似然函数的一阶偏导数

$$\begin{aligned} \ell'(\theta; X) &= \frac{\partial \ell(\theta; X)}{\partial \theta} \\ &= \frac{\partial}{\partial \theta} \log P(X; \theta) \\ &= \frac{P'(X; \theta)}{P(X; \theta)} \end{aligned} \quad (4.4.10)$$

注解: 这里利用了对数函数的求导公式:

$$\nabla \log f(x) = \frac{1}{f(x)} \nabla f(x) \quad (4.4.11)$$

其中 $P'(X; \theta)$ 表示函数 $P(X; \theta)$ 关于 θ 的一阶导数, 同理, 符号 $P''(x; \theta)$ 表示二阶导数。如果参数 θ 是一个标量参数, 关于参数的一阶导数和二阶导数也是一个标量。如果参数 θ 是一个参数向量, 关于参数的一阶偏导数就是一个向量, 二阶偏导数是一个矩阵。

Score function

对数似然函数关于参数的一阶导数称为得分函数 (Score function), 通常用符号 S 表示。

$$S(\theta) = \frac{\partial \ell(\theta; X)}{\partial \theta} \quad (4.4.12)$$

其中 θ 是模型的参数, 当模型存在多个参数时, θ 是参数向量, $S(\theta)$ 也是一个向量。 $S(\theta)$ 是一个关于 **观测样本和参数** 的函数。通常如果似然函数是凹 (concave) 的, 我们可以通过令 $S(\theta) = 0$ 求得参数的最优解。

$S(\theta)$ 是对数似然函数的一阶导数, 一阶导数描述的是函数在这一点的切线的斜率, 导数越大切线斜率越大, 所以 $S(\theta)$ 表示的是对数似然函数在某个 θ 值时模型的敏感度 (sensitive)。

$S(\theta)$ 是关于 X 的一个函数, 所以 $S(\theta)$ 也是一个随机变量, 我们可以研究它的期望与方差。首先来看一下 $S(\theta)$ 的期望, 在开始之前, 先给出有关积分计算的一些技巧。

一个函数的积分和求导是可以互换的, 并且概率质量 (密度) 函数的积分一定是等于 1 的, 所以有如下等式成立。

$$\int f'(x; \theta) dx = \frac{\partial}{\partial \theta} \int f(x; \theta) dx = \frac{\partial}{\partial \theta} 1 = 0 \quad (4.4.13)$$

类似地有:

$$\int f''(x; \theta) dx = \frac{\partial^2}{\partial \theta^2} \int f(x; \theta) dx = \frac{\partial}{\partial \theta} 1 = 0 \quad (4.4.14)$$

$S(\theta)$ 关于样本变量的期望一定是等于 0 的, 结合公式 (4.4.13) 可以推导出 $S(\theta)$ 的期望为:

$$\begin{aligned}
 \mathbb{E}_X[S(\theta)] &= \mathbb{E}_X[\nabla \ell(\theta; X)] \\
 &= \int [\nabla \ell(\theta; X)] P(X; \theta) dx \\
 &= \int [\nabla \log P(X; \theta)] P(X; \theta) dx \\
 &= \int \frac{\nabla P(X; \theta)}{P(X; \theta)} P(X; \theta) dx \\
 &= \int \nabla P(X; \theta) dx \\
 &= \nabla \int P(X; \theta) dx \\
 &= \nabla 1 \\
 &= 0
 \end{aligned} \tag{4.4.15}$$

$S(\theta)$ 的二阶矩 (second moment), 也就是其方差 (Variance), 被称为 Fisher information, 中文常翻译成费歇尔信息, 通常用符号 $I(\theta)$ 表示, $I(\theta)$ 是一个方阵, 通常称为信息矩阵 (information matrix)。

$$\begin{aligned}
 I(\theta) &= V(S(\theta)) \\
 &= \mathbb{E}_X[(S(\theta) - \mathbb{E}_X[S(\theta)])^2] \\
 &= \mathbb{E}_X[S(\theta)^2] \\
 &= \mathbb{E}_X[S(\theta)S(\theta)^T]
 \end{aligned} \tag{4.4.16}$$

实际上, $I(\theta)$ 和对数似然函数的二阶导数的期望值是有关系的, 我们先来看下对数似然函数的二阶导数可以在一阶导数的基础上再次求导得到。

$$\begin{aligned}
 \ell''(\theta; X) &= \frac{\partial}{\partial \theta} \ell'(\theta; X) \\
 &= \frac{\partial}{\partial \theta} \left[\frac{P'(X; \theta)}{P(X; \theta)} \right] \\
 &= \frac{P''(X; \theta)P(X; \theta) - [P'(X; \theta)]^2}{[P(X; \theta)]^2} \\
 &= \frac{P''(X; \theta)P(X; \theta)}{[P(X; \theta)]^2} - \left[\frac{P'(X; \theta)}{P(X; \theta)} \right]^2 \\
 &= \frac{P''(X; \theta)}{P(X; \theta)} - [\ell'(\theta; X)]^2
 \end{aligned} \tag{4.4.17}$$

然后我们看下对数似然函数二阶导数的期望值:

$$\begin{aligned}
 \mathbb{E}_X[\ell''(\theta; X)] &= \int \left[\frac{P''(X; \theta)}{P(X; \theta)} - [\ell'(\theta; X)]^2 \right] P(X; \theta) dx \\
 &= \int P''(X; \theta) dx - \int [\ell'(\theta; X)]^2 P(X; \theta) dx \\
 &= 0 - \int [S(\theta)]^2 P(X; \theta) dx \\
 &= -\mathbb{E}_X[[S(\theta)]^2] \\
 &= -I(\theta)
 \end{aligned} \tag{4.4.18}$$

因此, Fisher information 就等于对数似然函数二阶导数的期望的负数。

$$I(\theta) = -\mathbb{E}_X [\ell''(\theta; X)] \quad (4.4.19)$$

但参数 θ 是一个参数向量时, 对数似然函数的二阶偏导数就是一个矩阵 (方阵), 这个二阶偏导数矩阵称为 **海森矩阵 (Hessian matrix)**, 通常用符号 H 表示, 因此 $I(\theta)$ 经常也被表示成海森矩阵的期望的负数, 当然此时 $I(\theta)$ 也是一个矩阵, 称为 **信息矩阵 (information matrix)**。

$$I(\theta) = -\mathbb{E}_X [H(\theta)] \quad (4.4.20)$$

我们看到, 无论是通过 score function 的方差计算, 还是通过 Hessian 矩阵计算, $I(\theta)$ 都是一个期望值, 所以经常被称为期望化信息 (expected information)。信息量 $I(\theta)$ 是关于随机变量 X 的期望的函数, 已经对 X 求了期望, 所以信息量 $I(\theta)$ 最终的表达式中不再有随机变量 (样本) X , 它仅仅是一个关于参数 θ 的函数。

以上单条观测样本的信息量称为单位费歇尔信息量 (unit Fisher information), 如果有 N 个独立不同分布的 N 条独立观测样本, 它们的信息量就是 N 条单位信息量的求和。如果有 N 条独立同分布的观测样本, 它们的信息量就是 N 倍的单位信息量。因为单位信息量是变量的期望, 与具体的观测样本无关的, 所以当有多条观测样本时, 累加就可以了。

对于独立同分布的观测样本集 $\mathcal{D} = \{X_1, X_2, \dots, X_N\}$, 它的信息量为

$$I_{\mathcal{D}}(\theta) = NI_X(\theta) \quad (4.4.21)$$

$I_{\mathcal{D}}(\theta)$ 是正比于 N 的, 也就是说样本越多, 我们到关于参数的信息量就越大。

Fisher 信息是一种测量可观测随机变量 X 携带其概率所依赖的未知参数 θ 的信息量的方式。Fisher 是 score function (似然函数一阶偏导数) 方差, 也是似然函数二阶偏导数期望的负数, Fisher 信息越大似然函数的曲线越尖锐, 越容易得到参数的最优解。根据 CRLB 定理, 基于独立同分布观测样本 \mathcal{D} 的无偏参数估计量的方差的最小值为

$$V(\hat{\theta}) \geq \frac{1}{I_{\mathcal{D}}(\theta)} = \frac{1}{NI_X(\theta)} \quad (4.4.22)$$

从这个也可以看出, 当 Fisher 信息越大的时候, 参数估计量的方差越小, 方差越小自然就容易得到一个接近参数真值的估计值。同时它是正比于样本数量 N 的, 意味着随着样本的增加, 估计量的方差越来越小。

在 Fisher 信息量的实际应用中, 当需要计算一个独立同分布的观测样本对于参数的信息量 $I_{\mathcal{D}}(\theta)$ 时, 如果按照上面讲的求了观测变量的期望, 那么

$$I_{\mathcal{D}}(\theta) = NI_X(\theta) \quad (4.4.23)$$

此时就称 $I_{\mathcal{D}}(\theta)$ 是期望 (expected) 信息 (矩阵)。也就是不求期望, 直接就按照观测样本值计算, 此时得到的就是观测 (observed) 信息 (矩阵)。使用期望信息矩阵和观测信息矩阵分别计算出的估计量的方差会有些差别, 这在之后的广义线性模型的内容中会用到。

4.4.3 最大似然估计的特性

现在我们来看下最大似然估计量具有哪些特点, 首先回顾一下分布的均值参数和方差参数的最大似然估计量。

已知随机变量 X 的期望参数为 μ , 方差参数为 σ^2 , 两个参数的似然估计量分别记作 $\hat{\mu}_{ML}$ 和 $\hat{\sigma}_{ML}^2$ 。

均值参数的似然估计量

均值参数的最大似然估计量就等于样本的均值统计量,

$$\hat{\mu}_{ML} = \bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} \quad (4.4.24)$$

并且估计量 $\hat{\mu}$ 的期望和方差分别为

$$\begin{aligned}\mathbb{E}[\hat{\mu}_{ML}] &= \mu \\ V(\hat{\mu}_{ML}) &= \frac{\sigma^2}{N}\end{aligned}\tag{4.4.25}$$

显然, 对于均值参数的似然估计量有

- 根据弱大数定律, 它相合估计, 或者说一致性估计。
- 它是无偏估计量, 它的偏差为 0, 因此它的均方误差是 $MSE = \sigma^2/N$ 。
- 它的方差符合 CRLB 的下界, 因此它是最小方差无偏估计, 或者说是有效估计。
- 根据中心极限定理, 它有 **渐近正态性 (asymptotic normality)**, 其渐进服从正态分布 $N(\mu, \frac{\sigma^2}{N})$ 。

方差参数的似然估计量

随机变量 X 的方差参数的似然估计量就是样本的方差, 即

$$\hat{\sigma}_{ML}^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}\tag{4.4.26}$$

现在我们也来看下方差估计量的期望值, 在计算前, 先给出如下几个事实。

$$\mathbb{E}[\bar{X}] = \mu, \quad \mathbb{E}[X_i^2] = \mu^2 + \sigma^2, \quad \mathbb{E}[\bar{X}^2] = \mu^2 + \frac{\sigma^2}{N}\tag{4.4.27}$$

估计量 $\hat{\sigma}_{ML}^2$ 的期望为

$$\begin{aligned}\mathbb{E}[\hat{\sigma}_{ML}^2] &= \mathbb{E}\left[\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^N (X_i - \bar{X})^2\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^N (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right] \\ &= \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^N X_i^2 - 2\bar{X} \sum_{i=1}^N X_i + N\bar{X}^2\right] \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i^2 - \frac{2\bar{X} \sum_{i=1}^N X_i}{N} + \bar{X}^2\right] \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i^2 - 2\bar{X}^2 + \bar{X}^2\right] \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2\right] \\ &= \frac{N(\mu^2 + \sigma^2)}{N} - \left(\mu^2 + \frac{\sigma^2}{N}\right) \\ &= \frac{N-1}{N}\sigma^2\end{aligned}\tag{4.4.28}$$

可以看到 方差似然估计量的期望不等于方差参数真值, 因此它是一个有偏估计量。但是当 $N \rightarrow \infty$ 时, 它们是相等的, 因此 方差似然估计量是渐近无偏的, 同时它也是渐近正态性的。

虽然方差的似然估计量是有偏的, 但是可以做一个简单的变换得到一个无偏的估计量, 显然只需要乘上 $N/(N - 1)$ 即可。

$$\hat{\sigma}_{\text{无偏}}^2 = \frac{N}{N - 1} \hat{\sigma}_{ML}^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1} \quad (4.4.29)$$

当样本数量 N 足够大时 $\hat{\sigma}_{\text{无偏}}^2$ 与 $\hat{\sigma}_{ML}^2$ 其实没有太大的区别。

最大似然估计还有一个特别的性质, 它遵循 **不变原理**: 如果 $\hat{\theta}$ 是 θ 的最大似然估计, 那么对于任意关于 θ 的一映射函数 h , $H = h(\theta)$ 的最大似然估计是 $h(\hat{\theta})$ 。对于独立同分布的观测, 在一些适合的假设条件下, 最大似然估计量是相合的或者说一致的。

另一个有趣的性质是当 θ 是标量参数的时候, 在某些合适的条件下, 最大似然估计量具有 **渐近正态性质**。特别地, 可以看到 $(\hat{\theta} - \theta)/V(\hat{\theta})$ 的分布接近标准正态分布。因此, 如果我们还能够估计出 $V(\hat{\theta})$ 就能进一步得到基于正态近似的误差方差估计。当 θ 是向量参数, 针对每个分量都可以得到类似结论。

渐近正态性 (Asymptotic normality)

我们说一个估计量是渐近正态性的, 如果满足:

$$\sqrt{N}(\hat{\theta} - \theta_{\text{true}}) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I(\theta)}) \quad (4.4.30)$$

或者

$$\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta_{\text{true}}, \frac{1}{NI(\theta)}) \quad (4.4.31)$$

渐近正态性对应着中心极限定理, 最大似然估计是满足渐近正态性的。似然估计量不仅是渐近服从正态分布, 而且是以参数真实值为期望的正态分布, 这表明似然估计量依概率(正态分布)收敛于参数的真实值, 这符合一致性的定义。显然极大似然估计是一致性估计。由于似然估计是一致性估计, 似然估计量是渐近收敛于参数真实值的, 也就是估计量的偏差渐近为 0, 因此可以得出似然估计量是 **渐近无偏估计**。

我们知道估计量的 **MSE** 是由偏差和方差组成的 (公式 (4.4.3)), 无偏性是对估计量的偏差的评价, 而估计量的方差影响着估计值的稳定性, 方差越小估计量就越稳定, 并且最大似然估计量的方差符合 CRLB 的下界, 就等于费歇尔信息的倒数, 如果 θ 是参数向量, 估计量的方差为协方差矩阵, 此时费歇尔信息 $I(\theta)$ 为信息矩阵 (Information matrix)。

$$\text{Cov}(\hat{\theta}_{ML}) = [I(\theta)]^{-1} \quad (4.4.32)$$

这里我们省略证明过程, 有兴趣的读者可以参考其他资料。显然似然估计不仅仅是渐近无偏估计, 而且估计量的方差就等于 CRLB 定理的下界, 因此似然估计量是不仅仅是有效估计量, 而是其最小方差无偏估计 (Minimum Variance Unbiased Estimator, MVUE), **并且我们可以通过 $I(\theta)$ 量化衡量 MLE 估计量的方差**。

当我们用最大似然估计出一个参数的估计值后, 我们期望能量化评估出这个参数估计值的好坏, 大家常用的方法是计算观测值的误差:

$$\text{MSE} = \frac{1}{N} \sum_i^N (y_i - \hat{y}_i)^2 \quad (4.4.33)$$

这种方法衡量的是整个模型的预测效果, 并不能衡量出参数的估计值和参数的最优值之间的误差, 我们已经知道最大似然估计是无偏估计, 那么最终最大似然估计量的误差就可以用如下公式衡量:

$$\text{Standard Errors} = \sqrt{V(\hat{\theta}_{ML})} = \sqrt{\text{diag}([I(\theta)]^{-1})} \quad (4.4.34)$$

$[I(\theta)]^{-1}$ 是协方差矩阵, 其对角线元素是每个参数方差, 开根号后得到每个参数的标准差。

最后我们总结下最大似然估计拥有的特性:

- 最大似然估计量符合大数定律, 它是一致性 (consistency) 估计, 或者数相合估计。
- 由于满足一致性, 渐近收敛于参数真实值, 渐近偏差为 0, 因此它满足渐近无偏性 (unbiased)。
- 根据中心极限定理, 它有 **渐近正态性 (asymptotic normality)**, 其渐进服从正态分布 $N(\mu, \frac{\sigma^2}{N})$ 。
- 最大似然估计量的方差符合 CRLB 定理的下限, 所以是有效估计 (Efficient Estimator), 并且是最小方差无偏估计。

最大似然估计量这些特性都是 **渐近的**, 即当观测样本数量 N 足够大时才能显现出来, 好处是对观测变量所属的分布没有任何要求, 即不管观测变量服从什么概率分布, 最大似然估计都有这些渐近特性。有一个例外是, 如果观测变量的概率分布是正态分布, 则 **不再是渐近的, 而是精确的**。

4.5 置信区间

在统计学中, 由样本数据估计总体分布所含未知参数的真实值, 所得到的值, 称为估计值。最大似然为我们提供了一个利用样本估计总体未知参数的良好方法, 通过上一节的内容, 我们已经知道最大似然估计量拥有非常好的性质, 多数情况下, 能为我们提供一个良好的参数估计值。我们把这种估计结果使用一个点的数值表示“最佳估计值”方法, 称为点估计 (point estimation)。

我们知道估计量是一个随机变量, 通过现有的样本算出一个具体的估计值, 不同的样本算出的估计值也是不同的。最大似然估计量理论上也只是 **依概率收敛于参数真值的**, 因此通常我们使用某个具体的样本算出的估计值和真实值还是不一样的, 估计值和真实值之间到底相差多少, 点估计并没有给出。本节我们讨论统计学中另一种参数估计方法: **区间估计 (interval estimate)**, 也叫 **置信区间 (confidence interval)**, 相比于点估计, 它能给出有关估计值的更多信息。

样本的点估计值并不是完全等于总体参数真实值的, 虽然很接近, 但还是存在一定误差的。在统计学, 我们不能用“可能”、“大概”、“也许吧”这样的字眼去描述这个误差, 而是需要给出去一个量化的描述方法, 就像本章开篇引言说的那样。

统计推断除了结论之外, 还需要说明结论的不确定程度。—《统计学的世界》

我们已经知道, 样本的均值统计量 \bar{X} 可以作为总体均值参数 μ 的点估计量, 而统计量 \bar{X} 是一个随机量, 即不同的样本会得到不同的值, 其渐近服从正态分布。

$$\hat{\mu} = \bar{X} \sim \mathcal{N}(\mu, \sigma^2/N) \quad (4.5.1)$$

其中 μ 和 σ^2 分布是总体的均值参数和方差参数, N 是样本的容量。

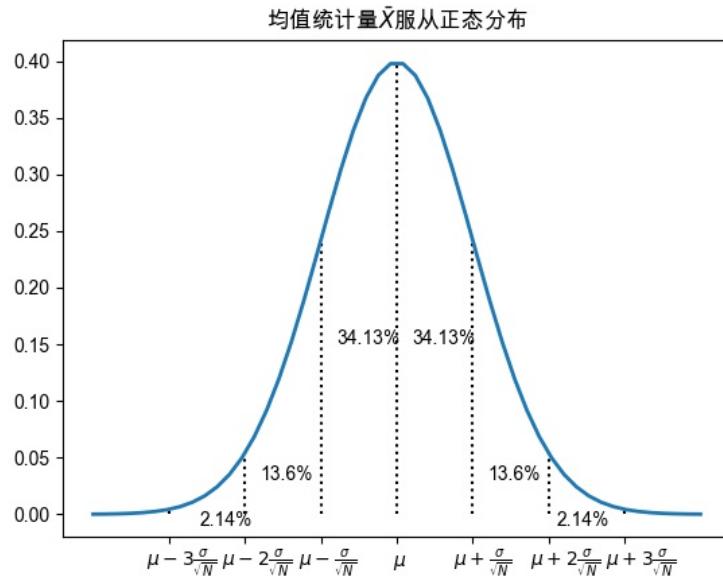
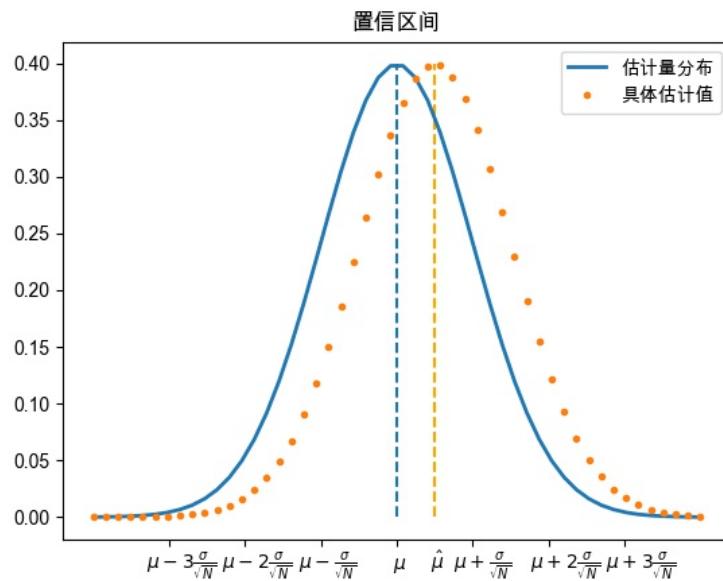
从 [图 4.5.1](#) 可以看出, 均值统计量 \bar{X} 的值有 68.2% 的概率落在区间 $\mu \pm \sigma/\sqrt{N}$ 范围内, 我们知道 μ 是总体的真实均值参数, 也就是说样本估计值 \bar{X} 有 68.2% 的概率和总体真实值 μ 之间的误差在一个标准差的范围 $\pm \sigma/\sqrt{N}$ 内。

但是总体均值参数 μ 是未知的, 无法得知区间 $\mu \pm \sigma/\sqrt{N}$ 的具体范围, 这时可以反转一下。既然 \bar{X} 有 68.2% 的概率落在区间 $\mu \pm \sigma/\sqrt{N}$, 反过来就是, μ 有 68.2% 的概率在区间 $\hat{\mu} \pm \sigma/\sqrt{N}$ 内, 如 [图 4.5.2](#) 所示, $\hat{\mu}$ 是总体均值 μ 的一个具体估计值 (均值统计量 \bar{X} 的一个具体值), $\hat{\mu}$ 落在区间 $[\mu - \sigma/\sqrt{N}, \mu + \sigma/\sqrt{N}]$ 也可以看成是 μ 在区间 $[\hat{\mu} - \sigma/\sqrt{N}, \hat{\mu} + \sigma/\sqrt{N}]$ 。

上面的例子中, 我们给出的置信区间是上下一个标准差的范围, 置信区间的范围可以根据实际情况调整。我们用 δ 表示区间的距离中心点的距离, 则这个区间可以表示成 $[\hat{\mu} - \delta, \hat{\mu} + \delta]$, 这个区间的概率记为 $1 - \alpha$, 则可以记为

$$P(\hat{\mu} - \delta \leq \mu \leq \hat{\mu} + \delta) = 1 - \alpha \quad (4.5.2)$$

区间 $[\text{估计值} \pm \text{误差范围}]$ 称为 **置信区间 (confidence interval)**, $1 - \alpha$ 称为 **置信度 (confidence level)**, 也叫 **置信系数 (confidence coefficient)**。而 $\alpha (0 < \alpha < 1)$ 则被称为 **显著 (性) 水平 (level of significance)**, α 的值通常是事先就确定好的, 显然 α 越大, 置信区间的范围就越小, 一般会选择 0.01, 0.05 这样的值。置信区间的

图 4.5.1: 均值统计量 \bar{X} 服从正态分布, 样本容量 N 越大其标准差越小。图 4.5.2: 估计值 $\hat{\mu}$ 与总体期望 μ 位置是相对的, 黄色曲线是蓝色曲线的一个平移。

端点被称为 置信限 (confidence limits) 或者临界值 (critical values) , $\hat{\mu} - \delta$ 称为置信下限 (lower confidence limit) , 而 $\hat{\mu} + \delta$ 称为置信上限 (upper confidence limit)。

在 α 确定的条件下, 置信区间的范围和 δ 相关, 而 δ 和估计量的标准误差相关, 标准误差越大, 置信区间越宽。换句话说, 估计量的标准误差越大, 对未知参数的真值进行估计的不确定性越大。因此, 估计量的标准误差常被喻为估计量的 精度, 即用估计量去测定真实的总体值有多精确。

相比于原本的似然估计, 置信区间给出了一个参数估计区间, 因此也称作 区间估计 (interval estimate), 顾名思义, 区间估计给出的是一个可能包含参数真值的区间。与之相对的, 点估计给出的是一个具体的估计 (点) 值, 相比单纯的点估计, 区间估计提供了更加丰富的信息。

置信区间利用了参数估计量的抽样分布, 当估计量是无偏估计量时, 估计量的期望就是总体参数的真实值, 注意置信区间是根据参数估计量给出的, 因此这个区间是随机 (量) 的, 而参数的真值是一个固定的数值, 不是随机值。因此置信区间解读成: 随机 (置信) 区间包含参数真值的概率是 $1 - \alpha$ 。不能说成: 参数真值落在这个区间的概率是 $1 - \alpha$ 。

概率分布的常见参数有均值参数 μ 和方差参数 σ , 这两个参数会一直贯穿本书的全部内容, 为了让大家更深刻的理解, 这里我们分别给出两个参数的区间估计的过程。比较特殊的一点是, 均值参数估计量 $\hat{\mu} = \bar{X}$ 的抽样分布有两种情况, 当已知总体方差 σ 或者样本数量足够大时, $\hat{\mu}$ 的抽样分布可以选择标准正态分布 (Z 统计量), 反之需要使用学生 t 分布 (T 统计量), 两种情况我们都简要介绍一下。

4.5.1 均值参数的 Z 区间估计

虽然已知均值参数估计量的抽样分布是 (渐近) 正态分布公式 (4.5.3), 但是在计算技术普及前, 非标准正态分布的概率值不是很方便计算, 所以通常会转成服从标准正态分布的 Z 统计量。

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0, 1) \quad (4.5.3)$$

其中 μ 是总体均值参数的真值, σ 是总体方差参数的真值, N 是观测样本的数量。根据置信区间的公式 (公式 (4.5.2)), 需要找到一个概率为 $1 - \alpha$ 的区间。

$$P(\delta_1 \leq Z \leq \delta_2) = 1 - \alpha \quad (4.5.4)$$

由于 Z 是标准正态分布, 区间 $[\delta_1, \delta_2]$ 是以 0 点为中心左右对称的, 可以记为

$$P(-\delta \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq \delta) = 1 - \alpha \quad (4.5.5)$$

α 的值是事前指定的, 假设为 5% , 则 $1 - \alpha = 95\%$, 根据标准正态分布概率密度的划分情况, 可以近似认为在两个标准差的范围, 而 Z (标准正态分布) 的标准差是 1, 因此有 $\delta = 2$ 。

$$P(-2 \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq 2) = 0.95 \quad (4.5.6)$$

进一步移项可得

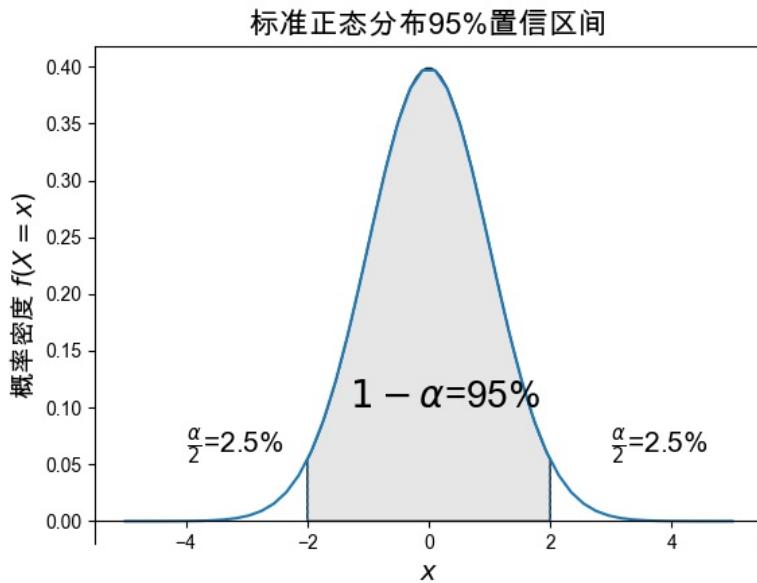
$$P(\hat{\mu} - \frac{2\sigma}{\sqrt{N}} \leq \mu \leq \hat{\mu} + \frac{2\sigma}{\sqrt{N}}) = 0.95 \quad (4.5.7)$$

$\hat{\mu}$ 是先一步利用最大似然估计得到的估计值, 在这里是已知的。如果总体的方差参数 σ 是已知的, 这里就已经结束了, 已经得到了 95% 的置信的区间。然而实际应用中, σ 通常是未知的, 此时如果你的样本数量足够多, 就可以使用 σ 的一个无偏估计值替代。

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}} \quad (4.5.8)$$

最后, 利用 Z 统计量得到的均值参数的 95% 置信区间为

$$\left[\hat{\mu} - \frac{2\hat{\sigma}}{\sqrt{N}}, \hat{\mu} + \frac{2\hat{\sigma}}{\sqrt{N}} \right] \quad (4.5.9)$$

图 4.5.3: 标准正态分布 95% 置信区间 $[-2, 2]$

4.5.2 均值参数的 T 区间估计

在节 4.2.2 讲过, 当总体的方差参数未知或者样本数量小于 30 的时候, 均值统计量的抽样分布可以用学生 t 分布替代。这时在得到对均值参数的置信区间时就要使用学生 t 分布代替标准正态分布, 实现起来比较简单, 是需要把 Z 统计量换成 T 统计量。

$$T = \frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}} \sim T(N - 1) \quad (4.5.10)$$

$$P_T(-\delta \leq \frac{\hat{\mu} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}} \leq \delta) = 95\% \quad (4.5.11)$$

对于 t 分布, 它的概率区间就不是用标准差来分割了, 需要查询 t 分布临界表或者用计算机去计算得到, t 分布的概率是和自由度 $(N - 1)$ 相关的, 假设 $N = 30$, 通过查表可得自由度为 29 的 t 分布 95% 的区间为边界 $\delta = 2.045$, 这比标准正态分布 (Z) 的 2 稍微大了一点。最后, 利用 T 统计量得到的均值参数的 95% 置信区间为

$$\left[\hat{\mu} - \frac{2.045\hat{\sigma}}{\sqrt{N}}, \hat{\mu} + \frac{2.045\hat{\sigma}}{\sqrt{N}} \right] \quad (4.5.12)$$

4.5.3 方差参数的区间估计

我们已经知道方差参数的无偏估计量 $\hat{\sigma}^2$ 是和卡方分布相关的, 如下统计量服从自由度为 $N - 1$ 的卡方分布。

$$\chi^2 = \frac{N\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N - 1) \quad (4.5.13)$$

获取方差估计量置信区间的过程和上面的均值参数的基本是一样的, 唯一注意的地方是, 卡方分布概率密度函数不再是对称的, 上界和下界不再对称。 $1 - \alpha$ 的概率区间, 相当于是在左右两边各扣除 $\alpha/2$ 的概率区间,

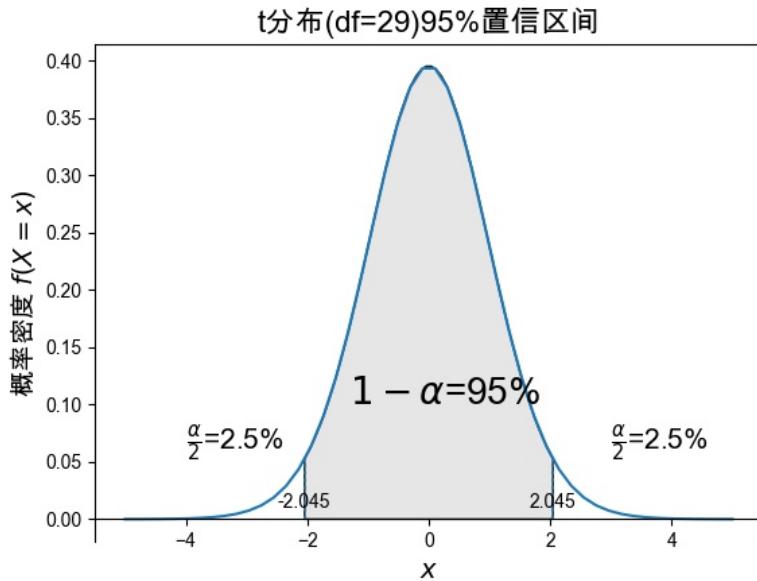


图 4.5.4: 自由度为 29 的 t 分布的 95% 置信区间 $[-2.045, 2.045]$, 相比标准正态分布的 95% 区间 $[-2, 2]$ 稍微大了一些

也就是在分布的左边去掉 $\alpha/2$ 的概率区间, 在分布的右边也去掉 $\alpha/2$ 的概率区间

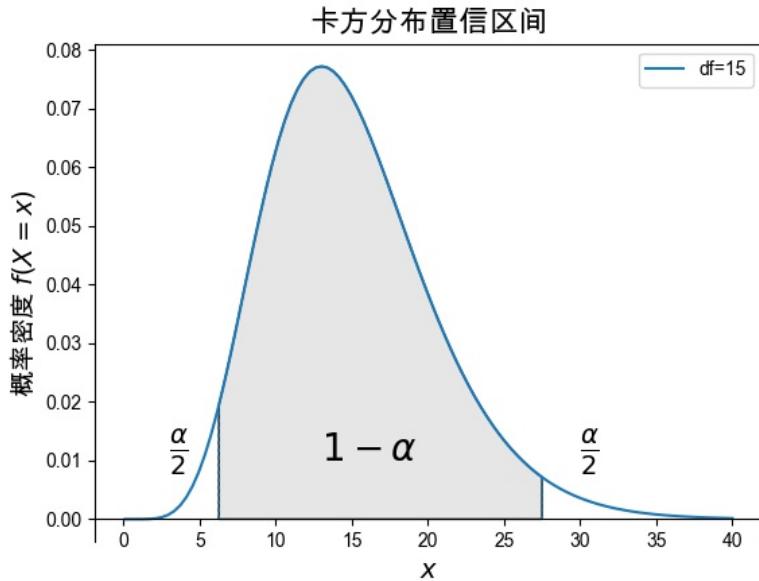
$$P\left(\delta_1 \leq \frac{N\hat{\sigma}^2}{\sigma^2} \leq \delta_2\right) = 1 - \alpha \quad (4.5.14)$$

然后通过查询卡方分布临界表得到分别得到左边界 δ_1 和右边界 δ_2 的值, 也可以利用 python 中 `scipy` 数学工具包计算得到。

```
import scipy.stats as st
# 显著水平为 5%
alpha = 0.05
# 自由度为 15
df = 15
# 左侧边界
delta_1 = st.chi2.ppf(alpha/2, 15)
# 右侧边界
delta_2 = st.chi2.ppf(1-alpha/2, 15)
```

最后调整下公式 (4.5.14) 得到方差参数的置信区间。

$$P\left[(N-1)\frac{\hat{\sigma}^2}{\delta_1} \leq \sigma^2 \leq (N-1)\frac{\hat{\sigma}^2}{\delta_2}\right] = 1 - \alpha \quad (4.5.15)$$

图 4.5.5: 卡方分布置信区间, 两侧各有 $\alpha/2$ 的区域被剔除。

4.6 简单假设检验

统计推断是利用样本数据来对总体得出结论。点估计是使用样本统计量来估计总体参数, 但点估计量并不是准确无误的。置信区间, 又叫区间估计, 给出了点估计量的不确定性程度。本节介绍统计推断中另一种推断的方法, 假设检验 (hypothesis testing)。假设检验 (hypothesis testing), 或者显著性检验 (significance testing) 是用来处理有关总体参数或者总体分布的断言。不同于点估计和区间估计, 假设检验不是用来估计总体参数的, 而是用来判断对于总体 (参数) 的某个假设是否成立。

在日常生活中, 经常会遇到这样一种情况, 我们已经对总体有了猜测或者断言, 需要去验证这个猜测是不是“正确”的, 或者说这个猜测有多大可能性是正确的。然而总体的真实情况我们是无法得知的, 这时就只能通过样本去验证这个猜测, 这就是假设检验做的事情。

举个例子说明下, 假设有一个学者发表了一篇关于国人身高的论文, 论文中声称国人的平均身高为 165cm。你对这个值有些怀疑, 你想验证下这个值是否可信。然而你又不可能统计出全国所有人民的身高去验证专家的结论是否正确。通常的做法是, 自己随机选择一些身高数据作为样本, 然后算出样本的平均值, 假设你算出来是 160cm, 和学者公布的 165cm 有些差异。然而这个差异能说明专家声明的 165cm 是错误的么? 我们知道样本的均值统计量是一个随机量, 不同的采样会得到不同的统计值。那么, 这 5cm 的差异是由于样本的随机性导致的, 还是专家的声明是错误的呢? 这可以通过假设检验给出结论。

假设检验 (hypothesis testing), 又叫显著性检验 (significance testing), 检验的过程一般可以抽象成四个步骤。

步骤 1. 陈述假设

通常我们把对总体的假设称为零假设 (null hypothesis), 通常用符号 H_0 表示, 读作“H 零”。 H_0 是对总体的一个假设或者说断言, 它是一个虚拟的假设。比如在我们的例子中, 零假设就是: 假设专家的声明是正确的, 即国人身高的总体均值是 165cm。这是我们做出的一个虚拟假设, 用符号表示记作:

$$H_0: \mu = 165 \quad (4.6.1)$$

零假设 H_0

在统计学显著性检验中, 被检验的断言叫作”零假设” (null hypothesis), 也叫初始假设。检验主要评估否定零假设的证据有多强。

和零假设相反的结论称为备择假设 (alternative hypothesis), 也可以叫做对立假设, 通常用符号 H_a 表示。如果零假设 H_0 不成立, 就意味着备择假设 H_a 是成立的, 通常二者是对立的。在我们的例子中, 备择假设为:

$$H_a: \mu \neq 165 \quad (4.6.2)$$

假设检验的过程, 就是先假设 H_0 是正确的, 然后在这个前提下寻找否定 H_0 的证据, 如果找到”证据”, 并且这个证据足够强烈, 就拒绝 (reject) H_0 , 接受 H_a ; 如果没有足够的”证据”, 就接受 (accept) H_0 。

经过前面的熏陶, 我们已经了解到, 在统计学中没有什么是绝对的, 一切都是通过概率来描述。这意味用来否定 H_0 的”证据”也不是绝对的, 亦然是”概率”的, 所以用的是接受 (accept)、拒绝 (reject) 这样的词, 而不是其他准确判定的词。因为我们找到的证据并不是百分百的证明 H_0 是错误的, 只能是从概率上认为 H_0 成立的可能性”比较小”, 所及拒绝了 H_0 选择了 H_a , 假设检验只是一种从概率上选择最有可能的结果, 而不是像数学上的证明一样给出绝对的对错。

步骤 2. 设定决策标准

假设检验是要找到否定 H_0 的”证据”, 通常这种”证据”就是在 H_0 成立的条件下发生了一件”不可能”发生的事件, 所谓的”不可能事件”, 就是一件概率很小的事件。那么这个”不可能”的程度是多少, ”概率很小”又有多少? 这就需要给出一个标准。这个标准称为显著水平 (level of significance)。

显著水平

显著水平 (level of significance, or significance level), 是判断小概率事件有多小的标准, 显著水平值越小, 意味这个事件发生概率越小, 越极端。通常用符号 α 表示。

显著水平 α 通常会设置成 5%、2%、1% 等值, 其含义是只要一个事件发生的概率小于等于 α 就认为这是一件极端的小概率事件。在假设检验中, 如果在 H_0 成立的条件下, 发生了一件概率小于等于 α 的事件, 认为 H_0 很可能是错误的, 此时会拒绝 H_0 。

步骤 3. 计算检验统计量

有了检验标准后, 就需要计算出一个值和这个标准比较, 计算这个值的统计量就称为检验统计量, 检验统计量有很多种, 一般会根据实际的问题场景选择合适的检验统计量, 然后计算出这个检验统计量的值以及理论上得到这个值的可能性 (概率), 这个概率值称为 P 值 (P-value), 最后把这个 P 值和检验标准值 α 进行比较, 并根据比较结果给出结论。

样本均值统计量 \bar{X} 的抽样分布是正态分布 $\mathcal{N}(\mu, \sigma^2/N)$, 通过样本统计量的抽样分布就能计算出样本统计值的发生概率。在我们的例子中, 在 H_0 成立的前提下, 身高总体分布的均值 (期望) 就是 $\mu = 165$, 总体的方差未知, 暂时用符号 σ^2 表示, 则抽样样本的均值统计量 \bar{X} 的抽样分布是

$$\bar{X} \sim \mathcal{N}(\mu = 165, \sigma^2/N) \quad (4.6.3)$$

以上抽样分布的方差 $Var(\bar{X}) = \sigma^2/N$ 是未知的, 其中 σ^2 是总体方差参数, N 是抽样样本容量, 通常样本容量 N 是已知的, 假设抽样样本容量是 100。这时还需要得到总体方差参数 σ^2 才可以, 根据点估计的知识, 可以用样本方差近似估计总体方差

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (\bar{x} - x_i)^2}{N - 1} \quad (4.6.4)$$

这里我们假设算出来的总体方差估计值是 $\hat{\sigma}^2 = 36.0$, 则样本均值统计量的抽样分布的方差为 $\sigma^2/N = 36.0/100 = 0.36$, 在 H_0 成立的条件下, 样本均值统计量 \bar{X} 的抽样分布就为

$$\bar{X} \sim \mathcal{N}(165, 0.36) \quad (4.6.5)$$

理论上样本统计量 \bar{X} 的期望是 165, 方差是 0.36。然后我们发现, 从抽样样本计算得到样本均值为 $\bar{X} = 160$ 。理论上样本结果越接近 165, 专家 (H_0) 是正确的可能性就越大; 样本均值结果偏离 165 越远, 专家 (H_0) 是错误的可能性就越大。

那么样本统计值偏离期望值多远才叫小概率事件呢? 总要有个判断标准。这个标准就是我们在上个步骤中制定的显著水平 α 。图 4.6.1 是均值统计量 \bar{X} 的抽样分布 (正态分布) 的概率分布曲线。我们把曲线下方的面积分成两个区域, 紧邻期望值两侧的中间区域称为 **置信区间**, 其面积是 $1 - \alpha$, $1 - \alpha$ 是这个区域的面积, 也是 \bar{X} 落在这个区间的概率值, 称为 **置信水平**。在假设检验中这个区域也叫作 **接受域**, 表示我们接受零假设 H_0 的区域, 样本统计值落在接受域的概率是 $1 - \alpha$ 。接受域两侧的阴影区域称为 **拒绝域**, 表示拒绝零假设 H_0 的区域, 其面积总和是 α , 样本统计值落在这个区间的概率是 α 。显著水平 α 就是对这个“极端小概率”事件的一个标准, 如果统计量 \bar{X} 的值落在这个区间, 我们就认为发生了小概率事件, 此时选择拒绝零假设 H_0 。

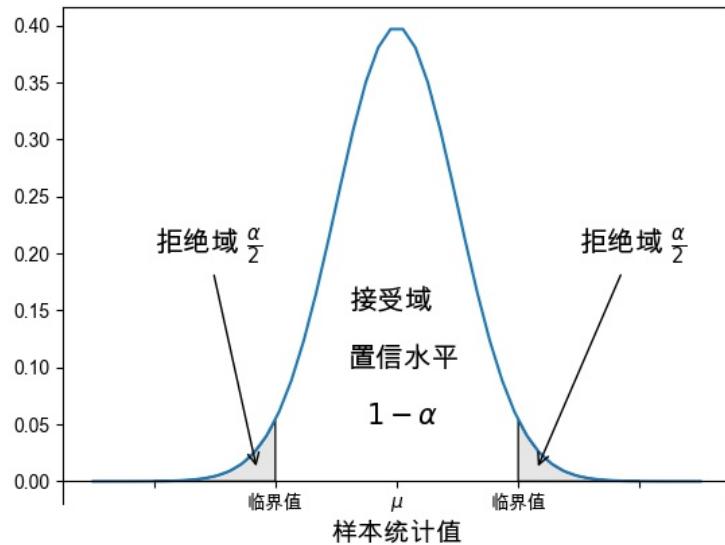


图 4.6.1: 标准正态分布的区域划分。阴影部分是拒绝域, 左右两部分的概率和为 α 。中间区域是接受域, 它的概率是 $1 - \alpha$ 。

下一步就是要算样本统计值 160 落在了抽样分布 $\mathcal{N}(165, 0.36)$ 的哪个区域, 是落在了接受域还是拒绝域, 落在不同的区域会导致我们对 H_0 做出不一样的选择。然而在计算机普及之前, 要计算出 160 在正态分布 $\mathcal{N}(165, 0.36)$ 哪个区域不是一件简单的事情, 因此通常并不直接使用均值统计量进行检验 (验证) 而是使用 Z 统计量, Z 统计量就是均值统计量转化成标准正态分布。

Z 统计量

如下统计量称为 Z 统计量, Z 统计量的抽样分布是 **标准正态分布**。

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0, 1) \quad (4.6.6)$$

有关 Z 统计量的推导我们在 [节 4.2.1](#) 和 [节 4.3.4](#) 都有讲到过, 可以回顾一下相关内容。Z 统计量其实就是把服从非标准正态分布的样本均值统计量转换为一个服从标准正态的分布的统计量, 标准正态的分布方便进行检验计算, 可以通过查表的方式得出 P 值。Z 统计量也可以用来衡量样本均值结果值距离期望值有多少个标准差的距离, 有时也叫作标准分。

现在我们把样本均值转成成 Z 的值, 通过计算可得 $Z = \frac{163 - 165}{\sqrt{0.36}} = -2/0.6 = -3.34$, 意味着我们的检验统

计量的样本值偏离其抽样分布理论期望值 3.34 个标准差 (Z 的期望值为 0, 标准差为 1) 远, 负号代表是负偏离, 小于期望值。如果是正数, 就是大于期望值, 是正偏离。

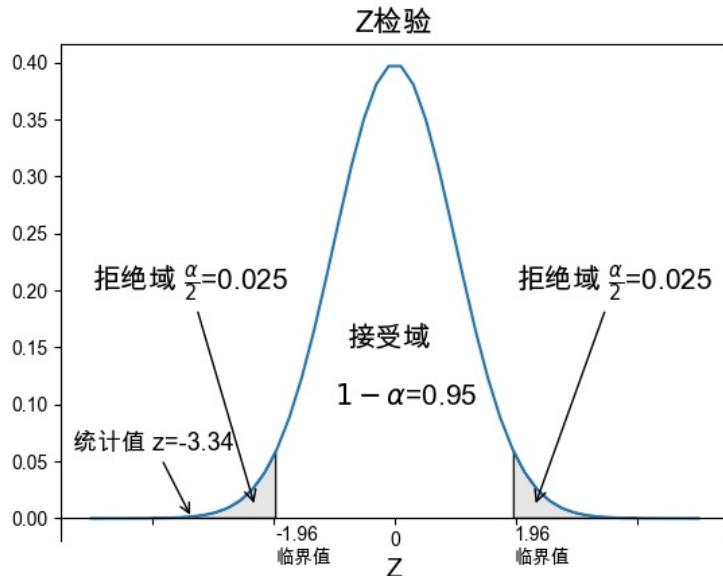


图 4.6.2: 当 $\alpha = 0.05$ 时, 左右边界分别是 -1.96 和 1.96 , 计算出的 $z = -3.34$ 正好落在了左侧拒绝域内。

步骤 4. 做出决策

样本均值 \bar{X} 结果值 163 偏离理论期望值 165 的原因可能有两种, 第一个可能的原因是, 正常的随机结果, 因为统计量 \bar{X} 本就是一个随机量, 不同样本会得到不同值, 出现不一致是正常的随机现象。第二个可能原因就是, H_0 是错误的, H_1 才是对的, 总体期望不是 165, 也会导致样本结果偏离理论期望值。那么如何判断是哪个原因导致现在这个结果呢? 很遗憾, 并没有准确的判断方法。我们只能根据概率“接受”其中的一个, 这也是假设检验的本质。

理论上, 样本结果值偏离理论期望值越远, 第二个原因的可能性越大。换句话说, 检验统计量 Z 的值越大, H_0 错误的可能性越大。我们知道统计量 \bar{X} 是服从正态分布的, 在一个正态分布中, 越偏离中心位置的值概率越小, 得到一个远离中心的值是一个概率很小很极端的事件。因此如果我们通过样本计算得到值在正态分布上是一个很小概率的事件, 就意味发生了一件很极端(概率很小)的事件, 而我们认为通常不会这么“巧合”。如果在 H_0 是正确的前提下, 发生了一件极端的事件, 我们更倾向于认为 H_0 是错误的。本例中计算的到 $Z = -3.34$, 那么要得到这样一个样本结果值 $|Z| \geq 3.34$ 的概率是多少呢?

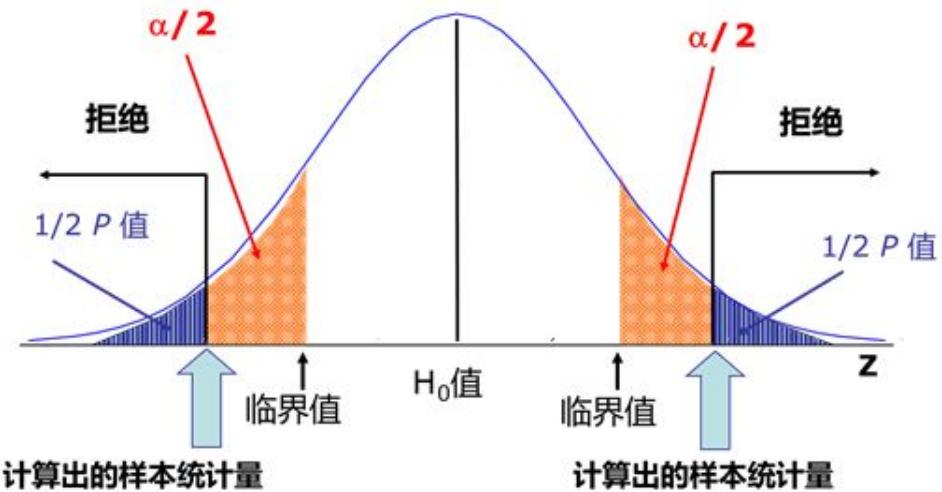
在正态分布中, 采样值落到区间 $[\mu, \mu \pm \sigma]$ 的概率大约是 68.27%, 落到 2 个标准差区间 $[\mu, \mu \pm 2\sigma]$ 的概率大约是 95.46%, 落到 3 个标准差区间 $[\mu, \mu \pm 3\sigma]$ 的概率大约是 99.73%, 参考 图 2.8.6。

我们的例子中计算得到 $Z = -3.34$, 通过查正态分布表可以得到 $P(|Z| \geq 3.34) = 0.08\%$, 其含义是, 正态分布得到一个偏离期望值 至少 3.34 的标准差距离的值的概率是 0.08%。这个概率值在假设检验中称作 P 值 (P-value)。

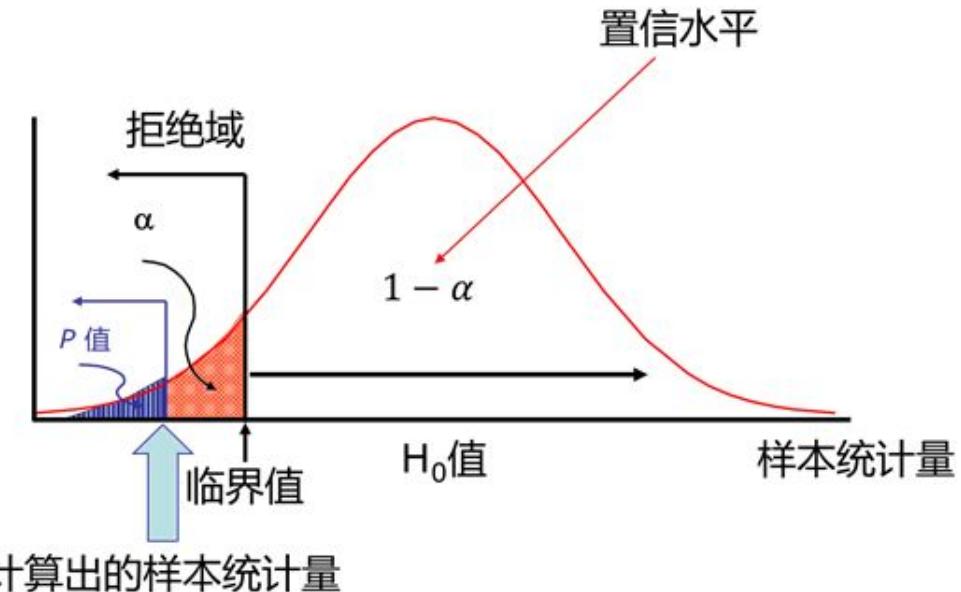
P 值

统计检验的 P 值 (P-value) 是在 H_0 为真的假设下, 所得到的样本统计值结果会像实际观察结果那么极端或者更极端的概率。P 值越小, 说明越极端, 否定 H_0 的证据就越强。注意, P 值算的不是一个点的概率 $P(|Z| = z)$, 而是这个点以及比这个点更极端的概率 $P(|Z| \geq z)$ 。

注意我们的备择假设是 $H_1: \mu \neq 165$, 不等于意味着大于或小于, 也就是本例中 H_1 是包含负偏离和正偏离

图 4.6.3: 双侧检验 P 值和 α 的关系

两个的, 需要计算检验统计量 Z 落在分布两侧的概率之和, 这种正负偏离一起算的检验称为双边检验。如果把备择假设 H_1 改成 $\mu < 165$, 就变成了单边检验, 在单边检验中拒绝域只有一侧, 此时只能计算 $Z \leq -3.4$ 的概率。

图 4.6.4: 单侧检验 P 值和 α 的关系

我们已经计算了 P 值 0.08%, 那这个 P 值是大还是小, 算不算极端事件, 需要有一个判断标准。这个标准就是步骤 2 中设置的显著性水平 α , 如果 $P \leq \alpha$, 则认为发生了极端事件, 此时我们拒绝零假设 H_0 , 接受备择假设 H_1 ; 如果 $P > \alpha$, 则认为没有发生极端事件, 样本统计值的偏离是正常的随机误差造成的, 此时接受零假设 H_0 。假设本例中, 我们设置显著水平 $\alpha = 1.0\%$, 显然 P 值 0.08 小于显著水平 0.1, 因此我们拒绝零假设 H_0 , 我们有理由认为专家是在胡扯。

决策错误

重要: 假设检验对总体断言的决策并不是百分百正确的, 对于零假设的接受或拒绝的决策是基于概率的, 所以是有可能做出错误的决策的, 显著水平 α 就是做出错误决策的概率的上限。

回顾整个检验过程, 从始至终我们都是不知道总体的真实情况的, 仅仅根据一份样本统计值做出的决策, 而决策的判定又是基于概率的, 因此假设检验给出的结论也有错误的可能。零假设的真实情况和检验结论之间存在四种可能结果。

表 4.6.1: 假设检验的四种决策结果

	接受零假设	拒绝零假设
零假设为真	正确 $1 - \alpha$	Type I 错误 α
零假设为假	Type II 错误 β	正确 $1 - \beta$

表 4.6.1 是用表格的形式给出 4 种情况, 其中两种结果是正确的, 另两种结果是错误的。

- 零假设为真, 并且决策结果是接受, 此时决策结果是正确的, 这个结果的概率是 $1 - \alpha$ 。
- 零假设为真, 然而决策结果是拒绝, 此时决策结果是错误的, 这个结果的概率是 α 。
- 零假设为假, 然而决策结果是接受, 此时决策结果是错误的, 这个结果的概率是 β , 此时称为 Type I 错误。
- 零假设为假, 并且决策结果是拒绝, 此时决策结果是正确的, 这个结果的概率是 $1 - \beta$, 此时称为 Type II 错误。

Type II 错误

如果检验的决策是接受零假设, 那么这个结果有可能是正确的也可能是错误的。如果零假设实际上是错误的, 那么我们就做了一个错误的决策, 此时称为 Type II 错误, 又叫 β 错误, β 表示做出错误决策的概率, 当然这个 β 的值我们是无法得知的。

假设检验的零假设通常是对事物或者总体已有的一个认知或者结论, 我们通过假设检验去论证这个认知是否正确, 如果假设检验的决策是 β 错误, 相当于我们的检验过程其实没有贡献什么, 并没有判断出来这个零假设是错误的。更可悲的是, 我们自己并不知道发生了 β 错误。

Type I 错误

同样的, 如果检验的决策是拒绝零假设, 也有可能是错误的决策。零假设是真实的, 但决策结果是拒绝零假设, 我们把这类型的错误称为 Type I 错误。幸运的是, 我们能掌控犯 Type I 错误的概率上限, Type I 错误发生的概率上限就是显著水平 α 。我们通过比较 P 值和 α 做出拒绝零假设决策, 因此 α 就是代表着我们做出拒绝零假设决策的概率, 也就是犯 Type I 错误的 **概率上限**, 注意 α 不是 Type I 错误的概率, 而是其理论上限, 可以通过减小显著水平 α 的值, 来降低 Type I 错误的概率。但是, 并不能一味的降低 α 的值, 随着 α 的降低, 我们拒绝零假设的条件就更加严苛, 减少了拒绝零假设的可能性, 因此也就减少了检验出错误零假设的能力 (power)。

假设检验的关键思想在于一个检验统计量 (test statistic) 及其在虚拟假设下的抽样分布, 根据观测数据算出检验统计量值决定是否接受 H_0 。其过程概括起来就是

- 对总体某个参数的值做出一个虚拟的假设, 称为零假设, 记作 H_0 。与 H_0 不同的结果是对立假设, 记作 H_a 。
- 选择一个和这个参数相关的检验统计量, 并根据样本和虚拟假设的参数值计算出这个检验统计量的值, 然后算出检验统计量值对应的 P 值。所谓 P 值就是, 在检验统计量的抽样分布下, 得到检验统计量值及其更极端值的概率。
- 根据 P 值和显著水平 α 做出接受还是拒绝零假设的决策。

习惯上会根据检验统计量 (抽样分布) 对检验过程进行命名, 比如利用 Z 统计量进行假设检验就称为 Z 统计量进行假设检验就称为 T 检验, 利用 χ^2 统计量进行检验就称为 χ^2 检验, 下面我们分别对这些检验进行简单的介绍。

4.6.1 Z 检验

和均值参数相关的统计量有两个 Z 统计量和 T 统计量, 当总体方差已知或者抽样样本足够多时, 使用 Z 统计量即可, 当总体方差未知并且抽样样本比较少时, 建议使用 T 统计量。本节我们先介绍用 Z 统计量对均值参数进行检验, 下一节讨论如何用 T 统计量对均值参数进行检验。

假设我们要对某个总体的分布的均值参数 μ 进行检验, 总体的方差参数 σ^2 认为是已知的。我们对均值参数 μ 做出一个虚拟的假设, 假设它的真实值为 μ^* , 零假设就是

$$H_0 : \mu = \mu^* \quad (4.6.7)$$

与零假设结果相反的对立假设为

$$H_a : \mu \neq \mu^* \quad (4.6.8)$$

然后我们得到一个容量为 N 的抽样样本 (观测样本), 利用这个样本可以得到 μ 的一个估计值, 记作 $\hat{\mu}$, 估计量 $\hat{\mu}$ 的标准误差为 σ/\sqrt{N} , 然后就可以计算出 Z 统计量的一个值, 记作 z 。

$$z = \frac{\hat{\mu} - \mu^*}{\frac{\sigma}{\sqrt{N}}} \quad (4.6.9)$$

最后看 z 落在了哪个区域, 如果落在接受域, 则接受零假设, 即认为零假设是正确的。反之, 如果落在拒绝域, 就拒绝零假设。判断 z 落在哪个区域有两种方法。

第一种方法, 在给定显著水平 α 的值后, 计算出临界值 δ_1, δ_2 , 这和上一节置信区间的方法是一样的, 可以算出接受域 (置信) 区间 $[\delta_1, \delta_2]$, 然后判断 z 值是否在区间 $[\delta_1, \delta_2]$ 内即可得出结论。从这里可以看出, 建设检验和置信区间本质上 (区间估计) 是一样的。

第二种方法, 先计算出 P 值, 即 $P(|Z| > z)$, 如果 $P \leq \alpha$ 则说明 z 值落在了拒绝域, 反之, 如果 $P > \alpha$ 则说明 z 值落在了接受域。

$$\begin{aligned} P\text{值} &= P(Z \geq z) + P(Z \leq -z) \\ &= 2\Phi(-z) \end{aligned} \quad (4.6.10)$$

公式中 Φ 是标准正态的分布的累积分布函数, 由于标注正态分布是对称的, 因此有 $P(Z \geq z) = P(Z \leq -z)$

4.6.2 T 检验

当总体方差参数 σ 未知并且抽样样本数量比较少时, 就用 T 检验替代 Z 。 T 检验和 Z 检验的过程是完全一样的, 甚至二者的统计量值计算公式都是一样的。

$$t = \frac{\hat{\mu} - \mu^*}{\frac{\hat{\sigma}}{\sqrt{N}}} \quad (4.6.11)$$

不一样的地方仅在于计算 P 值的时候, 要使用学生 t 分布的累积分布函数。我们用符号 T_d 表示自由度为 n 的学生 t 分布的累积分布函数, 则 P 值的计算方法为

$$\begin{aligned} P\text{值} &= P(T \geq t) + P(T \leq -t) \\ &= 2T_d(-t) \end{aligned} \quad (4.6.12)$$

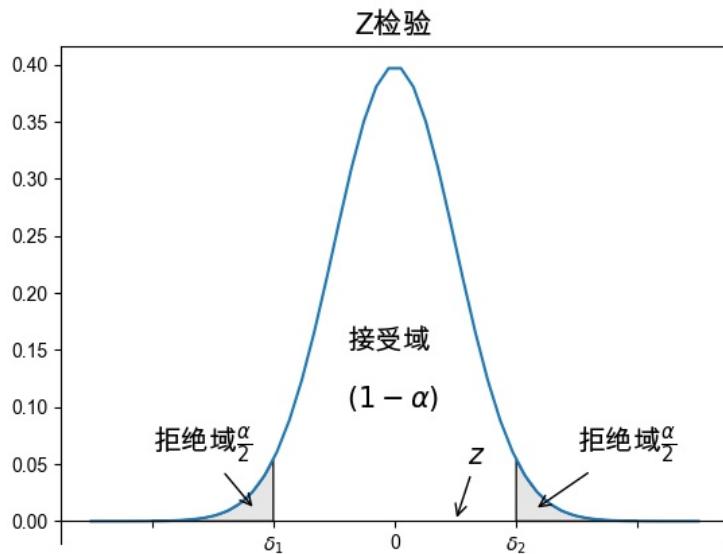


图 4.6.5: 双侧 Z 检验

4.6.3 卡方检验

卡方检验常用于对方差参数进行检验, 零假设是对方差参数的一个虚拟假设, 假设方差参数的值为 σ^* , 然后通过卡方检验决定是否接受这个假设。

设方差参数的零假设和对立假设分别为

$$\begin{aligned} H_0 : \sigma^2 &= \sigma^* \\ H_a : \sigma^2 &\neq \sigma^* \end{aligned} \quad (4.6.13)$$

然后利用容量为 N 的样本得到方差参数的一个无偏估计值 $\hat{\sigma}^2$, 有了 $\hat{\sigma}^2$ 和 σ^* 后, 可以计算出卡方统计量的值。

$$x = \frac{N\hat{\sigma}^2}{\sigma^*} \quad (4.6.14)$$

这里要注意 χ^2 分布不再是对称结构, 如果是双边检验, 无法直接计算出 P 值, 此时可以先算出接受 (置信) 域区间 $[\delta_1, \delta_2]$, 如 图 4.6.6 所示, 然后根据 χ^2 值是否落在这个区间做出决策。

事实上, 由于卡方分布是左偏的, 整个图形期望值距离左侧很近, 而右侧是一条长尾, 所以多数情况下, 卡方检验使用的是单 (右) 侧检验, 如 图 4.6.7 所示。此时可以计算出 P 值, 比较 P 值和显著水平 α 大小做出决策。

$$P\text{值} = P(\chi^2 \geq x) \quad (4.6.15)$$

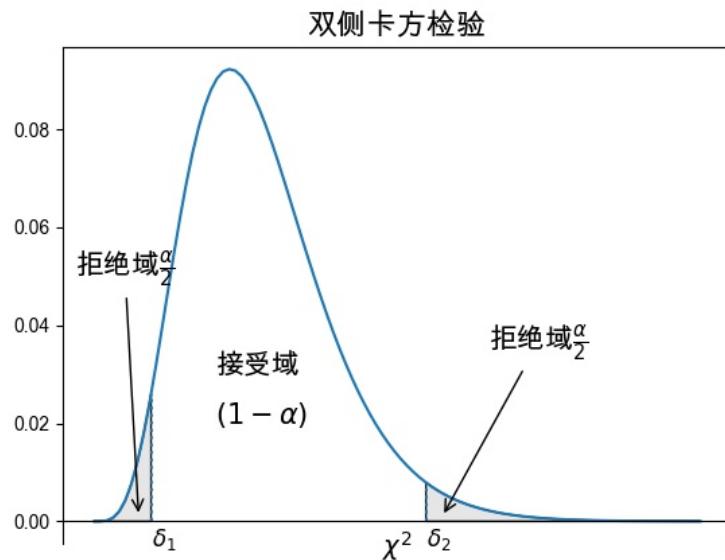


图 4.6.6: 双侧 χ^2 检验

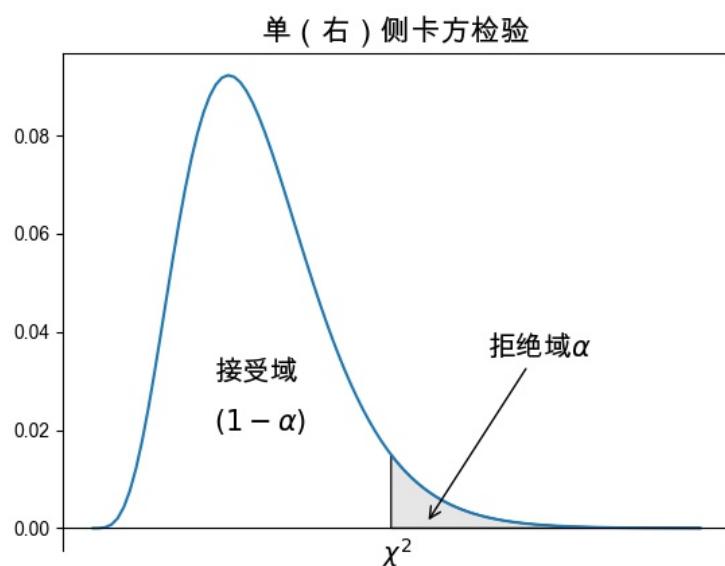


图 4.6.7: 单侧 χ^2 检验

指数族

本章我们介绍一类特殊概率分布，叫做指数族分布 (exponential family)。指数族分布并不是一个具体的概率分布，而是指一类分布，这类分布具有某些共同的特性，所以它们形成了一个概率分布族 (family)。很多常见的概率分布都属于指数族，比如高斯分布、二项分布、多项式分布、泊松分布、gamma 分布、beta 分布等等。

5.1 指数族的定义

顾名思义，指数族的含义就是这类概率分布的概率密度 (质量) 函数是一个指数函数。一个概率分布的概率密度 (质量) 函数如果具有如下的形式，这个概率分布就属于指数族分布。

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\{\phi(\theta)^T T(x)\} \quad (5.1.1)$$

其中 x 表示随机变量， θ 是分布的未知参数。 $T(x)$, $h(x)$, $\phi(\theta)$ 都是已知的函数，通常 $h(x)$ 被称为基础度量值 (base measure)， $T(x)$ 是充分统计量 (sufficient statistic)， $\phi(\theta)$ 是关于参数 θ 的一个函数。 $\phi(\theta)$ 和 $T(x)$ 可以是向量也可以是标量。如果两个都是标量 (scalar-value)， $\phi(\theta)^T T(x)$ 就是两者的乘积；如果是向量 (vector-value)， $\phi(\theta)^T T(x)$ 就是两者的内积。不管两者是标量还是向量， $\phi(\theta)^T T(x)$ 的结果都是一个实数数值。函数 $Z(\theta)$ 是这个分布的配分函数 (partition function)， $Z(\theta)$ 就是对分子的积分。配分函数通常出现在概率密度 (质量) 函数中，是为了使得这个函数的输出值符合概率约束，即使得函数的输出值在 $[0, 1]$ 范围内。

$$Z(\theta) = \ln \int h(x) \exp\{\phi(\theta)^T T(x)\} dx \quad (5.1.2)$$

通常公式 (5.1.1) 有多种变式，比如，我们令 $g(\theta) = \frac{1}{Z(\theta)}$ ，这样可以使得式子变得更加整洁。

$$p(x|\theta) = h(x)g(\theta) \exp\{\phi(\theta)^T T(x)\} \quad (5.1.3)$$

有时还会把 $Z(\theta)$ 移到指数的内部，其中 $A(\theta) = \ln Z(\theta)$ ，通常被称为对数配分函数 (log-partition function)。

$$p(x|\theta) = h(x) \exp\{\phi(\theta)^T T(x) - A(\theta)\} \quad (5.1.4)$$

也有一些资料会把 $h(x)$ 也移到指数内部，其中 $S(x) = \ln h(x)$ 。

$$p(x|\theta) = \exp\{\phi(\theta)^T T(x) + S(x) - A(\theta)\} \quad (5.1.5)$$

这些不同的表示都是同一个公式的变型而已, 它们是等价的。

规范形式 (The Canonical Form)

式中 $\phi(\theta)$ 是关于参数 θ 的一个函数, 在指数族中, 函数 $\phi(\cdot)$ 是一个双射函数 (也称一一对应或一一映射), 一一映射意味着函数 $\phi(\cdot)$ 是单调并且可导的, 其一定存在反函数。我们定义 $\eta = \phi(\theta)$, 因为 η 和 θ 的值是一一映射的, 所以二者是可以相互转化的。

$$\begin{aligned}\eta &= \phi(\theta) \\ \theta &= \phi^{-1}(\eta)\end{aligned}\tag{5.1.6}$$

鉴于此, 通常可以用 η 代替 θ 表示概率分布函数的未知参数, 得到一个更简单的形式。

$$p(x|\eta) = \exp\{\eta^T T(x) + S(x) - A(\eta)\}\tag{5.1.7}$$

通常把参数 η 称为自然参数 (natural parameter) 或者规范参数 (canonical parameter)。原来的形式中 $A(\theta)$ 是一个关于 θ 的函数, 我们通过代入 $\theta = \phi^{-1}(\eta)$ 一定可以转化成关于自然参数 η 的函数 $A(\eta)$ 。

使用规范参数 (自然参数) η 表示的形式称为指数族分布的规范形式 (canonical form), 或者叫自然形式 (natural form), 在规范形式下, 分布的参数是 η 。

部分指数族分布 $\phi(\cdot)$ 是恒等函数, 也就是 $\eta = \phi(\theta) = \theta$, 这样的分布天然具有指数族的规范形式。事实上, 对于指数族中的任意分布, 都可以通过参数转化函数 $\phi(\theta)$ 把原始参数 θ 转化成标准参数 η , 然后以 η 作为模型参数, 进而得到规范形式 (canonical form)。下面我们列举一些属于指数族分布的例子。

5.1.1 伯努利分布

伯努利分布的概率质量函数为:

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}\tag{5.1.8}$$

其中 θ 表示这个概率分布的参数, 我们可以把右侧改写一下:

$$\begin{aligned}p(x|\theta) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\{\ln[\theta^x (1-\theta)^{1-x}]\} \\ &= \exp\{x \ln \theta + (1-x) \ln(1-\theta)\} \\ &= \exp\left\{x \ln \left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta)\right\}\end{aligned}\tag{5.1.9}$$

和公式 (5.1.7) 对比下, 可以发现有:

$$\begin{aligned}\eta &= \phi(\theta) = \ln\left(\frac{\theta}{1-\theta}\right) \\ T(x) &= x \\ A(\eta) &= -\ln(1-\theta) = \ln(1+e^\eta) \\ S(x) &= 0\end{aligned}\tag{5.1.10}$$

函数 $\ln\left(\frac{\theta}{1-\theta}\right)$ 被称为 logit 函数:

$$\eta = \phi(\theta) = \text{logit}(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)\tag{5.1.11}$$

logit 函数的反函数是 logistic 函数。

$$\theta = \text{logistic}(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}\tag{5.1.12}$$

5.1.2 类别分布

伯努利分布是只有两个取值的离散随机变量的概率分布, 当随机变量的取值扩展到多个(大于2个并且有限集)的时候, 就是称为类别分布, 也可以认为是单一观测(一个样本, 一次实验)的多项式分布。其概率质量函数为:

$$p(x|\theta) = \prod_{k=1}^m \theta_k^{x_k} \quad (5.1.13)$$

其中 m 表示变量有 m 种取值, 注意 $x_k \in \{0, 1\}$, 表示变量是否为第 k 个值, 当变量值是第 k 个值时 $x_k = 1$, 否则为 0。 θ_k 表示 $x_k = 1$ 的概率, 并且有 $\sum_{k=1}^m \theta_k = 1$ 。

同样我们需要把上式变型成指数族形式。

$$p(x|\theta) = \prod_{k=1}^m \theta_k^{x_k} = \exp\left\{\sum_{k=1}^m x_k \ln \theta_k\right\} \quad (5.1.14)$$

然而我们注意到, 其中 m 个参数 θ_k 是冗余的, 因为有 $\sum_{k=1}^m \theta_k = 1$, 其中 θ_m 可以用 $\theta_m = 1 - \sum_{k=1}^{m-1} \theta_k$ 表示, 模型只需要 $m-1$ 个参数, 而不需要 m 个参数。

$$\begin{aligned} p(x|\theta) &= \exp\left\{\sum_{k=1}^m x_k \ln \theta_k\right\} \\ &= \exp\left\{\sum_{k=1}^{m-1} x_k \ln \theta_k + \left(1 - \sum_{k=1}^{m-1} x_k\right) \ln\left(1 - \sum_{k=1}^{m-1} \theta_k\right)\right\} \\ &= \exp\left\{\sum_{k=1}^{m-1} x_k \ln\left(\frac{\theta_k}{1 - \sum_{j=1}^{m-1} \theta_j}\right) + \ln\left(1 - \sum_{k=1}^{m-1} \theta_k\right)\right\} \\ &= \exp\left\{\sum_{k=1}^{m-1} x_k \ln\left(\frac{\theta_k}{\theta_m}\right) + \ln\left(1 - \sum_{k=1}^{m-1} \theta_k\right)\right\} \\ &= \exp\left\{\phi(\theta)^T T(x) - A(\theta)\right\} \end{aligned} \quad (5.1.15)$$

上式中的 $\sum_{k=1}^{m-1} x_k \ln\left(\frac{\theta_k}{\theta_m}\right)$ 可以看做是向量 $\phi(\theta) = [\phi(\theta_1), \dots, \phi(\theta_k), \dots, \phi(\theta_{m-1})]$ 和向量 $T(x) = [x_1, \dots, x_k, \dots, x_{m-1}]$ 的内积。和公式 (5.1.7) 对比下, 可以发现有:

$$\begin{aligned} \eta &= \phi(\theta) = [\phi(\theta_1), \dots, \phi(\theta_k), \dots, \phi(\theta_{m-1})], \phi(\theta_k) = \ln\left(\frac{\theta_k}{\theta_m}\right) \\ T(x) &= [x_1, \dots, x_k, \dots, x_{m-1}] \\ A(\eta) &= -\ln\left(1 - \sum_{k=1}^{m-1} \theta_k\right) = \ln\left(\sum_{k=1}^m e^{\eta_k}\right) \\ S(x) &= 0 \end{aligned} \quad (5.1.16)$$

用 η 表示 θ 有:

$$\theta_k = \frac{e^{\eta_k}}{\sum_{j=1}^m e^{\eta_j}} \quad (5.1.17)$$

这个函数被称为 softmax 函数。

5.1.3 泊松分布

泊松 (Poisson) 分布的概率质量函数为:

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \quad (5.1.18)$$

我们同样对它进行改写:

$$\begin{aligned} p(x|\theta) &= \frac{\exp\{\ln[\theta^x e^{-\theta}]\}}{x!} \\ &= \frac{\exp\{x \ln \theta - \theta\}}{x!} \\ &= \exp\{x \ln \theta - \theta - \ln x!\} \end{aligned} \quad (5.1.19)$$

和公式 (5.1.7) 对比可得:

$$\begin{aligned} \eta &= \phi(\theta) = \ln \theta \\ T(x) &= x \\ A(\eta) &= \theta = e^\eta \\ S(x) &= -\ln x! \end{aligned} \quad (5.1.20)$$

η 和 θ 的关系为:

$$\theta = e^\eta \quad (5.1.21)$$

5.1.4 高斯分布

这里我们只考虑单维高斯模型, 高斯模型有两个参数, 分别是均值参数 μ 和方差参数 σ^2 , 高斯分布的概率密度函数为:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad (5.1.22)$$

我们将其转化成指数族的标准形式。

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2 - \ln \sigma - \frac{1}{2}\ln(2\pi)\right\} \end{aligned} \quad (5.1.23)$$

和公式 (5.1.7) 对比可得:

$$\begin{aligned} \eta &= \phi(\theta) = [\mu/\sigma^2, -1/2\sigma^2] \\ T(x) &= [x, x^2] \\ A(\eta) &= \frac{\mu^2}{2\sigma^2} + \ln \sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2) \\ S(x) &= -\frac{1}{2}\ln(2\pi) \end{aligned} \quad (5.1.24)$$

注意单变量高斯模型是含有两个参数的, 所以 η 和 $T(x)$ 都是一个长度为 2 的向量。多维高斯模型同样也属于指数族, 可以自己推导下。

5.1.5 其它常见指数族

请参考: 维基百科 https://en.wikipedia.org/wiki/Exponential_family#Table_of_distributions

5.2 指数族的期望与方差

在数学和统计学中, 矩 (moment) 是对变量分布和形态特点的一组度量。n 阶矩被定义为一变量的 n 次方与其概率密度函数 (Probability Density Function, PDF) 之积的积分。在文献中 n 阶矩通常用符号 μ_n 表示, 直接使用变量计算的矩被称为原始矩 (raw moment), 移除均值后计算的矩被称为中心矩 (central moment)。变量的一阶原始矩等价于数学期望 (expectation)、二至四阶中心矩被定义为方差 (variance)、偏度 (skewness) 和峰度 (kurtosis)。

—摘自百度百科

通俗的讲, 矩 (moment) 是描述一个随机变量的一系列指标, 变量的期望 (Expectation, 或者叫均值, Mean) 和方差 (Variance) 属于其中最简单的两个指标, 这里只讨论这两种。

指数族有一个特点, 就是我们可以通过对 $A(\eta)$ 求导来得到 $T(x)$ 的矩, 比如其一阶导数是 $T(x)$ 的期望, 二阶导数是 $T(x)$ 的方差。在指数族分布中 $A(\eta) = \ln \int h(x) \exp\{\eta^T T(x)\} dx$, 其一阶导数为:

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{d}{d\eta} \left\{ \ln \int h(x) \exp\{\eta^T T(x)\} dx \right\} \\ &= \frac{\int T(x) \exp\{\eta^T T(x)\} h(x) dx}{\int \exp\{\eta^T T(x)\} h(x) dx} \\ &= \int T(x) \exp\{\eta^T T(x) - A(\eta)\} h(x) dx \\ &= \mathbb{E}[T(x)] \end{aligned} \tag{5.2.1}$$

我们看到 $A(\eta)$ 的一阶导数正好等于 $T(x)$ 的期望 (均值), 对于伯努利分布、多项分布、泊松分布、高斯分布等这些 $T(x) = x$ 的分布来说, $T(x)$ 的均值就是分布的均值。

比如上面的示例中, 对于伯努利分布, 有 $A(\eta) = \ln(1 + e^\eta)$, 其一阶导数为:

$$\begin{aligned} \frac{dA}{d\eta} &= \frac{d}{d\eta} \ln(1 + e^\eta) \\ &= \frac{e^\eta}{1 + e^\eta} \\ &= \frac{1}{1 + e^{-\eta}} \\ &= \mu \end{aligned} \tag{5.2.2}$$

对于高斯分布有:

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-2\eta_2) \tag{5.2.3}$$

其中, $\eta_1 = \mu/\sigma^2$, $\eta_2 = -1/2\sigma^2$, 我们计算 η_1 的偏导数:

$$\begin{aligned} \frac{\partial A}{\partial \eta_1} &= \frac{\eta_1}{2\eta_2} \\ &= \frac{\mu/\sigma^2}{1/\sigma^2} \\ &= \mu \end{aligned} \tag{5.2.4}$$

现在我们看下 $A(\eta)$ 的二阶导数:

$$\begin{aligned}
 \frac{d^2 A}{d\eta^2} &= \int T(x) \exp\{\eta T(x) - A(\eta)\} (T(x) - A'(\eta)) h(x) dx \\
 &= \int T(x) \exp\{\eta T(x) - A(\eta)\} (T(x) - \mathbb{E}[T(x)]) h(x) dx \\
 &= \int T(x)^2 \exp\{\eta T(x) - A(\eta)\} h(x) dx - \mathbb{E}[T(x)] \int T(x) \exp\{\eta T(x) - A(\eta)\} h(x) dx \\
 &= \mathbb{E}[T(x)^2] - (\mathbb{E}[T(x)])^2 \\
 &= \text{Var}[T(x)]
 \end{aligned} \tag{5.2.5}$$

$A(\eta)$ 的二阶导数正好是 $T(x)$ 的方差, 对于 $T(x) = x$ 的分布, 就是分布的方差。

比如对于高斯分布, 对于 η_1 的二阶偏导数为:

$$\begin{aligned}
 \frac{\partial A}{\partial \eta_1} &= -\frac{1}{2\eta_2} \\
 &= \sigma^2
 \end{aligned} \tag{5.2.6}$$

总结一下, 对于指数族分布, 我们可以通过对 $A\eta$ 求导来计算分布中 $T(x)$ 期望和方差, 当然通过高阶导数还能计算出更多的矩 (Moment)。

此外, 我们发现函数 $A\eta$ 的二阶导数是 $T(x)$ 的方差, 我们都知道方差肯定是大于等于 0 的, 一个函数的二阶导数大于等于 0, 证明这个函数是一个凸函数 (convex, 碗状的), 对于凸函数, 一阶导数和参数 η 之间是一一对应关系, 并且这种对应关系是可逆的。我们定义 $A(\eta)$ 的一阶导数用符号 μ 表示, 则有 $u \triangleq \mathbb{E}[T(x)]$, μ 和 η 之间的关系可以用如下函数表示:

$$\mu = \frac{dA}{d\eta} \tag{5.2.7}$$

并且这个函数是可逆的, 也就是说已知 μ 就能求出 η ; 反过来, 已知 η 就能求出 μ 。比如对于伯努利分布:

$$\begin{aligned}
 \eta &= \frac{\mu}{1 - \mu} \\
 \mu &= \frac{1}{1 + e^{-\eta}} \text{ (logistic function)}
 \end{aligned} \tag{5.2.8}$$

对于多项式分布:

$$\begin{aligned}
 \eta_i &= \ln \left(\frac{\mu_i}{1 - \sum_{i=1}^{m-1} \mu_i} \right) \\
 \mu_i &= \frac{e^{\eta_i}}{\sum_{j=1}^m e^{\eta_j}} \text{ (softmax function)}
 \end{aligned} \tag{5.2.9}$$

由于 μ 和 η 是可逆的, 所以对于指数族分布, 也可以用 μ 去定义分布模型, 也就是用 μ 去当做模型的参数。事实上, 我们常见的分布都是这么做的, 比如伯努利分布、高斯分布等等。

5.3 最大似然估计

现在我们讨论下指数族的最大似然估计, 我们知道指数族的自然参数 η 和特定分布的原始参数 θ 是一一对应的, 二者是存在可逆关系的, 所有只要我们能估计出自然参数 η , 就一定能通过逆函数 $\phi(\cdot)^{-1}$ 得到分布的真实参数 θ 的估计值, 也就是说对于指数族, 我们只需要推导自然参数的估计量 $\hat{\eta}$ 即可。

我们用符号 \mathcal{D} 表示随机变量的一个观测样本集, 样本集的规模是 N , 并且样本集是满足 IID(独立同分布) 的。

首先回顾一下指数族分布的标准形式：

$$p(x|\eta) = \exp\{\eta^T T(x) - A(\eta) + S(x)\} \quad (5.3.1)$$

我们知道样本的似然就是所有样本发生的联合概率：

$$\begin{aligned} L(\eta; \mathcal{D}) &= p(\mathcal{D}|\eta) \\ &= p(x_1, \dots, x_N|\eta) \\ &= \prod_{i=1}^N p(x_i|\eta) \\ &= \prod_{i=1}^N \exp\{\eta^T T(x_i) - A(\eta) + S(x_i)\} \\ &= \exp\{\eta^T \sum_{i=1}^N T(x_i) - NA(\eta) + \sum_{i=1}^N S(x_i)\} \end{aligned} \quad (5.3.2)$$

对比一下，我们发现指数族分布的联合概率仍然是指数族：

$$\begin{aligned} T(x) &\implies \sum_{i=1}^N T(x_i) \\ A(\eta) &\implies NA(\eta) \\ S(x) &\implies \sum_{i=1}^N S(x_i) \end{aligned} \quad (5.3.3)$$

现在我们为似然函数加上对数，得到对数似然函数：

$$\begin{aligned} \ell(\eta; \mathcal{D}) &= \ln L(\eta; \mathcal{D}) \\ &= \eta^T \sum_{i=1}^N T(x_i) - NA(\eta) + \sum_{i=1}^N S(x_i) \end{aligned} \quad (5.3.4)$$

我们对参数 η 求导：

$$\nabla_\eta \ell = \sum_{i=1}^N T(x_i) - N \nabla_\eta A(\eta) \quad (5.3.5)$$

上述公式中的 $\nabla_\eta A(\eta)$ 表示对函数 $A(\eta)$ 关于 η 求导，这里函数 $A(\eta)$ 是一个关于 η 的函数。我们令这个导数为 0，可得：

$$\nabla_\eta A(\eta) = \frac{1}{N} \sum_{i=1}^N T(x_i) \quad (5.3.6)$$

由公式 (5.2.1) 我们知道 $A(\eta)$ 的一阶导数等于 $T(x)$ 的期望 $\mathbb{E}[T(x)]$ ，即 $\nabla_\eta A(\eta) = \mathbb{E}[T(x)]$ 。我们令 $\mu \triangleq \nabla_\eta A(\eta) = \mathbb{E}[T(x)]$ ，结合公式公式 (5.3.6) 有：

$$\mu_{ML} = \mathbb{E}[T(x)] = \frac{1}{N} \sum_{i=1}^N T(x_i) \quad (5.3.7)$$

从公式 (5.3.7) 可以看出，指数族分布理论期望值(均值参数)等于样本的期望值(平均值)。均值参数的最大似然估计值，只和样本的统计量 $\sum_{i=1}^N T(x)$ 有关，而不再依赖样本的其它信息，所以 $\sum_{i=1}^N T(x)$ (或者说 $T(x)$) 是指数族的充分统计量。对于满足 $T(x) = x$ 的分布，比如伯努利分布、多项式分布、泊松分布等等，样本

的均值就是 $T(x)$ 的均值, 样本的均值就是均值参数的最大似然估计值。同理, 对于单变量的高斯分布, 样本的方差就是方差参数的最大似然估计值。

我们知道 μ 和 η 是一一对应的, 可以通过一个函数进行互相计算, 最大似然估计给出了 μ_{ML} 的估计值, 我们就是可以换算出 η_{ML} 。前文说过, 事实上对于很多常见分布是直接用 μ 作为参数的, 所以有了最大似然的估计值 μ_{ML} 就直接是模型的参数估计值。公式 (5.3.7) 也直接说明了当样本数量趋近无穷大时, 最大似然估计值和 μ 的真实值是一致的。

5.4 最大似然估计与 KL 散度的关系

本节我们讨论一下指数族的最大似然估计和 KL 散度的关系, 在开始前我们先回顾一下 KL 散度的定义。

KL 散度 (Kullback–Leibler divergence) KL 散度 (Kullback–Leibler divergence, 简称 KLD), 在信息系统中称为相对熵 (relative entropy), 在连续时间序列中称为 randomness, 在统计模型推断中称为信息增益 (information gain), 也称信息散度 (information divergence)。KL 散度是两个概率分布 P 和 Q 差别的非对称性的度量, 可以理解成是用来度量两个分布的相似性。一般用符号 $D_{KL}(P \parallel Q)$ 表示。

对于离散随机变量, 概率分布 P 和 Q 的 KL 散度按照下式定义:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \quad (5.4.1)$$

或者:

$$D_{KL}(P \parallel Q) = - \sum_i P(i) \ln \frac{Q(i)}{P(i)} \quad (5.4.2)$$

即按照概率 P 求得 P 和 Q 的对数商的平均值 (期望), 其中对数的底可以是任意的。KL 散度仅当概率 P 和 Q 各自总和均为 1, 且对于任何 i 皆满足 $Q(i) > 0, P(i) > 0$ 时才有定义。式中出现 $0 \ln 0$ 的情况, 其值按 0 处理。

对于连续随机变量, 其概率分布 P 和 Q 可按积分方式定义为:

$$D_{KL}(P \parallel Q) = \int P(x) \ln \frac{P(x)}{Q(x)} dx \quad (5.4.3)$$

相对熵的值为非负数 $D_{KL}(P \parallel Q) \geq 0$, 由吉布斯不等式可知, 当且仅当 $P = Q$ 时 $D_{KL}(P \parallel Q)$ 为零。尽管从直觉上 KL 散度是个度量或距离函数, 但是它实际上并不是一个真正的度量或距离。因为 KL 散度不具有对称性: 从分布 P 到 Q 的距离通常并不等于从 Q 到 P 的距离。

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P) \quad (5.4.4)$$

我们可以根据信息理论量重写对数似然函数, 其中 x_m 为随机变量的一个可能取值, $\hat{p}_{\mathcal{D}}(x_m)$ 表示在样本中变量值为 x_m 的样本出现的比例, 乘以 N 后就是出现的次数。我们用 $\hat{p}_{\mathcal{D}}(x_m)$ 表示从样本中的到的经验分布。

此外, 定义 n_m 表示样本中 x_m 出现的次数, 则有 $n_m = N\hat{p}_{\mathcal{D}}(x_m)$ 。

$$\begin{aligned}
 \ell(\eta; \mathcal{D}) &= \sum_{i=1}^N \log p(x^{(i)}; \eta) \\
 &= \sum_{x_m \in \mathcal{X}} \log p(x_m; \eta)^{n_m} \\
 &= N \sum_{x_m \in \mathcal{X}} \hat{p}_{\mathcal{D}}(x_m) \log p(x_m; \eta) \\
 &= N \sum_{x_m \in \mathcal{X}} \hat{p}_{\mathcal{D}}(x_m) [\log p(x_m; \eta) - \log \hat{p}_{\mathcal{D}}(x_m) + \log \hat{p}_{\mathcal{D}}(x_m)] \\
 &= N \sum_{x_m \in \mathcal{X}} \hat{p}_{\mathcal{D}}(x_m) [\log \frac{p(x_m; \eta)}{\hat{p}_{\mathcal{D}}(x_m)} + \log \hat{p}_{\mathcal{D}}(x_m)] \\
 &= N \underbrace{\sum_{x_m \in \mathcal{X}} \hat{p}_{\mathcal{D}}(x_m) \log \frac{p(x_m; \eta)}{\hat{p}_{\mathcal{D}}(x_m)}}_{\text{负的 KL 散度}} + N \underbrace{\sum_{x_m \in \mathcal{X}} \hat{p}_{\mathcal{D}}(x_m) \log \hat{p}_{\mathcal{D}}(x_m)}_{\text{经验分布的信息熵}} \\
 &= N(H(\hat{p}_{\mathcal{D}}) - D(\hat{p}_{\mathcal{D}} \parallel p(x; \eta)))
 \end{aligned} \tag{5.4.5}$$

我们可以忽略熵项, 因为它是经验分布的函数, 与参数 η 无关, 在极大化过程中其值是固定值。因此, **最大化似然等同于最小化经验分布与真实分布的信息差异** $D(\hat{p}_{\mathcal{D}} \parallel p(\cdot; \eta))$ 。

回想一下, 当两个分布是相同的分布时, KL 散度为零。在多项式情况下, 由于我们在有限空间 \mathcal{X} 上优化了所有分布的集合, 我们可以精确地匹配分布, 例如, 令 $p(\cdot; \eta) = \hat{p}_{\mathcal{D}}$, 即可使 KL 散度为零, 得到精确匹配。然而, 在大多数有趣的问题中, 我们无法完全匹配数据分布(如果可以, 我们只会过度拟合)。相反, 我们通常优化由 η 参数化的受限类分布, 来得到近似解。

线性回归模型

在统计分析中，经常需要分析变量之间的关系，最常见的场景就是研究某个变量如何取决于其它一个或多个变量，通常会把研究的目标变量称为响应变量、输出变量、被解释变量，把影响目标变量的其它变量称为预测变量、输入变量、解释变量等。习惯上，用符号 X 表示输入变量，用符号 Y 表示输出变量，通常二者间存在一定的关系，统计学家研究内容就是找到二者之间合适的函数关系， $y = f(x)$ ，然而二者之间最真实的关系是不得知的，统计学家也只是尽可能的找到一个近似的关系。线性回归模型是统计学中最基本的模型，也是机器学习领域的入门模型，得益于它的简单，应用十分广泛。本书的主题，广义线性模型就是线性回归模型的扩展，在讨论广义线性模型之前，先简单回顾一下线性回归模型。

6.1 最小二乘

6.1.1 最小误差

通常情况下，机器学习就是要找到输出变量 Y 和输入变量 X 的函数关系，变量之间最简单的函数关系，就是线性函数关系。在线性回归模型中，假设 Y 与 X 之间是一个线性函数的关系。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (6.1.1)$$

由于输入特征数据 X 通常是多维的，所以 $f(x)$ 是一个多元一次函数，其中 x_j 是第 j 维输入数据， x_j 可以是任意实数值。 $\beta_j (j > 0)$ 是 x_j 的系数， β_0 是这个线性方程的截距，我们把 β_j 看做这个函数的未知参数，其值暂时是未知的。公式 (??) 看上去不是很简洁，通常为了公式的简洁表达，我们令 $\beta^T = [\beta_0, \beta_1, \dots, \beta_p]$, $x^T = [1, x_1, x_2, \dots, x_p]$ ，然后把公式 (??) 看成是特征向量 x 和参数向量 β 的内积形式。线性回归模型的简洁表示为：

$$y = \beta_0 \times 1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = x^T \beta \quad (6.1.2)$$

注解：为了满足向量内积的形式，我们人为的增加一列特征数据 $x_0 = 1$ ，所有输入样本的 x_0 都为常量 1。

fg_29_1 是单一维度输入变量的线性回归模型的图形化展示，为了方便图形化展示，我们假设输入特征变量 X 只有一维。图中蓝色的点代表一些 (x_i, y_i) 的样本点，下标 i 是样本点的编号。线性回归模型的本质就

是找到一条直线 $y = x^T \beta$ (比如图中的红色直线)，并且这条直线和样本点的“走势”是一致的，这样我们就可以用这条直线去预测新的样。比如根据图中蓝色样本点的分布，我们找到红色直线和样本的“走势”一致，这样当输入一个新的 x_{new} 时，就用直线上的点 $y_{new} = x_{new}^T \beta$ 作为预测值 $\hat{y} = y_{new} = x_{new}^T \beta$ 。

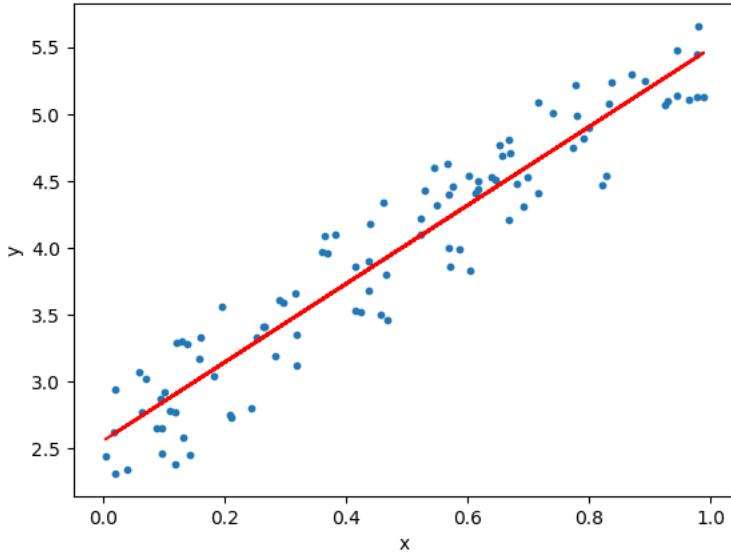


图 6.1.1: 单一输入特征的回归模型

然而空间上存在无数条直线，要如何确定和样本点“走势”相同的直线呢？我们的最终目的是用这条之间预测新的样本点，那么理论上预测最准的直线是最优的直线。对于一条样本数据 (x_i, y_i) ，模型的预测值是 $\hat{y}_i = x_i^T \beta$ ，显然，最优的直线就是 **预测误差最小的直线**。我们把样本的真实值 y_i 和预测值 \hat{y}_i 之间的差值定义成残差 (residual)，我们的目标就是找到一条令所有观测样本残差最小的直线。通常我们使用所有样本残差的平方和 (residual sum of squares, RSS) 做为整体的误差。

$$J(\beta) = \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (6.1.3)$$

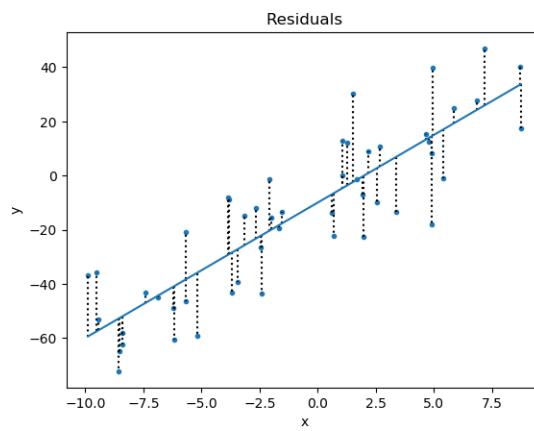


图 6.1.2: 线性回归的残差

我们认为令 RSS 取得最小值的直线的最优的直线，所以我们通过极小化 RSS 来确定这条最优的直线，由于

直线是由参数 β 决定的, 所以要确定这条直线就是等价于找到参数 β 的值。因此:

$$\hat{\beta} = \arg \max_{\beta} \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \arg \max_{\beta} \sum_{i=1}^N (x_i^T \beta - y_i)^2 \quad (6.1.4)$$

注解: 通常机器学习的过程, 都是先根据场景或数据定义一个参数化的模型函数 $y = f(x, \beta)$, 模型函数是对单条数据样本的建模。但它是含有未知参数的, 我们需要找到一个最优的参数值使得这个模型函数尽可能好的拟合数据样本。因此就需要定义一个评价不同参数值模型好坏的标准或者说函数, 通常我们称这个评价函数为目标函数 (object function), 然后通过极大 (小) 化目标函数求得参数的最优解。比如似然函数就是目标函数的一种, 除此之外, 还可以定义某种损失 (误差) 函数 (cost function loss function | error function) 作为目标函数, 比如线性回归的平方损失、逻辑回归的交叉熵损失等等。

以“残差平方和最小”确定直线位置的方法被称为最小二乘法。用最小二乘法除了计算比较方便外, 得到的估计量还具有优良特性, 这种方法对异常值非常敏感。

6.1.2 参数估计

显然对于线性回归模型, RSS 是一个关于 β 的二次函数, 我们知道二次函数一定是存在唯一的一个极值点的, 所以参数 β 一定存在唯一解, 并且在极值点函数的导数为 0。所以我们可以直接求出 RSS 的导数, 并令导数为 0 的方法求得 β 。

$$\frac{\partial J}{\partial \beta} = \sum_{i=1}^N 2x_i(x_i^T \beta - y_i) \quad (6.1.5)$$

上述偏导结果中包含所有样本的求和符号 $\sum_{i=1}^N$, 为了简单表达我们用矩阵和向量乘积的方式替换求和符号。我们用符号 X 表示训练样本集中所有的输入数据的矩阵, 用符号 y 表示训练样本集中所有输出数据的向量。上述偏导用矩阵符号表示为:

$$\frac{\partial J}{\partial \beta} = 2X^T(X\beta - y) \quad (6.1.6)$$

然后我们令导数为 0, 可以得到:

$$X^T X \beta = X^T y \quad (6.1.7)$$

公式 (??) 通常被称为正规方程组 (normal equations), 理论上, 我们可以根据这个等式得到参数 β 的解析解

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (6.1.8)$$

然而, 这个解析解需要求矩阵 $X^T X$ 的逆矩阵, 矩阵存在逆矩阵需要满足两个条件: (1) 是方阵, (2) 是满秩的。虽然是 $X^T X$ 方阵, 但是未必满秩, 那么也就不存在逆矩阵, 当逆矩阵不存在时也没办法计算解析解。当矩阵 X 存在 (行或列) 共线性时, $(X^T X)$ 一定是不满秩的。

$(X^T X)$ 不存在逆矩阵不代表参数 β 无解, 上面已经讲过损失函数 $J(\beta)$ 是二次函数, 一定存在唯一的极值点, 所以参数 β 一定有全局最优解的。当无法求得解析解时, 我们可以使用迭代法求解, 比如基于一阶导数的梯度下降法和基于二阶导数的牛顿法。

6.2 线性回归的概率解释

现在我们尝试在概率的框架下解释线性回归模型。在概率的框架下, 认为输入变量 X 和输出变量 Y 都是随机变量, 两个变量的联合概率为

$$P(X, Y) = P(X)P(Y|X) \quad (6.2.1)$$

在回归问题中, 是给定输入 x , 模型输出 y 的值, 特征变量 X 的值是已知的确定的, 所以不需要边缘概率 $P(X)$, 只需要得到条件概率 $P(Y|X)$ 即可。换句话说, 不需要对联合概率 $P(X, Y)$ 进行建模, 只需要建模条件概率 $P(Y|X)$ 。在概率的框架下, 用条件概率 $P(Y|X) = f(x)$ 去表达这种关系, 即当 $X = x$ 时, 变量 $Y = y$ 的概率为 $P(Y = y|X = x)$ 。

6.2.1 高斯分布

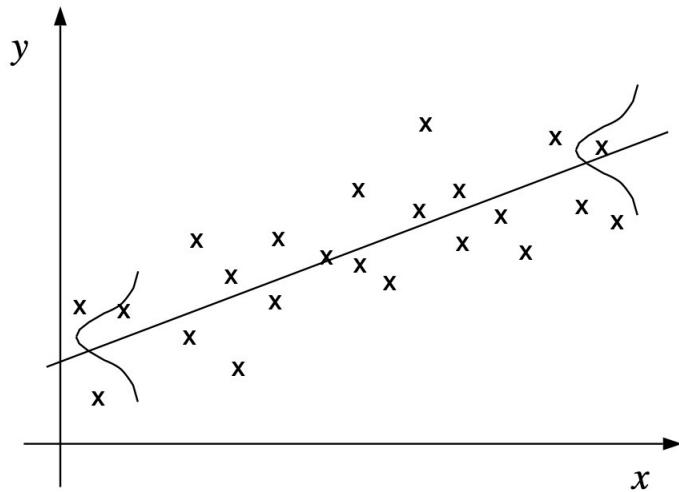


图 6.2.1: 线性回归模型表示成条件均值函数加上一个高斯噪声。

观察 fg_29_12, 变量 Y 的值, 可以看做是线性预测器 $x^T \beta$ 的值加上一个误差项 (噪声) ϵ 。

$$y = x^T \beta + \epsilon \quad (6.2.2)$$

对于噪声, 最普遍常用的就是高斯噪声, 这里假设 ϵ 是一个均值为 0, 方差为 σ^2 的高斯噪声, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 。线性部分 $x^T \beta$ 是一个数值量, 不是随机量。期望为 0 的高斯项 ϵ 加上一个数值项 $x^T \beta$ 得到的就是一个期望值为 $x^T \beta$ 随机变量, 因此输出变量 Y 是一个高斯变量。

$$Y \sim \mathcal{N}(x^T \beta, \sigma^2) \quad (6.2.3)$$

因此条件概率 $P(Y|X)$ 是均值为 $x^T \beta$ 方差为 σ^2 的高斯分布, 条件概率分布 $P(Y|X)$ 的概率密度函数为:

$$p(y|x; \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \beta^T x)^2\right\} \quad (6.2.4)$$

注解: 实际上 Y 的概率分布要根据数据的实际分布确定, 并不是一定要高斯分布。只不过在线性回归这个模型中我们假设 Y 是服从高斯分布的。如果你的数据不是(近似)高斯分布, 那么就不应该使用线性回归模

型。在后面的章节中我们会介绍当 Y 是其它分布时, 应该怎么处理, 在广义线性模型的框架下, 输出变量 Y 可以扩展到指数族分布。

在传统线性回归模型中, 方差项 σ^2 被假设为常量 1。 β 是模型的参数, 是需要利用观测数据进行学习的。高斯版的线性回归模型, 是一个概率模型, 可以使用最大似然估计参数 β , 下一节详细阐述如何利用最大似然估计参数 β 。

在得到参数 β 的最大似然估计值 $\hat{\beta}_{ML}$ 后, 可以用高斯变量 Y 的期望值作为模型的预测值, 模型输入一个 x 值, 输出 $\hat{y} = \mathbb{E}[P(Y|X = x; \hat{\beta}_{ML})] = x^T \hat{\beta}_{ML}$ 。

实际上, 线性回归模型比我们看上去的要广泛, 线性组合部分 $x^T \beta$ 只要求对 β 是线性的, 并不要求对 x 是线性的, 所以可以为 x 加上一个非线性的函数 $\phi(\cdot)$ 使模型具有拟合非线性数据的能力。

$$y = \beta^T \phi(x) + \epsilon = \beta^T x' + \epsilon \quad (6.2.5)$$

$\phi(x)$ 可以看做是对特征数据的预处理 $x \Rightarrow \phi(x)$, 转化之后的 $x' = \phi(x)$ 作为模型的输入特征, 并不影响线性回归模型的定义和计算。

高斯假设的线性回归模型是建立在两个很强的假设之上的, (1) 条件概率 $P(Y|X)$ 是高斯分布, 并且 (2) 不同的 x 条件下方差是相同的。然而这种假设在很多时候是不满足的, 尤其是第 (2) 点, 这也是线性回归的模型的局限性。

6.2.2 参数估计

假设我们有一个成对的观测数据集 $\mathcal{D} = \{(x_i, y_i); i = 1, 2, \dots, N\}$, 其中 x_i 是一条输入变量 X 的观测样值, y_i 是对应的输出变量 Y 的观测值, 注意 x_i 是一个 p 维的向量 (vector), 而 y_i 是一个标量 (scalar)。样本集中的样本都是满足独立同分布 (IID) 的。样本集的联合概率可以写成:

$$\begin{aligned} P(\mathcal{D}) &= \prod_{i=1}^N P(y_i|x_i; \beta) \\ &= 2\pi\sigma^{2-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i^T \beta)^2\right\} \end{aligned} \quad (6.2.6)$$

我们知道观测样本集的联合概率就是似然函数, 我们可以通过最大似然估计法估计出模型的未知参数 β , 为了计算简单, 通常我们采用极大化对数似然函数的方法估计参数。线性回归模型的对数似然函数为:

$$\ell(\beta; x, y) = \underbrace{-\frac{N}{2} \ln(2\pi\sigma^2)}_{\text{常量}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i^T \beta)^2 \quad (6.2.7)$$

由于我们假设方差 σ^2 是常量, 所以上述公式的第一项是一个常量, 在极大化对数似然函数时不影响最终的求解, 所以是可以去掉的。

$$\begin{aligned} \hat{\beta}_{ML} &= \arg \max_{\beta} \ell(\beta; x, y) \\ &\triangleq \arg \max_{\beta} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i^T \beta)^2 \right\} \end{aligned} \quad (6.2.8)$$

我们发现这和最小二乘法的损失函数 $J(\beta)$ 是等价的, 对数似然函数中的 $\sum_{i=1}^N (y_i - x_i^T \beta)^2$ 就是残差的平方和 (residual sum of squares, RSS)。同理, 我们可以使用迭代法求的最优解。

广义线性模型

线性回归模型是算法领域的入门模型，是每个新人入门的必须课。在线性回归模型中假设响应变量 Y 是由两部分组成：系统组件 (system component) 和误差组件 (error component)。其中系统组件是一个线性预测器 $\eta = x^T \beta$ ，误差组件是一个服从标准正态分布的随机量 $\epsilon \sim \mathcal{N}(0, 1)$ 。

$$\begin{aligned} y &= \beta_0 + x_1 \beta_1 + x_2 \beta_2 + \cdots + x_p \beta_p + \epsilon \\ &= x^T \beta + \epsilon \end{aligned} \tag{7.1}$$

虽然线性预测器 η 是一个数值变量，但误差项 ϵ 是一个高斯随机变量，响应变量 Y 作为二者的加和，也是一个高斯随机量，并有 $\mathbb{E}[Y] = \eta = x^T \beta$ 。

$$Y \sim \mathcal{N}(x^T \beta, 1) \tag{7.2}$$

因此，在线性回归中，可以把响应变量 Y 解释成一个高斯随机变量。

那么，响应变量 Y 是不是可以解释成其它概率分布的随机变量呢？比如，伯努利变量、泊松变量等等，答案显然是可以的。如果把 Y 解释成伯努利变量，得到就是逻辑回归模型，如果把 Y 解释成泊松变量，得到的就是泊松回归模型，等等，还有很多种类的回归模型。实际上，对于常见的概率分布，都有对应的回归模型。但是在早期，这些回归模型，都是独立开发，独立应用的。虽然这些模型都是使用最大似然估计进行参数估计，但每种模型都需要独立对似然函数就行求导等操作。

直到 1972 年，John Nelder 和 Robert Wedderburn 提出了一种统一的框架：广义线性模型 (Generalized linear models,GLM)。GLM 将多种统计回归模型归一到一个框架下，并且提出了一个统一的参数估计算法：迭代重加权最小二乘法 (iteratively reweighted least squares method,IRLS)。在 GLM 框架中，误差项的概率分布可以是指数族分布中的任意一种，因此，在 GLM 中，响应变量 Y 可以解释成指数族分布中的任意一种。线性预测器部分保持不变，仅仅是误差项扩展到了指数族分布，因此称为 **广义线性模型**。

本章我们正式讨论广义线性模型，GLM 是建立在指数族分布的技术上，因此我们首先介绍下指数族概率分布的标准形式，然后再给出 GLM 的定义。

7.1 指数族分布

7.1.1 自然指数族

在节 5.1 我们讨论了指数族分布, 所有指数族的概率密度 (质量) 函数都可以写成如下的形式。

$$p(y|\theta) = \exp\{\theta^T T(y) - A(\theta) + S(y)\} \quad (7.1.1)$$

其中 θ 称为自然参数 (natural parameter) 或者规范参数 (canonical parameter), 其代表了模型中所有的未知参数。通常指数族分布会有两个参数, 一个代表位置 (location) 的参数, 一个代表尺度 (scale) 的参数。位置参数和分布的期望相关, 尺度参数和分布的方差相关。

本章我们讨论的广义线性模型并不使用上述形式的指数族, 而是指数族的一个子集, 自然指数族 (natural exponential family), 自然指数族是满足 $T(y) = y$ 的指数族。

$$p(y|\theta) = \exp\{\theta^T y - A(\theta) + S(y)\} \quad (7.1.2)$$

指数族的这个形式被称为自然形式 (natural form) 或者规范形式 (canonical form), 虽然指数族中大部分分布都可以写成上述自然形式, 但是也有一些分布, 虽然属于指数族, 但是不能写成上述自然形式, 比如对数正态分布 (LogNormal distribution)。

重要: 这里有个容易搞混的地方, 虽然参数 θ 叫规范参数 (canonical parameter), 但是必须满足 $T(y) = y$ 的形式才叫做规范形式 (canonical form)。

指数族分布中, 有的分布只有一个参数, 有的分布有两个参数, 规范参数 θ 包含了分布所有的原始参数, 当分布只有一个参数时, θ 就是一个标量参数, 当分布有两个参数时, θ 就是一个二元向量参数。并且指数族分布的两个参数分别和分布的期望与方差相关, 分别代表了位置 (location) 与尺度 (scale)。

规范参数 θ 和指数族分布的原始参数是存在一一映射的, 规范参数 θ 可以是一个标量参数, 也可以包含两个参数的向量, 对于单参数的指数族分布, 原始参数通常就是分布的期望 μ , 此时 θ 是 μ 的函数。对于双参数的指数族分布, 原始参数通常就是分布的期望 μ 和方差 σ^2 , 此时 θ 是含有两个参数的向量, 并且 θ 是期望 μ 和 σ^2 的函数。

指数族的规范形式 (公式 (7.1.2)) 规范参数 θ 包含了所有参数, 这不方便处理。因此我们把参数拆分一下, 在规范形式的基础上再引入一个代表尺度 (scale) 的参数 ϕ 。

$$p(y|\theta) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (7.1.3)$$

这种形式的指数族通常被称为指数分散族 (exponential dispersion family, EDF), $a(\phi)$ 称为分散函数 (dispersion function), 是已知的。 ϕ 称为分散参数 (dispersion parameter)。 θ 仍然称作自然参数 (natural parameter) 或者规范参数 (canonical parameter)。

公式 (7.1.3) 形式的指数族, 其实就是对参数 θ 进行了拆分, 把期望参数和方差参数拆分开。使得自然参数 θ 仅和期望 μ 相关, 分散参数 ϕ 和分布的方差参数相关。分拆后, 规范参数 θ 仅和分布的期望参数 μ 相关, 并且和 μ 之间存在一一映射的函数关系, 换句话说, θ 和 μ 可以互相转化。

$$\begin{aligned} \theta &= f(\mu) \\ \mu &= f^{-1}(\theta) \end{aligned} \quad (7.1.4)$$

分散参数 (dispersion parameter)

在最初的 GLM 论文中 (Nelder and Wedderburn, 1972) 把 $a(\phi)$ 称为尺度因子 (scale factor), 并且没有给参数 ϕ 单独命名。后来在 1974 年 Royal Statistical Society 发布了首个 GLM 的软件工具包 (Generalized Linear Interactive Modelling, GLIM), 在 GLIM 中把 $a(\phi)$ 定义成:

$$a(\phi) = \frac{\phi}{w} \quad (7.1.5)$$

其中 w 是样本的先验权重 (prior weight), ϕ 称为尺度参数 (scale parameter), 这就是导致了对 ϕ 命名产生了歧义。因为“scale”这个词在统计学还有其它用法, 容易产生混淆, 所以在 1980s(McCullagh and Nelder) 初版的 GLM 书籍中, 把 ϕ 命名为“dispersion parameter”, 后来也就沿用了这种叫法。但是由于 GLIM 流行了很久, 导致“scale”的叫法还存在很多资料中。

在很多 GLM 的工具包中, 都会把 $a(\phi)$ 定义成如下形式:

$$a(\phi)_i = \frac{\phi}{w_i} \quad (7.1.6)$$

其中 w_i 是观测样本的权重, 一般是已知的。不同的样本可以拥有不同的权重值, 比如进行参数估计时, 对于某些样本设置成 $w_i = 0$, 这就相当于抛弃了这些样本。

$a(\phi)$ 的函数形式并没有严格的要求, 其函数形式并不重要, 本质上 $a(\phi)$ 就是代表了分散参数 (dispersion parameter), 所以通常直接令 $a(\phi) = \phi$ 。如果你需要不同的样本有不同的值, 那么就使用公式 (7.1.6) 的形式。公式 (7.1.6) 的形式中, 当所有样本拥有相同的 w 权重时, 就等价于 $a(\phi) = \phi$ 。

通常在 GLM 中, 只有参数 θ 作为模型的未知参数, 此时称为单参数模。单参数模型指的是模型中只有 θ 是未知参数, 而 ϕ 是已知的, 反之, 如果 θ 和 ϕ 都是未知的, 则成为双参数模型。指数族的某些分布, 是不存在分散参数的, 比如对于伯努利分布、泊松分布、二项式分布等等离散分布。

自然参数 θ 和分布的期望相关, 它是期望的一个函数。而分散参数 ϕ 和分布的方差相关, 它影响着方差的大小, 具体的关系在之后的内容中会详细说明。

累积函数 (cumulant function)

我们知道在公式 (7.1.1) 的指数族形式中 $A(\theta)$ 称为累积函数 (cumulant function), 可以用 $A(\theta)$ 的导数求出分布的矩, 一阶导数是分布的期望, 二阶导数是分布的方差。然而在 GLM 中我们使用的是公式 (7.1.3) 的形式, 其中 $b(\theta)$ 本质上就是 $A(\theta)$, 二者关系是:

$$A(\theta) = \frac{b(\theta)}{a(\phi)} \quad (7.1.7)$$

所以我们同样把 $b(\theta)$ 被称为累积函数 (cumulant function), 并且它同样和分布的矩 (moments) 有关。

$$\mathbb{E}[Y] = b'(\theta) = \mu \quad (7.1.8)$$

$$V(Y) = A''(\theta) = a(\phi)b''(\theta) \quad (7.1.9)$$

由于 $b(\theta)$ 是在 $A(\theta)$ 的基础上拆分出去 $a(\phi)$, 所以 $b(\theta)$ 的二阶导数不再分布的方差, 需要再乘上 $a(\phi)$ 才能得到分布的方差。

方差结构

在 EDF (指数分散族, Exponential Dispersion Family) 中, 分布的方差可以表示成两部分的乘积 (公式 (7.1.9)), 一部分是分散函数 $a(\phi)$, 另一部分是累计函数的二阶导数 $b''(\theta)$ 。

$$V(Y) = b''(\theta)a(\phi) = \nu(\mu)a(\phi) \quad (7.1.10)$$

累积函数 $b(\theta)$ 是一个关于 θ 的函数, 其二阶导数要么是一个常数, 要么是一个关于自然参数 θ 的函数。而自然参数 θ 和均值参数 μ 存在一一对关系, 所以一定可以把 θ 替换成 μ 。

我们定义累计函数 $b(\theta)$ 的二阶导数为方差函数 (variance function), 方差函数是一个关于期望 μ 的函数。

$$b''(\theta) = \nu(\mu) \quad (7.1.11)$$

方差函数 $\nu(\mu)$ 存在两种情况:

1. 方差函数是一个常量值, $\nu(\mu) = b''(\theta) = \text{constant}$, 此时分布的方差与均值无关。
2. 方差函数是一个关于均值 μ 的函数, $\nu(\mu) = b''(\theta) = f(\theta) = f(\mu)$, 此时分布的方差与均值有关

方差函数 (variance function)，是一个平滑函数，它把分布的均值参数 μ 和分布的方差关联在一起。如果其值一个常数值，说明均值和方差是独立无关的；反之，如果是 μ 的函数，说明均值和方差是相关联的。在高斯分布中， $b''(\theta) = 1$ ，所以方差和均值是相互独立的，对于其他分布，这是不成立的，高斯分布是特例。

影响方差的，除了方差函数 $\nu(\mu)$ 以外，还有分散参数 $a(\phi) = \phi$ ，它起到一个缩放的作用。参数 θ 和 ϕ 本质上是位置 (locate) 和尺度 (scale) 参数，位置参数反映数据的均值，尺度参数反映数据方差。

当 $a(\phi) = 1$ 时，分布的方差可以通过 $\nu(\mu)$ 计算得到，模型只有一个未知参数 μ (或者说是 θ ，因为 μ 和 θ 是可以转换的)，此时就是单参数指数族分布。当 $a(\phi) = \phi$ 时，分布就多了一个未知参数 ϕ ，此时就是双参数指数族分布。之后我们会看到，高斯分布是一个双参数分布，而最大似然估计是无法同时估计出参数 θ 和 ϕ 的，需要做一些改动才行，在以后的章节中我们会讨论这个问题。

在经典线性回归模型中，输入特征数据 x 通过线性组合 $\eta = \beta^T x$ 影响着响应变量 Y (高斯分布) 的均值 $\mu = \eta = \beta^T x$ ，所有的观测样本共用参数 β (对于任意 x ，都是同样的 β 值)，当 x 不同时，高斯变量 Y 拥有不同的均值 μ ，通过这种方式实现了条件概率分布 $p(Y|X)$ 的表达。但是对于高斯变量 Y 的方差参数 σ^2 并没有假设为未知参数，而是假设其为已知的值，并且对于任意的观测样本 x 都是一样的值。然而，在 GLM 的框架下，是可以允许不同观测样本有不同的方差，而这是通过 $a(\phi)$ 实现的。此时函数 $a(\phi)$ 通常被定义成如下的形式：

$$a(\phi) = \frac{\phi}{w_i} \quad (7.1.12)$$

通常对于所有的观测样本来说， ϕ 是一个相同的，而 w 可以根据不同的观测样本取不同的值，下标 i 表示样本编号。 w 被称为先验权重 (prior weight)，通常是指根据额外的先验信息确定的。如果所有观测样本具有相同的方差假设，那么 w 值通常就是 1；反之， w 可以是和样本相关的，不同的样本采用不同的值。

表 7.1.1: 常见分布的方差函数

分布	方差函数 $\nu(\mu)$	约束	导数 $\partial\nu(\mu)/\partial\mu$
Gaussian	1	$\begin{cases} \mu \in \mathcal{R} \\ y \in \mathcal{R} \end{cases}$	0
Bernoulli	$\mu(1 - \mu)$	$\begin{cases} 0 < \mu < 1 \\ 0 \leq y \leq 1 \end{cases}$	$1 - 2\mu$
Binomial(k)	$\mu(1 - \mu/k)$	$\begin{cases} 0 < \mu < k \\ 0 \leq y \leq k \end{cases}$	$1 - 2\mu/k$
Poisson	μ	$\begin{cases} \mu > 0 \\ y \geq 0 \end{cases}$	1
Gamma	μ^2	$\begin{cases} \mu > 0 \\ y > 0 \end{cases}$	2μ
Inverse Gaussian	μ^3	$\begin{cases} \mu > 0 \\ y > 0 \end{cases}$	$3\mu^2$
Negative binomial(α)	$\mu + \alpha\mu^3$	$\begin{cases} \mu > 0 \\ y \geq 0 \end{cases}$	$1 + 2\alpha\mu$
Power(k)	μ^k	$\begin{cases} \mu > 0 \\ k \neq 0, 1, 2 \end{cases}$	$k\mu^{k-1}$

7.1.2 示例：高斯分布

高斯分布的概率密度函数为：

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}\right\} \quad (7.1.13)$$

将其改写成指数分散族的形式：

$$f(y) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right\} \quad (7.1.14)$$

和公式 (7.1.3) 进行对比, 各个标准项为:

$$\begin{aligned}\theta &= \mu \\ b(\theta) &= \frac{1}{2}\mu^2 \\ a(\phi) &= \sigma^2\end{aligned}\tag{7.1.15}$$

高斯分布的均值和方差为:

$$\begin{aligned}\mathbb{E}[Y] &= b'(\theta) = \mu \\ \text{Var}(Y) &= b''(\theta)a(\phi) = \sigma^2\end{aligned}\tag{7.1.16}$$

对于高斯分布来说, 方差和均值是独立无关的。

7.1.3 示例: 伯努利分布

7.2 广义线性模型

在 GLM 框架中, 我们假设响应变量 Y 是服从指数族分布的, 我们的目的是通过输入变量 X 预测响应变量 Y 的值, 并且 GLM 是线性模型, 也就是通过输入变量 X 的线性组合预测 Y 。

线性预测器

既然是线性模型, 所以显然, 我们需要把输入变量 X 进行组成一个线性组合。

$$\eta = \beta^T x + b\tag{7.2.1}$$

x 可以是一个向量, η 是关于 x 的一个线性函数。为了简洁性, 通常会人为的为 x 加入一维常量值 1, 并且把截距参数 b 算在 β 中, 这样上述线性函数可以写成向量内积的形式。

$$\eta = \beta^T x\tag{7.2.2}$$

连接函数

回顾一下我们的初衷, 我们需要用输入变量 X 的线性结果 η 去预测输出变量 Y 的值, Y 是一个指数族的随机变量, 对于一个随机变量, 其值可以是其值域空间中任意的一个值, 只不过每个值的概率可能不同 (当然对于均匀分布其每个值的概率是相同的)。但是, 我们期望得到 Y 的一个具体的值, 显然随机变量 Y 的期望值是最好不过选择。

$$\mu = \mathbb{E}[Y]\tag{7.2.3}$$

现在, 我们需要通过 η 得到 μ , 然后把 μ 作为模型的输出值, 也就是模型预测出的 Y 的值。那要如何做到呢? 显然, 可以定义一个函数, 将两者联系起来。

$$\begin{aligned}\eta &= g(\mu) \\ \mu &= g^{-1}(\eta)\end{aligned}\tag{7.2.4}$$

通常把函数 g 称为连接函数 (link function), 连接函数 g 是用来连接线性预测器 η 和均值 μ 的。连接函数的反函数 g^{-1} 可以称为响应函数 (response function), 或者激活函数 (active function), 连接函数可以有很多种选择。在高斯线性模型 (传统线性回归模型) 中, 连接函数是恒等函数 $\eta = g(\mu) = \mu$ 。在泊松分布中, 均值 μ 必须是正的, 所以 $\eta = \mu$ 不再适用, 因为 $\eta = \beta^T x$ 的取值范围值整个实数域。对于泊松分布, 连接函数可以选择对数函数 $\eta = \log \mu$, 此时 $\mu = e^\eta$ 确保了 μ 为正数。连接函数本质上, 就是把实数域范围的 η 转换到特定分布合法的 μ 值空间上。

广义线性模型

显然, 一个广义线性模型有三个关键组件:

1. 一个线性预测器 $\eta = \beta^T x$, 被称为系统组件 (systematic component)。
2. 一个指数族分布作为响应变量 Y 概率分布 $p(Y; \theta)$, 被称为随机组件 (random component)。
3. 一个连接函数 g 使得 $\eta = g(\mu)$, μ 是 Y 的期望, 连接函数描述系统组件和随机组件之间的关系。

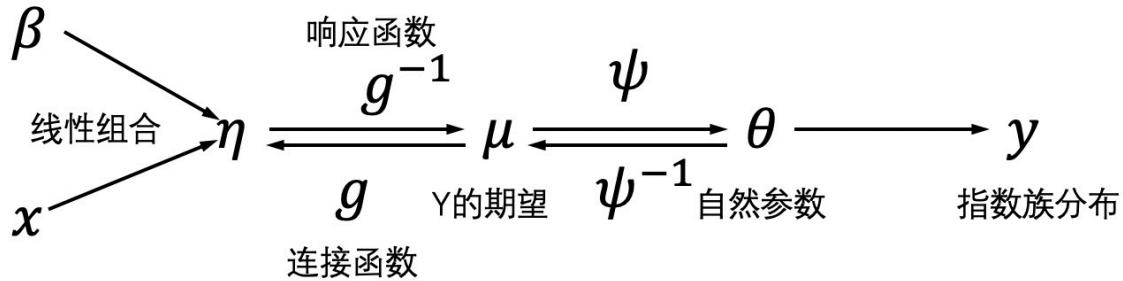


图 7.2.1: 广义线性模型变量之间的关系

广义线性模型是对经典线性模型的扩展, 将输出变量 Y 的条件概率分布扩展到指数族分布, 图 7.2.1 展示了广义线性模型框架下各个变量之间的关系。

- 输入变量 X 和系数 β 组成一个线性关系, $\eta = \beta^T x$, η 被称为线性预测器 (linear predictor), β 是定义的未知参数。
- 在广义线性模型的框架下, 响应变量 Y 被看做是随机变量, 并且其概率分布是指数族分布的一种, θ 是分布的自然参数。 θ 和 μ 存在一一映射关系, 我们用函数 ψ 表示这种关系。
- 通过一个连接函数 (link function) 将 η 和变量 Y 的期望 μ 关联起来, $\eta = g(\mu)$, 函数 g 称为连接函数。 g 的反函数就是激活函数 (active function), $\mu = g^{-1}(\eta)$, 有的资料中也称为响应函数 (response function)、均值函数 (mean function)。激活函数可以是线性的, 也可以是非线性的, 比如, 经典线性回归模型的激活函数为 $\mu = g^{-1}(\eta) = \eta$, 逻辑回归模型的激活函数为 $\mu = g^{-1}(\eta) = \text{sigmoid}(\eta)$ 。

线性预测器 η 和指数族分布的期望 μ 存在函数关系, $\mu = g^{-1}(\eta)$ 。指数族分布的自然参数 θ 又和期望 μ 存在着一一映射的关系, $\theta = \psi(\mu)$ 。因此, 指数族分布的自然参数 θ 一定是可以转换成一个关于 η 的函数, 响应变量 Y 的概率分布函数可以转化成和 η 相关。

$$\begin{aligned}
 p(y|\theta) &= \exp \left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \\
 &= \exp \left\{ \frac{\psi(\mu)y - b(\psi(\mu))}{a(\phi)} + c(y, \phi) \right\} \\
 &= \exp \left\{ \frac{\psi(g^{-1}(\eta))y - b(\psi(g^{-1}(\eta)))}{a(\phi)} + c(y, \phi) \right\} \\
 &= \exp \left\{ \frac{\psi(g^{-1}(\beta^T x))y - b(\psi(g^{-1}(\beta^T x)))}{a(\phi)} + c(y, \phi) \right\} \\
 &= P(Y|X; \beta)
 \end{aligned} \tag{7.2.5}$$

至此, 我们把输入变量 X 和响应变量 Y 的概率分布函数连接到了一起, 得到了条件概率分布 $P(Y|X)$ 的概率分布, 公式 (7.2.5) 就是广义线性模型的一般形式。

规范连接 (canonical link)

观察公式 (7.2.5) , 如果连接函数 g 和 ψ 相同, 那么 ψ 和 g^{-1} 就互为反函数, 二者可以抵消掉, 此时满足 $\theta = \eta$, 上式就可以简化成如下形式。

$$p(Y|X; \beta) = \exp \left\{ \frac{(\beta^T x)y - b(\beta^T x)}{a(\phi)} + c(y, \phi) \right\} \quad (7.2.6)$$

当连接函数使得 $\eta = \theta$ 时, 称为规范连接 (canonical link) 函数。实际上规范连接函数满足 $\eta = g(\mu) = \psi(\mu) = \theta$, 换句话说, 对于一个特定的指数族分布, 其规范连接函数为 $g = \psi$ 。使用规范连接函数可以带来很多统计属性, 最直接的就是可以简化参数估计的计算过程。

传统线性回归模型

传统的线性回归模型就是假设响应变量 Y 服从高斯分布, 高斯分布的概率密度函数用指数族的形式表示为:

$$f(y) = \exp \left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \quad (7.2.7)$$

和公式 (7.1.3) 进行对比, 各个标准项为:

$$\begin{aligned} \theta &= \mu \\ b(\theta) &= \frac{1}{2}\mu^2 \\ a(\phi) &= \sigma^2 \end{aligned} \quad (7.2.8)$$

可以看到, 高斯分布的自然参数 θ 和其期望 μ 之间的关系为恒等函数, 即 $\theta = \mu$ 。在传统线性回归模型中, 连接函数采用的也是恒等函数, 因此传统线性回归模型采用的是标准连接函数。此时满足 $\theta = \mu = \eta = \beta^T x$, 模型预测值 \hat{y} 为:

$$\hat{y} = \mu = \eta = \beta^T x \quad (7.2.9)$$

此外, 传统线性回归模型中假设所有样本的都具有相同的方差, 并且方差是常量 1 , $\sigma^2 = 1$ 。

然而, 当无法合理地假设数据是正态分布的或者响应变量的结果集有限集时, 传统的线性回归模型是不合适的。此外, 在许多情况下, 传统线性回归模型的同方差假设是站不住脚的, 此时传统线性回归模型也是不合适的。GLM 允许对传统线性回归模型进行扩展, 以突破这些限制。我们可以根据 y 的数据分布选择合适的指数族概率分布, 并且调整连接函数把实数域的 η 值映射到 y 的值域空间中。同时我们能够开发一种适用于所有 GLM 框架下模型的参数估计算法, 以应对不同情况下的参数估计算法。

表 7.2.1: 常见连接函数

名称	连接函数	激活函数 (反连接)	μ 的空间
Identity	$\eta = \mu$	$\mu = \eta$	$\mu \in \mathcal{R}$
Logit	$\eta = \ln\{\mu/(1 - \mu)\}$	$\mu = e^\eta/(1 + e^\eta)$	$\mu \in (0, 1)$
Log	$\eta = \ln(\mu)$	$\mu = e^\eta$	$\mu > 0$
Negative binomial(α)	$\eta = \ln\{\mu/(\mu + 1/\alpha)\}$	$\mu = e^\eta/\{\alpha(1 - e^\eta)\}$	$\mu > 0$
Log-complement	$\eta = \ln(1 - \mu)$	$\mu = 1 - e^\eta$	$\mu < 1$
Log-log	$\eta = -\ln\{-\ln(\mu)\}$	$\mu = \exp\{-\exp(-\eta)\}$	$\mu \in (0, 1)$
Complementary log-log	$\eta = \ln\{-\ln(1 - \mu)\}$	$\mu = 1 - \exp\{-\exp(\eta)\}$	$\mu \in (0, 1)$
Probit	$\eta = \Phi^{-1}(\mu)$	$\mu = \Phi(\eta)$	$\mu \in (0, 1)$
Reciprocal	$\eta = 1/\mu$	$\mu = 1/\eta$	$\mu \in \mathcal{R}$
Power($\alpha = -2$)	$\eta = 1/\mu^2$	$\mu = 1/\sqrt{\eta}$	$\mu > 0$
Power(α) $\left\{ \begin{array}{l} \alpha \neq 0 \\ \alpha = 0 \end{array} \right.$	$\eta = \begin{cases} \mu^\alpha \\ \ln(\mu) \end{cases}$	$\mu = \begin{cases} \eta^{1/\alpha} \\ \exp(\eta) \end{cases}$	$\mu \in \mathcal{R}$
Odds power(α) $\left\{ \begin{array}{l} \alpha \neq 0 \\ \alpha = 0 \end{array} \right.$	$\eta = \begin{cases} \frac{\mu/(1-\mu)^{\alpha-1}}{\alpha} \\ \ln\left(\frac{\mu}{1-\mu}\right) \end{cases}$	$\mu = \begin{cases} \frac{(1+\alpha\eta)^{1/\alpha}}{1+(1+\alpha\eta)^{1/\alpha}} \\ \frac{e^\eta}{1+e^\eta} \end{cases}$	$\mu \in (0, 1)$

表 7.2.2: 连接函数的导数

名称	连接函数	一阶导数 $\Delta = \partial\eta/\partial\mu$	二阶导数
Identity	$\eta = \mu$	1	0
Logit	$\eta = \ln\{\mu/(1 - \mu)\}$	$1/\{\mu(1 - \mu)\}$	$(2\mu - 1)\Delta^2$
Log	$\eta = \ln(\mu)$	$1/\mu$	$-\Delta^2$
Negative binomial(α)	$\eta = \ln\{\alpha\mu/(1 + \alpha\mu)\}$	$1/(\mu + \alpha\mu^2)$	$-\Delta^2(1 + 2\alpha\mu)$
Log-complement	$\eta = \ln(1 - \mu)$	$-1/(1 - \mu)$	$-\Delta^2$
Log-log	$\eta = -\ln\{-\ln(\mu)\}$	$-1/\{\mu \ln(\mu)\}$	$\{1 + \ln(\mu)\}\Delta^2$
Complementary log-log	$\eta = \ln\{-\ln(1 - \mu)\}$	$\{(\mu - 1) \ln(1 - \mu)\}^{-1}$	$-\{1 + \ln(1 - \mu)\}\Delta^2$
Probit	$\eta = \Phi^{-1}(\mu)$	$1/\phi\{\Phi^{-1}(\mu)\}$	$\eta\Delta^2$
Reciprocal	$\eta = 1/\mu$	$-1/\mu^2$	$-2\Delta/\mu$
Power($\alpha = -2$)	$\eta = 1/\mu^2$	$-2/\mu^3$	$-3\Delta/\mu$
Power(α) $\begin{cases} \alpha \neq 0 \\ \alpha = 0 \end{cases}$	$\eta = \begin{cases} \mu^\alpha \\ \ln(\mu) \end{cases}$	$\begin{cases} \alpha\mu^{\alpha-1} \\ 1/\mu \end{cases}$	$\begin{cases} (\alpha - 1)\Delta/\alpha \\ -\Delta^2 \end{cases}$
Odds power(α) $\begin{cases} \alpha \neq 0 \\ \alpha = 0 \end{cases}$	$\eta = \begin{cases} \frac{\mu/(1-\mu)^{\alpha-1}}{\alpha} \\ \ln\left(\frac{\mu}{1-\mu}\right) \end{cases}$	$\begin{cases} \frac{\mu^{\alpha-1}}{(1-\mu)^{\alpha+1}} \\ \frac{1}{\mu(1-\mu)} \end{cases}$	$\begin{cases} \Delta\left(\frac{1-1/\alpha}{1-\mu} + \alpha + 1\right) \\ \mu\Delta^2 \end{cases}$

表 7.2.3: 激活函数的导数

连接函数名称	激活函数 (反连接)	一阶导数 $\Delta = \partial\mu/\partial\eta$	二阶导数
Identity	$\mu = \eta$	1	0
Logit	$\mu = e^\eta/(1 + e^\eta)$	$\mu(1 - \mu)$	$\Delta(1 - 2\mu)$
Log	$\mu = e^\eta$	μ	Δ
Negative binomial(α)	$\mu = e^\eta/\{\alpha(1 - e^\eta)\}$	$\mu + \alpha\mu^2$	$\Delta(1 + 2\alpha\mu)$
Log-complement	$\mu = 1 - e^\eta$	$\mu - 1$	Δ
Log-log	$\mu = \exp\{-\exp(-\eta)\}$	$-\mu \ln(\mu)$	$\Delta\{1 + \ln(\mu)\}$
Complementary log-log	$\mu = 1 - \exp\{-\exp(\eta)\}$	$(\mu - 1) \ln(1 - \mu)$	$\Delta\{1 + \ln(1 - \mu)\}$
Probit	$\mu = \Phi(\eta)$	$\phi(\eta)$	$-\Delta\eta$
Reciprocal	$\mu = 1/\eta$	$-\mu^2$	$-2\Delta\mu$
Power($\alpha = -2$)	$\mu = 1/\sqrt{\eta}$	$-\mu^3/2$	$3\Delta^2/\mu$
Power(α) $\begin{cases} \alpha \neq 0 \\ \alpha = 0 \end{cases}$	$\mu = \begin{cases} \eta^{1/\alpha} \\ \exp(\eta) \end{cases}$	$\begin{cases} \frac{1}{\alpha}\mu^{1-\alpha} \\ \mu \end{cases}$	$\begin{cases} \Delta(1/\alpha - 1)/\mu^\alpha \\ \Delta \end{cases}$
Odds power(α) $\begin{cases} \alpha \neq 0 \\ \alpha = 0 \end{cases}$	$\mu = \begin{cases} \frac{(1+\alpha\eta)^{1/\alpha}}{1+(1+\alpha\eta)^{1/\alpha}} \\ \frac{e^\eta}{1+e^\eta} \end{cases}$	$\begin{cases} \frac{\mu(1-\mu)}{1+\alpha\eta} \\ \mu(1 - \mu) \end{cases}$	$\begin{cases} \Delta\left(1 - 2\mu - \frac{\alpha}{1+\alpha\eta}\right) \\ \Delta(1 - 2\mu) \end{cases}$

7.3 例子

表 7.3.1: 常见 GLM 表 (1)

	Normal(Gaussian) $N(\mu, \sigma^2)$	Bernoulli $B(\mu)$	Binomial $B(N, \mu)$
Range of y	real: $(-\infty, +\infty)$	$\{0, 1\}$	$\{0, \dots, N\}$
$f(y)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\mu^y (1-\mu)^{1-y}$	$\binom{N}{y} \mu^y (1-\mu)^{N-y}$
EDF	$\exp\left\{\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\ln 2\pi\sigma^2}{2}\right\}$	$\exp\left\{y \ln \frac{\mu}{1-\mu} + \ln(1-\mu)\right\}$	$\exp\left[\frac{y \ln(\frac{\mu}{1-\mu}) + \ln(1-\mu)}{1/N} + \ln(\binom{N}{y})\right]$
$\theta = \psi(\mu)$	$\theta = \mu$	$\theta = \ln\left(\frac{\mu}{1-\mu}\right) = \text{logit}(\mu)$	$\theta = \ln\left(\frac{\mu}{1-\mu}\right)$
$\mu = \psi^{-1}(\theta)$	$\mu = \theta$	$\mu = \frac{1}{1+e^{-\theta}} = \text{sigmoid}(\theta)$	$\mu = \frac{1}{1+e^{-\theta}}$
$b(\theta)$	$\frac{\theta^2}{2}$	$\ln(1 + e^\theta)$	$\ln(1 + e^\theta)$
$b(\mu)$	$\frac{\mu^2}{2}$	$-\ln(1 - \mu)$	$-\ln(1 - \mu)$
Link name	Identity	Logit	Logit
Link function	$\eta = \mu$	$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$
Mean function	$\mu = \eta$	$\mu = \frac{1}{1+e^{-\eta}}$	$\mu = \frac{N}{1+e^{-\eta}}$
$\nu(\mu) = b''(\theta)$	1	$\mu(1 - \mu)$	$\mu(1 - \mu)$
$a(\phi)$	σ^2	1	$\frac{1}{N}$

表 7.3.2: 常见 GLM 表 (2)

	Categorical $\text{Cat}(K, \mu)$	Poisson $\text{Poisson}(\mu)$
Range of y	$\{1, \dots, K\}$	integer 0, 1, 2, ...
$f(y)$	$\prod_k \mu_k^{y_k}$	$\exp\{y \ln \mu - \ln \mu\}$
EDF	$\exp\left\{\sum_{k=1}^{K-1} x_k \ln\left(\frac{\mu_k}{\mu_K}\right) + \ln\left(1 - \sum_{k=1}^{K-1} \mu_k\right)\right\}$	$\exp\{y \ln \mu - \ln \mu\}$
$\theta = \psi(\mu)$	$\theta_k = \ln\left(\frac{\mu_k}{\mu_K}\right)$	$\theta = \ln \mu$
$\mu = \psi^{-1}(\theta)$	$\mu_k = \frac{e^{\theta_k}}{\sum_{j=1}^K e^{\theta_j}}$	$\mu = e^\theta$
$b(\theta)$	$\ln\left(\sum_{k=1}^K e^{\theta_k}\right)$	e^θ
$b(\mu)$	$-\ln\left(1 - \sum_{k=1}^{K-1} \mu_k\right)$	$\ln \mu$
Link name	Logit	
Link function	$\eta_k = \ln\left(\frac{\mu_k}{\mu_K}\right)$	
Mean function	$\mu_k = \frac{e^{\eta_k}}{\sum_k e^{\eta_k}}$	
$\nu(\mu) = b''(\theta)$	$\mu_k(1 - \mu_k)$	mu
$a(\phi)$	1	1

参数估计

本节我们介绍两种 GLM 模型的参数估计算法，我们将统一的以指数族的形式展现算法过程，这样适用于指数族中的所有具体分布，我们的目标是让大家对 GLM 的基础理论有个全面的了解，同时我们会着重强调算法成立的假设及其一些限制。

传统上，对于单参数的指数族分布可以运用梯度下降法和牛顿法进行参数估计，梯度下降法的优点是算法实现简单，缺点是收敛速度不如牛顿法。梯度下降法和牛顿法在形式上是非常相似的，二者都是沿着目标函数的负梯度方向寻找最优解，不同的是传统梯度下降法利用一阶导数，而牛顿法利用二阶导数，牛顿法相对于梯度下降法收敛速度会更快，但是由于二阶导数的引入也使得牛顿法的计算复杂度增加很多，甚至很多时候无法计算。在用牛顿法对指数族模型进行参数估计时，不同的分布拥有不同的梯度表达式，所以每种分布都需实现一个适合自己的牛顿法。这里我们同时会介绍牛顿法的一个变种算法，迭代重加权最小平方法 (iteratively reweighted least squares, IRLS)。GLM 框架下的模型都可以以统计的形式运用 IRLS 算法进行参数估计，这是 GLM 非常有吸引力的一点。IRLS 算法的另一个特点是不需要对估计参数 β 进行初始化设置。

8.1 最大似然估计

最大似然估计是应用十分广泛的一种参数估计方法，其核心思想是通过极大化最大似然函数找到参数的最优解，GLM 中参数估计就是使用的最大似然方法。注意在 GLM 中，最大似然方法不能同时估计线性协变量参数 β 和分散参数 ϕ ，在 GLM 中的最大似然估计通常是假设分散参数 ϕ 已知的情况下，估计协变量参数 β 。

假设响应变量 Y 是 GLM 中的指数族分布，协变量 X 和响应变量 Y 的一个样本集为 $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，样本之间是相互独立的，本书中的最大似然估计都是建立在样本独立的假设之上。所有样本的联合概率为

$$f(y; \theta, \phi) = \prod_{i=1}^N f(y_i; \theta, \phi) \quad (8.1.1)$$

θ 是指数族分布的自然参数， ϕ 是分散参数。样本的联合概率又被称为样本集的似然函数，

$$L(\theta, \phi; y) = \prod_{i=1}^N f(\theta, \phi; y_i) \quad (8.1.2)$$

公式 (8.1.1) 和公式 (8.1.2) 的区别在于, 前者是一个概率密度 (质量) 函数, 是在给定 θ, ϕ 的条件下关于变量 Y 的函数; 后者是一个似然函数, 其表达是在给定观测样本 y 的条件下关于未知参数 θ, ϕ 的函数。

因为似然函数是一个连乘形式, 所以通常我们会对其进行一个对数转换 (log-transform) 进而得到一个连加的形式, 连加的形式更方便进行计算。连乘变成连加有两个好处, (1) 更容易求导和极大化操作; (2) 似然函数是概率连乘, 而概率都是小于 1 的, 大量小于 1 的数字连乘产生更小的数字, 甚至趋近于 0, 而计算机的浮点数精度是通常无法处理这么小的数字的, 所以加对数更方便计算机进行数值处理。

加了对数的似然函数被称为对数似然函数, 通常用符号 ℓ 表示, 对数似然函数是最大似然估计 (ML) 的核心, GLM 模型的对数似然估计函数为

$$\ell(\theta, \phi; y) = \sum_{i=1}^N \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad (8.1.3)$$

θ_i 是自然参数, $b(\theta_i)$ 是累积函数, 它描述的是分布的矩 (moment); ϕ 是分散参数 (dispersion parameter), 影响着分布的方差; $c(\cdot)$ 是归一化项。归一化项不是 θ 的函数, 而是简单地缩放基础密度函数的范围使得整个函数的积分 (或求和) 为 1。在上一节我们讨论过, 指数族分布的期望与方差可以通过 $b(\theta)$ 的导数求得。

$$\begin{aligned} \mu &= \mathbb{E}[Y] = b'(\theta) \\ V(Y) &= b''(\theta)a(\phi) \end{aligned} \quad (8.1.4)$$

并且我们知道, 均值参数 μ 和自然参数 θ 是存在一个可逆的函数关系的, 也就是说 μ 可以看做是关于 θ 的一个函数, 反之, θ 也可看做是一个关于 μ 的函数。基于这个事实, 我们可以把 $b''(\theta)$ 看做是一个关于 μ 的函数, 记作 $\nu(\mu)$ 。

$$b''(\theta) \triangleq \nu(\mu) \quad (8.1.5)$$

因此, 方差 $V(Y)$ 就可以被看成是函数 $\nu(\mu)$ 和分散函数 $a(\phi)$ 的乘积, 通常我们把 $\nu(\mu) = b''(\theta)$ 称为方差函数 (variance function), 注意: 虽然叫方差函数, 但方差函数的值不是方差本身。有时 $b''(\theta)$ 会是一个常数量 (constant), 比如高斯分布, 此时分布的方差为:

$$V(Y) = \text{constant} \times a(\phi) \quad (8.1.6)$$

这时, 分布的方差就不会受到均值的影响了。另外方差函数 $\nu(\mu)$ 可以通过简单方式求得。

$$\nu(\mu) = b''(\theta) = (b'(\theta))' = (\mu(\theta))' = \frac{\partial \mu}{\partial \theta} \quad (8.1.7)$$

显然, 当 μ 与 θ 之间的映射函数是线性函数时, 一阶偏导 $\frac{\partial \mu}{\partial \theta}$ 就是一个常数值。另外, 我们知道反函数的导数就等于原函数导数的倒数, 所以有:

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{\nu(\mu)} \quad (8.1.8)$$

在 GLM 框架下, 输入变量 X 和其系数 β 组成一个线性预测器 $\eta = \beta^T x$ 。 η 和分布的均值 (期望) 通过连接函数 (已知的) 连接在一起。

$$\begin{aligned} \eta &= \beta^T x = g(\mu) \\ \mu &= g^{-1}(\eta) \end{aligned} \quad (8.1.9)$$

其中 β 和 x 都是一个向量, $\beta^T x = \sum_j \beta_j x_j$, 因此有:

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad (8.1.10)$$

线性预测器 η 的值空间并没有特别的限定, 其值空间是整个实数域 $\eta \in R$ 。而 μ 的取值范围是特定分布相关的, 不同指数族分布, μ 的取值范围是不同的, 比如高斯分布 $\mu \in R$, 二项分布 $\mu \in [0, 1]$ 。因此, 连接函数的一个目的就是将线性预测器的值映射到响应变量期望参数的范围。

现在让我们回到最大似然估计, 最大似然估计的思想是使得似然函数取得最大值的参数值为模型的最优解。根据微分理论, 函数取得极值的点其一阶偏导数为 0, 然而导数为 0 的点不一定是最大值的点, 也可能是驻点、最小值点, 所以最大似然估计要求似然函数最好是凹函数。利用求导的链式法则对 GLM 模型的对数似然函数进行求导, 注意, 参数 β 是一个向量, 所以这里是偏导数,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^N \left\{ \frac{y_i - b'(\theta_i)}{a(\phi)} \right\} \left\{ \frac{1}{\nu(\mu_i)} \right\} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij}\end{aligned}\quad (8.1.11)$$

其中 i 是观测样本的编号, j 是参数向量的下标。 x_{ij} 表示第 i 条观测样本的第 j 列特征值, y_i 是响应变量的观测值, $a(\phi)$ 通常认为是已知的。 μ_i 是 y_i 的期望, 也是模型的预测值, 方差函数 $\nu(\mu_i)$ 是关于 μ_i 的函数, 因此也可以算出。 $\frac{\partial \mu}{\partial \eta}$ 是响应函数(或者说是激活函数)关于 η_i 的导数, 在确定了连接函数的形式后也是可以算出的。

公式 (8.1.11) 是 GLM 标准形式下对数似然函数的一阶偏导数, GLM 框架下的任意模型都可以按照此公式计算偏导数, 只需要按照特定的分布和连接函数替换相应组件即可。

对数似然函数的一阶导数又叫得分统计量(score statistic, Fisher score), 或者得分函数(score function), 常用符号 U 表示。

$$\begin{aligned}U_j &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)g(\mu_i)'} x_{ij}\end{aligned}\quad (8.1.12)$$

U 的表达式中只有 y_i 是随机变量的样本, 其它都是数值变量, U 是一个关于样本的函数, 所以它是一个统计量(statistic), 得分函数(score function)有时也叫作得分统计量(score statistic), 统计量也是随机变量。我们知道 $\mathbb{E}[y_i] = \mu_i$, 而统计量 U 是变量 y 的函数, 因此 U 期望值为:

$$\begin{aligned}\mathbb{E}_y[U_j] &= \mathbb{E}_y \left[\frac{\partial \ell}{\partial \beta_j} \right] \\ &= \mathbb{E}_y \left[\sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \right] \\ &= \sum_{i=1}^N \frac{\mathbb{E}[y_i] - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= 0\end{aligned}\quad (8.1.13)$$

统计量 U 的方差 $\mathcal{J} = \mathbb{E}\{(U - \mathbb{E}[U])(U - \mathbb{E}[U])^T\} = \mathbb{E}[UU^T]$ 又被称为费希尔信息(Fisher information), 或者

信息矩阵 (information matrix)。

$$\begin{aligned}
 \mathcal{J}_{jk} &= \mathbb{E}[U_j U_k] \\
 &= \mathbb{E}_y \left[\sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \cdot \sum_{l=1}^N \frac{y_l - \mu_l}{a(\phi)\nu(\mu_l)} \left(\frac{\partial \mu}{\partial \eta} \right)_l x_{lk} \right] \\
 &= \mathbb{E}_y \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{(y_i - \mu_i)^2}{[a(\phi)\nu(\mu_i)]^2} x_{ij} x_{ik} + \mathbb{E}_y \left[\sum_{i=1}^N \sum_{l=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \frac{y_l - \mu_l}{a(\phi)\nu(\mu_l)} \left(\frac{\partial \mu}{\partial \eta} \right)_l x_{lk} \right]_{l \neq i} \\
 &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{\mathbb{E}_y[(y_i - \mu_i)^2]}{[a(\phi)\nu(\mu_i)]^2} x_{ij} x_{ik} + \underbrace{\left[\sum_{i=1}^N \sum_{l=1}^N \frac{\mathbb{E}_y[(y_i - \mu_i)(y_l - \mu_l)]}{a(\phi)^2 \nu(\mu_i) \nu(\mu_l)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \left(\frac{\partial \mu}{\partial \eta} \right)_l x_{lk} \right]_{l \neq i}}_0
 \end{aligned} \tag{8.1.14}$$

$\mathbb{E}[(y_i - \mu_i)(y_l - \mu_l)]$ 是 y_i 与 y_l 的协方差, 根据样本独立性假设, 有 $y_i \perp\!\!\!\perp y_l (l \neq i)$ 成立, 因此 y_i 与 y_l 的协方差为 0, 即 $\mathbb{E}[(y_i - \mu_i)(y_l - \mu_l)] = 0$ 。而 $\mathbb{E}_y[(y_i - \mu_i)^2]$ 表示 y_i 的方差, 有 $\mathbb{E}_y[(y_i - \mu_i)^2] = V(y_i) = a(\phi)\nu(\mu_i)$ 。最终化简为

$$\mathcal{J}_{jk} = \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{1}{a(\phi)\nu(\mu_i)} x_{ij} x_{ik} \tag{8.1.15}$$

在最大似然估计的理论中, 通过令 $U = 0$ 求得参数估计值, 这个等式被称为估计等式 (estimating equation), 有的资料中也叫正规方程 (normal equation)。

$$U_j = \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} = 0 \tag{8.1.16}$$

在这个方程中, 有 $\mu_i = r(\eta_i) = r(x_i^T \beta)$, 函数 $r(\cdot)$ 是连接函数的反函数, 称为响应函数, 是已知的。协变量系数 β 是方程的未知量, 也是模型的未知参数, 是我们想要求解的。分散函数 $a(\phi)$ 通常被认为是已知的, 假设 $a(\phi) = \phi$, 并且 ϕ 与样本无关, 即所有样本具有相同的值。当 $U_j = 0$ 时, 有

$$\begin{aligned}
 U_j &= 0 \\
 \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} &= 0 \\
 \phi \sum_{i=1}^N \frac{y_i - \mu_i}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} &= 0 \\
 \sum_{i=1}^N \frac{y_i - \mu_i}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} &= 0
 \end{aligned} \tag{8.1.17}$$

显然, 在样本具有相同分散参数 ϕ 的假设之下, 协变量参数 β 的最大似然估计是不受 ϕ 影响的。

现在我们以传统线下回归模型为例, 演示下如何利用估计等式进行参数求解。传统线性回归模型也是 GLM 的一员, 相当于响应变量 y_i 是高斯变量, $y_i \sim \mathcal{N}(\mu_i, \sigma^2 = 1)$, 并且连接函数是恒等函数 $\eta_i = \mu_i$, 响应函数作为连接函数的反函数, 自然也是恒等函数, 即 $\mu_i = \eta_i$, 因此响应函数的导数是常量 1。

$$\frac{\partial \mu_i}{\partial \eta_i} = 1 \tag{8.1.18}$$

传统线性回归模型中, 方差是常量, $V(y_i) = \sigma^2 = 1$, 因此有

$$\begin{aligned}
 \nu(\mu) &= 1 \\
 a(\phi) &= \sigma^2 = 1
 \end{aligned} \tag{8.1.19}$$

各项代入到估计方程中, U_j 简化为

$$U_j = \sum_{i=1}^N (y_i - \mu_i)x_{ij} = 0 \quad (8.1.20)$$

上式是单个参数 β_j 的得分统计量 U_j , 转成向量为

$$\mathbf{U} = (\mathbf{y} - \mathbf{u})^T \mathbf{X} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{X} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{0} \quad (8.1.21)$$

移项可得参数的估计值为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (8.1.22)$$

我们发现标准连接函数的高斯模型, 估计等式 $\mathbf{U} = \mathbf{0}$ 可以得到解析解, 这是高斯模型独有的特性, 其它模型或者连接函数是不具备这个特性的。

在 GLM 中, 估计等式要想得到解析解, 需要满足两个条件:

1. 连接函数的是标准连接函数的。
2. 连接函数是线性函数。

根据标准连接函数的定义, 标准连接函数的使得 $\theta_i = \eta_i$, 此时有 $\partial\theta_i/\partial\mu_i = \partial\eta_i/\partial\mu_i$, 得分统计量 U 可以得到简化。

$$\begin{aligned} U_j &= \frac{\partial \ell}{\beta_j} = \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \underbrace{\left(\frac{\partial \eta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)}_{\text{抵消掉}} \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)} x_{ji} \end{aligned} \quad (8.1.23)$$

当采用标准连接函数时, 得分统计量中, 响应函数的导数和连接函数的的导数互相抵消掉, 得分统计量 U 得到简化。再根据上文所述, $a(\phi)$ 不影响参数估计结果, 可以去掉。

$$U_j = \sum_{i=1}^N (y_i - \mu_i)x_{ji} \quad (8.1.24)$$

转换成矩阵的形式为

$$\mathbf{U} = (\mathbf{y} - \mathbf{u})^T \mathbf{X} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{u} = 0 \quad (8.1.25)$$

移项可得

$$\mathbf{X}^T \mathbf{u} = \mathbf{X}^T \mathbf{y} \quad (8.1.26)$$

当 μ 与 η 是线性关系时, 比如 $\mu_i = \alpha\eta_i = \alpha(x_i^T \boldsymbol{\beta})$, 公式 (8.1.26) 才能求得解析解。

$$\mathbf{X}^T \mathbf{u} = \mathbf{X}^T \alpha(\mathbf{X} \boldsymbol{\beta}) = \mathbf{X}^T \mathbf{y} \quad (8.1.27)$$

$$\hat{\boldsymbol{\beta}} = (\alpha \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (8.1.28)$$

在 GLM 中, 能同时满足这两个条件的, 只有高斯模型, 其它的模型都不符合第二点。对于无法取得解析解的模型, 可以用数值法求解。最常用的数值法有梯度下降法和牛顿法, 梯度下降法仅利用似然函数的一阶导数, 而牛顿法同时利用似然函数的一阶导数和二阶导数, 下节我们介绍最大似然估计的数值求解法。

8.2 泰勒级数

最大似然的求解需要求解正规方程, 然而在 GLM 中, 正规方程并不是一定存在解析解的, 需要满足一些限制条件才行, 解析解的方式并不具备通用性, 我们需要采用更一般的方法, 逼近法, 也叫迭代法、数值法。迭代法又可以简单分为一阶导(梯度下降法系列)和二阶导(牛顿法系列), 实际这两种都可以通过泰勒级数(Taylor series)进行推导。泰勒级数有很多个名字, 泰勒公式(Taylor formula)、泰勒级数(Taylor series)、泰勒展开(Taylor explanation)、泰勒定理(Taylor theory)等, 都是一回事。

设 n 是一个正整数。如果定义在一个包含 x_0 的区间上的函数 f , 在 x_0 处 $n+1$ 次可导, 那么对于这个区间上的任意 x , 都有

$$\begin{aligned} f(x)_{Taylor} &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n \\ &= f(x_0) + \frac{f'(x_0)}{1!} (x - x_0) + \frac{f^{(2)}(x_0)}{2!} (x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n + R_i(x) \end{aligned} \quad (8.2.1)$$

其中 $f^{(n)}$ 表示 f 的 n 阶导数, 泰勒展开表达的就是 $f(x)$ 可以用其附近的点 $f(x_0)$ 近似的表示。

提示: 注意, 本书讨论的迭代求解算法默认目标函数都是凸函数, 也就是函数有唯一的极值点。关于非凸函数以及带约束的优化问题, 请读者参考其它资料。

8.3 梯度下降法

我们把对数似然函数按照泰勒公式进行展开, 但是我们只展开到一阶导数, 把更高阶导数的和看做一个常数量 constant。

$$f(x)_{Taylor} = f(x_0) + f'(x_0)(x - x_0) + \text{constant} \quad (8.3.1)$$

现在我们把对数似然函数按照上式进行展开:

$$\ell(\beta^{(t+1)}) = \ell(\beta^t) + \ell'(\beta^t)(\beta^{(t+1)} - \beta^t) + \text{constant} \quad (8.3.2)$$

假设 $\beta^{(t+1)}$ 是对数似然函数的极值点, 也就是参数的最优解, β^t 是其附近的一个点。现在把这个式子进行简单的移项和变换,

$$\ell(\beta^{(t+1)}) - \ell(\beta^t) = \ell'(\beta^t)(\beta^{(t+1)} - \beta^t) + \text{constant} \quad (8.3.3)$$

显然 $\ell(\beta^{(t+1)})$ 应该是大于等于 $\ell(\beta^t)$ 的, 因此有

$$\ell(\beta^{(t+1)}) - \ell(\beta^t) = \ell'(\beta^t)(\beta^{(t+1)} - \beta^t) + \text{constant} \geq 0 \quad (8.3.4)$$

对上述公式进行移项处理, 可得:

$$\beta^{(t+1)} \geq \beta^t - \frac{\text{constant}}{\ell'(\beta^t)} \quad (8.3.5)$$

我们给参数 β 设置一个初始值, 然后通过上式不停的迭代计算新的 β , t 表示迭代计算的轮次, 直到等号成立的时候, 就找到了参数的最优解。

通常我们把一阶导 $\ell'(\beta^t)$ 称为梯度(gradient), 公式 (8.3.5) 说明只要 $\beta^{(t+1)}$ 沿着 β^t 的负梯度方向进行移动, 我们终将能达到极值点。注意 $\frac{\text{constant}}{\ell'(\beta^t)}$ 的绝对值的大小影响着前进的速度, 其方向(正负号)决定目标函数是

否向着极大值点移动。所以和下面的公式是等价的, α 称为学习率 (learning rate), 是一个人工设置参数, 控制的迭代的速度。

$$\beta^{(t+1)} = \beta^t - \alpha \ell'(\beta^t) \quad (8.3.6)$$

利用公式 (8.3.6) 进行参数迭代求解的方法就称为梯度上升法, 梯度上升法的核心就是让参数变量沿着负梯度的方向前进。虽然理论上最终一定能到达极值点, 但是实际上会受到学习率参数 α 的影响, 学习率可以理解成每次迭代前进的步长 (step size), 步长越大前进的越快, 收敛性速度就越快; 反之, 步长越小, 收敛越慢。但是步长如果大了, 就会造成震荡现象, 即一步迭代就越过了终点 (极值点), 并且在极值点附近往返震荡, 永远无法收敛。为了保证算法能一定收敛, 通常会为 α 设定一个较小的值。关于 α 的更多讨论请参考其它资料。

待处理: 图反了, 重新换个图。

画图参考:

https://zh.d2l.ai/chapter_optimization/gd-sgd.html

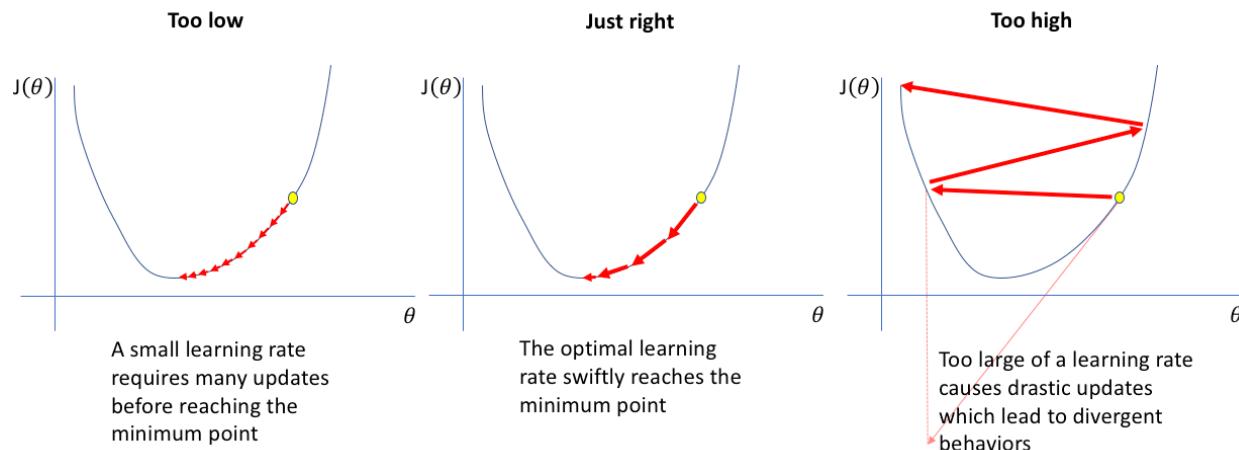


图 8.3.1: 梯度下降法中学习率的影响 (图片来自网络)

8.4 牛顿法

梯度下降法虽然也能收敛到最优解, 但是如果学习率设置 (通常人工设置) 不合理, 可能会造成收敛速度太慢或者无法收敛的问题, 其收敛速度难以有效的控制。现在我们讨论另一中迭代算法, 牛顿-拉夫森方法 (Newton-Raphson), 一般简称牛顿法。

8.4.1 算法推导

还是从泰勒展开公式开始, 让我们考虑二阶泰勒展开:

$$\ell(\beta^{(t+1)}) = \ell(\beta^t) + \ell'(\beta^t)(\beta^{(t+1)} - \beta^t) + \frac{1}{2}\ell''(\beta^t)(\beta^{(t+1)} - \beta^t)^2 + \text{constant} \quad (8.4.1)$$

我们知道目标函数在极值点处的导数应该为 0, 所以如果 $\beta^{(t+1)}$ 是极值点, 那么有 $\ell'(\beta^{(t+1)}) = 0$ 。我们对公式 (8.4.1) 进行求导, 注意 $\beta^{(t+1)}$ 才是函数未知量, β_t 和 $\ell(\beta^t)$ 都是已知量。

$$\ell'(\beta^{(t+1)}) = \ell'(\beta^t) + \ell''(\beta^t)(\beta^{(t+1)} - \beta^t) = 0 \quad (8.4.2)$$

通过移项可得:

$$\beta^{(t+1)} = \beta^t - \frac{\ell'(\beta^t)}{\ell''(\beta^t)} \quad (8.4.3)$$

这个迭代等式中, 需要同时使用到对数似然函数的一阶导和二阶导数, 二阶偏导数可以在一阶导数的基础上再次求导得到, 上一节已经讲过, 对数似然函数的一阶导数又称为得分统计量。

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \quad (8.4.4)$$

我们对 U_j 继续求导就是对数似然函数的二阶导数。

$$\begin{aligned} & \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right) \\ &= \frac{\partial U_j}{\partial \beta_k} \\ &= \sum_{i=1}^N \frac{1}{a(\phi)} \left(\frac{\partial}{\partial \beta_k} \right) \left\{ \frac{y_i - \mu_i}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{jn} \right\} \\ &= \sum_{i=1}^N \frac{1}{a(\phi)} \left[\left(\frac{\partial \mu}{\partial \eta} \right)_i \left\{ \left(\frac{\partial}{\partial \mu} \right)_i \left(\frac{\partial \mu}{\partial \eta} \right)_i \left(\frac{\partial \eta}{\partial \beta_k} \right)_i \right\} \frac{y_i - \mu_i}{\nu(\mu_i)} + \frac{y_i - \mu_i}{\nu(\mu_i)} \left\{ \left(\frac{\partial}{\partial \eta} \right)_i \left(\frac{\partial \eta}{\partial \beta_k} \right)_i \right\} \left(\frac{\partial \mu}{\partial \eta} \right)_i \right] x_{jn} \\ &= - \sum_{i=1}^N \frac{1}{a(\phi)} \left[\frac{1}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 - (\mu_i - y_i) \left\{ \frac{1}{\nu(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{\partial \nu(\mu_i)}{\partial \mu} - \frac{1}{\nu(\mu_i)} \left(\frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right\} \right] x_{jn} x_{kn} \end{aligned} \quad (8.4.5)$$

对数似然函数的二阶偏导数是一个矩阵, 这个矩阵又叫海森矩阵 (Hessian matrix), 常用符号 H 表示。牛顿法的迭代公式可以写成如下形式,

$$\beta^{(t+1)} = \beta^{(t)} - H(\beta^{(t)})^{-1} U(\beta^{(t)}) \quad (8.4.6)$$

和梯度下降法的公式 (8.3.6) 对比下发现, 两者非常相似, 不同的是牛顿法用 Hessian 矩阵的逆矩阵 $H(\beta^{(t)})^{-1}$ 替代了学习率参数, 避免了需要人工设置学习率的问题。相比梯度下降法, 牛顿法收敛速度更快, 并且也没有震荡无法收敛的问题。

观察下公式 (8.4.5), GLM 的海森矩阵计算难度是比较大的, 为了解决这个问题, 有时候会用海森的矩阵的期望 $\mathbb{E}[H]$ 替代。从公式 (8.4.5) 可以看到, 海森矩阵是一个关于样本的函数, 所以可以对海森矩阵求关于 y 的

期望。

$$\begin{aligned}
 \mathbb{E}_y[H]_{jk} &= \mathbb{E}_y \left[-\sum_{i=1}^N \frac{1}{a(\phi)} \left[\frac{1}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 - (\mu_i - y_i) \left\{ \frac{1}{\nu(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{\partial \nu(\mu_i)}{\partial \mu} - \frac{1}{\nu(\mu_i)} \left(\frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right\} \right] x_{ij} x_{ik} \right] \\
 &= -\sum_{i=1}^N \frac{1}{a(\phi)} \left[\frac{1}{\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 - (\mu_i - \mathbb{E}[y_i]) \left\{ \frac{1}{\nu(\mu_i)^2} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{\partial \nu(\mu_i)}{\partial \mu} - \frac{1}{\nu(\mu_i)} \left(\frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right\} \right] x_{ij} x_{ik} \\
 &= -\sum_{i=1}^N \frac{x_{ij} x_{ik}}{a(\phi) \nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2
 \end{aligned} \tag{8.4.7}$$

在参数的迭代过程中使用 $\mathbb{E}[H]$ 和使用 H 在参数收敛效果上没有太大区别，二者是类似的，但是 $\mathbb{E}[H]$ 的计算要简化了很多。原始海森矩阵 H 的计算依赖观测样本 y_i ，所以通常会把原始海森矩阵称为观测海森矩阵 (observed Hessian matrix, OHM)，他的期望矩阵称为期望海森 (expected Hessian matrix, EHM)。

$$\beta^{(t+1)} = \beta^{(t)} - \mathbb{E}[H(\beta^{(t)})]^{-1} U(\beta^t) \tag{8.4.8}$$

对比下信息矩阵公式 (8.1.15) 和期望海森公式 (8.4.7)，二者只差一个负号，是相反数的关系，这和我们在 [节 4.4.2](#) 讨论的结论是一致的。

$$\mathcal{J} = -\mathbb{E}[H] \tag{8.4.9}$$

可以看到在 GLM 中，信息矩阵 \mathcal{J} 可以通过对数似然函数的海森矩阵 H 得到。通常把负的 观测海森矩阵， $-H$ ，称为观测信息矩阵 (observed information matrix, OIM)，把负的 期望海森矩阵， $-\mathbb{E}[H]$ ，称为期望信息矩阵 (expected information matrix, EIM)。牛顿法的迭代过程可以用 EIM 代替 OIM 以简化计算过程。

$$\beta^{(t+1)} = \beta^{(t)} + \mathcal{J}(\beta^{(t)})^{-1} U(\beta^t) \tag{8.4.10}$$

我们这里描述的 Newton-Raphson 算法不支持分散参数 ϕ 的估计，通常在进行协变量参数 β 的最大似然估计时，认为 ϕ 是已知量。

在 [节 4.4.2](#) 讨论过，参数的最大似然估计量是一个统计量，并且其渐进服从正态分布，其方差可以通过信息矩阵 \mathcal{J} 计算得到。最终，Newton-Raphson 提供了如下功能：

1. 为所有 GLM 成员模型提供一个参数估计算法。
2. 附带产出参数估计量的标准误 (standard errors)，可通过信息矩阵得到。

8.4.2 标准连接函数

前文讲过，但模型采用标准连接函数时，得到统计量可以简化。现在我们看下标准连接函数对牛顿法的影响。当模型采用标准连接函数时，观测信息矩阵 (OIM) 会退化成期望信息矩阵 (EIM)，此时在牛顿算法中，两种矩阵是等价的。

根据标准连接函数的定义，当采用标准连接函数时自然参数 θ 就等于线性预测器 η ，即 $\theta = \eta$ 此时 U 可以简

化为:

$$\begin{aligned}
 U_j &= \frac{\partial \ell}{\beta_j} = \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\
 &= \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\
 &= \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\
 &= \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)} x_{ij}
 \end{aligned} \tag{8.4.11}$$

观测海森矩阵是对数似然函数的二阶导数, 也是 U 的一阶导数, 因此有

$$\begin{aligned}
 H_{jk} = U'_j &= \frac{\partial U_j}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_k} \\
 &= \sum_{i=1}^N \frac{-x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_{ik} \\
 &= - \sum_{i=1}^N \frac{x_{ij} x_{ik}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \\
 &= - \sum_{i=1}^N \frac{x_{ij} x_{ik}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \\
 &= - \sum_{i=1}^N \frac{x_{ij} x_{ik}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \\
 &= - \sum_{i=1}^N \frac{x_{ij} x_{ik}}{a(\phi) \nu(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\
 &= \mathbb{E}_y [H_{jk}]
 \end{aligned} \tag{8.4.12}$$

这和公式 (8.4.7) 是一样的。

重要: 在 GLM 中, 采用标准连接函数时, 观测海森矩阵和期望海森矩阵是相同的, 也就是观测信息矩阵 OIM 和期望信息矩阵 EIM 是相同的。

8.4.3 迭代初始值的设定

要实现 Newton-Raphson 迭代法, 我们必须对参数初始值有一个猜测。但目前没有用于获得良好参数初值的全局机制, 有一个相对合理的解决方案是, 利用线性预测器中的“常数项系数”获得初始值。这里的“常数项”指的是线性预测器中截距部分

$$\eta = \beta_0 \times 1 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{8.4.13}$$

其中 β_0 就是常数项系数。如果模型包含常数项, 则通常的做法是找到仅包含常数项系数的模型的估计值。我们令:

$$\eta = \beta_0 \tag{8.4.14}$$

然后令对数似然函数的一阶导数公式 (8.1.11) 为 0, 找到 β_0 的解析解。

$$\sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i = 0 \quad (8.4.15)$$

通过上式是可以得到 β_0 的一个估计值的。比如, 如果是逻辑回归模型, 则有

$$\begin{aligned} a(\phi) &= 1 \\ \nu(\mu) &= \mu(1 - \mu) \\ \mu &= \text{sigmoid}(\eta_i) = \text{sigmoid}(\beta_0) \\ \frac{\partial \mu}{\partial \eta} &= \frac{\partial}{\partial \eta} \text{sigmoid}(\eta) = \mu(1 - \mu) \end{aligned} \quad (8.4.16)$$

代入到公式 (8.4.15) 可得:

$$\begin{aligned} \sum_{i=1}^N \frac{(y_i - \mu_i)}{\mu_i(1 - \mu_i)} \mu_i(1 - \mu_i) &= 0 \\ \Downarrow \\ \sum_{i=1}^N (y_i - \mu_i) &= 0 \\ \Downarrow \\ \sum_{i=1}^N \left(y_i - \frac{1}{1 + e^{-\beta_0}} \right) &= 0 \\ \Downarrow \\ \underbrace{\frac{1}{N} \sum_{i=1}^N y_i}_{\text{均值} \bar{y}} &= \frac{1}{1 + e^{-\beta_0}} \\ \Downarrow \text{sigmoid 反函数求解} \\ \hat{\beta}_0 &= \ln \left(\frac{\bar{y}}{1 - \bar{y}} \right) \end{aligned} \quad (8.4.17)$$

然后我们就用 $\beta = (\hat{\beta}_0, 0, 0, \dots, 0)^T$ 作为 Newton-Raphson 算法首次迭代时参数向量的初始值。如果模型中没有常量项系数, 或者我们无法通过解析法求解纯常数项系数模型, 则必须使用更复杂的方法, 比如使用搜索方法寻找合理的初始点来开始 Newton-Raphson 算法。

8.5 迭代重加权最小二乘 (IRLS)

使用牛顿法对 GLM 中的模型进行参数估计时, 需要把每个模型的对数似然函数通过 β 进行参数化, 然后求出对数似然函数的偏导数, 并且在迭代开始前需要给 β 一个初始值, 这种方法过于繁琐, 本节我们介绍牛顿法在 GLM 中的一个变种算法, 迭代重加权最小二乘 (iteratively reweighted least square, IRLS) 算法, IRLS 算法是“GLM”的一个通用型参数估计算法, 可用于任意的指数族分布和连接函数, 并且不需要对 β 进行初始化。

8.5.1 算法推导

采用期望海森矩阵的牛顿法的参数迭代等式为

$$\beta^{(t+1)} = \beta^{(t)} + [\mathcal{J}^{(t)}]^{-1} U^{(t)} \quad (8.5.1)$$

等式两边同时乘以信息矩阵 \mathcal{J} ,

$$\mathcal{J}^{(t)} \beta^{(t+1)} = \mathcal{J}^{(t)} \beta^{(t)} + U^{(t)} \quad (8.5.2)$$

假设协变量参数 β 的数量是 p , 则信息矩阵 \mathcal{J} 是一个 $p \times p$ 的方阵, 其中每个元素 \mathcal{J}_{jk} 为

$$\mathcal{J}_{jk} = \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{x_{ij} x_{ik}}{a(\phi) \nu(\mu_i)} \quad (8.5.3)$$

仔细观察 \mathcal{J}_{jk} 的计算公式, 假设有一个 $N \times N$ 的对角矩阵, 每个对角元素为

$$W_{ii} = \frac{1}{a(\phi) \nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \quad (8.5.4)$$

方阵 \mathcal{J} 就相当于三个矩阵的乘法

$$\mathcal{J} = X^T W X \quad (8.5.5)$$

这个等式我们先记录下, 之后再使用。现在看下 $\mathcal{J}\beta$ 的结果是什么。

参数 β 是一个 $p \times 1$ 的列向量, 下标 j 表示行坐标, 下标 k 表示列坐标。方阵 \mathcal{J} 和列向量 β 相乘的计算过程是方阵 \mathcal{J} 的每个行向量 \mathcal{J}_j 和列向量 β 进行~~内~~积运算, 行向量 \mathcal{J}_j 和列向量 β 的~~内~~积结果为

$$\begin{aligned} \mathcal{J}_j \beta &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{x_{ij} x_i \beta}{a(\phi) \nu(\mu_i)} \\ &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{x_{ij} \eta_i}{a(\phi) \nu(\mu_i)} \end{aligned} \quad (8.5.6)$$

公式 (8.5.2) 的右侧就是两个 $p \times 1$ 的列向量相加, 每个元素 j 的计算过程是

$$\begin{aligned} \mathcal{J}_j^{(t)} \beta^{(t)} + U_j^{(t)} &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{x_{ij} \eta_i}{a(\phi) \nu(\mu_i)} + \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi) \nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{x_{ij}}{a(\phi) \nu(\mu_i)} \left\{ (y_i - \mu_i) \left(\frac{\partial \eta}{\partial \mu} \right)_i + \eta_i^{(t)} \right\} \end{aligned} \quad (8.5.7)$$

我们令

$$Z_i = \left\{ (y_i - \mu_i) \left(\frac{\partial \eta}{\partial \mu} \right)_i + \eta_i^{(t)} \right\} \quad (8.5.8)$$

Z 是一个 $N \times 1$ 的向量

$$Z = \left\{ (y - \mu) \left(\frac{\partial \eta}{\partial \mu} \right) + \eta^{(t)} \right\} \quad (8.5.9)$$

公式 (8.5.2) 的右侧等价于

$$\mathcal{J}^{(t)} \beta^{(t)} + U^{(t)} = X^T W^{(t)} Z^{(t)} \quad (8.5.10)$$

最终公式 (8.5.2) 等价于

$$(X^T W^{(t)} X) \beta^{(t+1)} = X^T W^{(t)} Z^{(t)} \quad (8.5.11)$$

通过移项可以得到参数 β 的迭代公式

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)} = \mathcal{J}^{-1} X^T W^{(t)} Z^{(t)} \quad (8.5.12)$$

其中

$$\begin{aligned} W^{(t)} &= \text{diag} \left\{ \frac{1}{a(\phi)\nu(\mu)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \right\}_{(N \times N)} \\ &= \text{diag} \left\{ \frac{1}{V(\mu)(g')^2} \right\}_{(N \times N)} \quad \text{对角矩阵} \\ Z^{(t)} &= \left\{ (y - \mu) \left(\frac{\partial \eta}{\partial \mu} \right) + \eta^{(t)} \right\}_{(N \times 1)} \\ &= \left\{ (y - \mu)g' + \eta^{(t)} \right\}_{(N \times 1)} \end{aligned} \quad (8.5.13)$$

$a(\phi)$ 是分散函数, $\nu(\mu)$ 是方差函数, $\frac{\partial \mu}{\partial \eta}$ 是响应函数 r 的导数, 等价于连接函数 g 的导数的倒数。 $\frac{\partial \eta}{\partial \mu}$ 是连接函数 g 对 μ 的导数。 W 和 Z 的计算都依赖 η , 而计算 η 又需要 β , 所以需要迭代的方式更新 β 。

公式 (8.5.12) 就是参数向量的更新公式, 它在形式上等价于加权的最小二乘法, 其中 W 相当于权重矩阵, 并且每一次迭代都要重新计算 W , 所以我们把这个算法称为迭代重加权最小二乘法 (Iteratively reweighted least square, IRLS), “reweighted” 指的就是每次迭代重新计算权重矩阵, Z 被称为工作响应 (working response)。

8.5.2 算法过程

收敛性判断

在迭代的过程中, 我们可以检查参数 β 的相对变化来决定是否结束算法。

$$\sqrt{\frac{(\beta^{new} - \beta^{old})^T (\beta^{new} - \beta^{old})}{\beta^{oldT} \beta^{new}}} < \epsilon \quad (8.5.14)$$

也可以通过相对偏差 (deviance) 来判断。

$$\left| \frac{D(y - \mu^{new}) - D(y, \mu^{old})}{D(y, \mu^{old})} \right| < \epsilon \quad (8.5.15)$$

关于偏差的概念我们将在下一章详细介绍。

迭代初始值的设定

对比下 Newton-Raphson 算法的参数迭代公式 (公式 (8.4.6)) 和 IRLS 算法的参数迭代公式 (公式 (8.5.12)), 可以发现 IRLS 算法并不需要直接在 $\beta^{(t)}$ 的基础上进行参数迭代, IRLS 算法的参数迭代仅仅依赖 μ 和 η , 因此与 Newton-Raphson 算法不同的是, IRLS 不需要对参数向量 β 进行初始值的猜测, 只需要给 μ 和 η 赋予一个初始值即可。

- 对于二项式分布, 可以令 $\mu_i^{(0)} = k_i(y_i + 0.5)/(k_i + 1)$, $\eta_i^{(0)} = g(\mu_i^{(0)})$ 。
- 对于非二项式分布, 可以令 $\mu_i^{(0)} = y_i$, $\eta_i^{(0)} = g(\mu_i^{(0)})$ 。

IRLS 算法在更新时, 只依赖期望 μ 和线性预测器 η , 鉴于这一特性, 可以使用期望 μ 对 GLM 模型的概率函数进行参数化, 而不需要细化到 β , 这可以极大的降低 GLM 模型概率函数的复杂性。

示例代码

```

import numpy as np
from typing import Optional, Union
import abc

class Link(abc.ABC):

    def link(self, mu: np.ndarray) -> Union[np.ndarray, float]:
        """
        连接函数

        :param mu:
        :return:
        """
        pass

    def gradient(self, mu: np.ndarray) -> Union[np.ndarray, float]:
        """
        连接函数的导数

        :param mu:
        :return:
        """
        pass

    def response(self, eta: np.ndarray) -> Union[np.ndarray, float]:
        """
        响应函数, 连接函数的反函数

        :param eta:
        :return:
        """
        pass

class Distribution(abc.ABC):
    name = "default"

    def __init__(self, phi=1):
        self._phi = phi

    def variance(self, mu: np.ndarray) -> Union[np.ndarray, float]:
        """
        方差函数

        :param mu:
        :return:
        """
        pass

    def phi(self):
        """
        分散函数  $a(\phi)$ 

        :return:
        """
        return self._phi

```

(下页继续)

(续上页)

```

class GLM:

    def __init__(self, p, link: Link, family: Distribution):
        """
        :param p: beta 参数数量
        :param link: 连接函数
        :param family: 响应变量的分布
        :return:
        """

        self.beta = np.zeros(shape=(p, 1))
        self.link = link
        self.family = family

    def convergence(self, old_beta, cur_beta):
        pass

    def init_mu(self, y):
        return (y + y.mean()) / 2 # 均值参数初始化

    def fit(self, X: np.ndarray, y: np.ndarray):
        """
        IRLS 算法

        :param X:
        :param y:
        :return:
        """

        mu = self.init_mu(y) # 均值参数初始化
        eta = self.link.link(mu) # 线性预测器初始化
        beta = self.beta.copy()
        # 直到收敛
        while True:
            v = self.family.variance(mu) # 方差函数
            g_gradient = self.link.gradient(mu) # 连接函数的导数
            W = 1 / (self.family.phi() * v * g_gradient * g_gradient) # 计算权重矩阵
            Z = eta + (y - mu) * g_gradient
            old_beta = beta
            beta = np.linalg.inv(X.T * W * X) * X.T * W * Z # 更新参数向量
            eta = X * beta # 计算新的线性预测器
            mu = self.link.response(eta) # 计算模型预测值/期望参数
            if self.convergence(old_beta, beta):
                break

        self.beta = beta

```

在 GLM 框架提出之前, 各个模型已经被提出并广泛应用于, 比如线性回归模型、逻辑回归模型、泊松回归模型等等, 这些模型都是先于 GLM 框架的。在 GLM 提出前, 这些模型都是利用最大似然进行参数估计的, 并且一般都是利用牛顿法进行求解的。IRLS 算法是伴随着 GLM 诞生的, 要明白的是 IRLS 本身也是建立在最大似然的基础上的。在 IRLS 提出前, GLM 中的每个模型都需要单独的运用牛顿法求解, IRLS 是对牛顿法在 GLM 中的一种统一化的抽象。牛顿法中, 每种模型需要单独计算对数似然函数对协变量参数 β 的一、二阶导数, 比较复杂。而 IRLS 建立在 GLM 统一的表达式上, 不需要为每个模型单独求参数的导数, 只需要替换相应的连接函数、连接函数导数、方差即可。

重要: IRLS 仍然属于最大似然估计, 它是牛顿法在 GLM 中的一种简化。为了区分, 很多资料把采用牛顿法的最大似然估计称为“完全最大似然估计 (full maximum likelihood estimation)”。完全最大似然估计法需要针对每种不同模型单独求解对数似然函数的导数, 而 IRLS 不需要。

8.6 估计量的标准误差

在 节 4.4.3 讨论最大似然估计时, 讲过参数的最大似然估计量是一个统计量, 而统计量是一个随机量, 统计量的概率分布称为抽样分布 (sampling distribution)。期望参数的似然估计量 $\hat{\mu}$ 的抽样分布是高斯分布。

$$\hat{\mu}_{ML} \sim \mathcal{N}(\mu_{true}, \mathcal{J}^{-1}) \quad (8.6.1)$$

期望参数的似然估计量 $\hat{\mu}$ 渐近服从高斯分布, 抽样分布的期望值就是参数真实值 $\mathbb{E}[\hat{\mu}_{ML}] = \mu_{true}$, 其协方差矩阵是信息矩阵的逆 $Var(\hat{\mu}_{ML}) = \mathcal{J}^{-1}$ 。这就意味着均值参数似然估计值的标准误差为

$$SE(\hat{\mu}_{ML}) = \sqrt{\mathcal{J}^{-1}} \quad (8.6.2)$$

在 GLM 中, 期望参数 μ_i 和线性预测器 $\eta_i = \beta^T x_i$ 通过连接函数连接到一起, 协变量参数 β 取代了期望参数 μ_i 。协变量参数 β 的最大似然估计量的抽样分布同样是高斯分布, 这里我们省略证明过程, 详细的证明过程可参考 节 10.1.2。

$$\hat{\beta}_{ML} \sim \mathcal{N}(\beta_{true}, \mathcal{J}^{-1}) \quad (8.6.3)$$

在 IRLS 算法的迭代过程中已经计算出了 $\mathcal{J}(\beta)^{-1} = -\mathbb{E}[H(\beta)]^{-1} = (X^T W X)^{-1}$, 所以使用 IRLS 算法可以很方便的得到估计量的标准误差。

$$SE(\hat{\beta}) = \sqrt{\mathcal{J}^{-1}} = \sqrt{\text{diag}[(X^T W X)^{-1}]} \quad (8.6.4)$$

8.7 分散参数的估计

我们已经知道, 在所有样本拥有相同分散参数 ϕ 的假设之下, 协变量参数 β 的最大似然估计不会受到 ϕ 的影响。但这并不意味着分散参数 ϕ 就没有价值了, 首先样本具有相同 ϕ 的假设未必总是成立的, 其次协变量参数 β 的最大似然估计量的标准误差的计算是依赖 ϕ 的。似然估计量 $\hat{\beta}_{ML}$ 的标准误差是通过信息矩阵 \mathcal{J} 计算得到的, 而 \mathcal{J} 的计算依赖 ϕ 。

在之前的讨论中, 我们都是假设 ϕ 是已知量, 通常可以根据人工经验值指定。然而人工经验不总是靠谱的, 很多时候我们需要从实际数据中去探索 ϕ 的合理值。但是 IRLS 算法并没有提供对 ϕ 的估计, 我们需要用一些其它的方法去估计。

最容易想到的方法, 就是在得到 β 的最大似然估计值之后, 再次利用最大似然估计对 ϕ 进行估计。要对 ϕ 进行最大似然估计, 就需要对 GLM 的对数似然函数求 ϕ 的导数。

$$\ell(\theta, \phi; y) = \sum_{i=1}^N \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad (8.7.1)$$

在对数似然函数中, $c(y_i, \phi)$ 也是包含 ϕ 的, 在 GLM 的不同分布中, 它形式是不尽相同的, 每种分布模型需要单独去针对 ϕ 求偏导, 这种方法比较繁琐, 这里我们暂且不表, 本节我们介绍一种简单且常用的方法。

在 GLM 中, 估计 ϕ 的最常用方法是利用皮尔逊卡方统计量, 皮尔逊卡方统计量的计算公式为

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{a(\phi) \nu(\hat{\mu}_i)} \quad (8.7.2)$$

皮尔逊卡方统计量, 顾名思义, 它也是一个统计量, 并且它的期望值是 $\mathbb{E}[\chi^2] = N - p$, N 是样本的数量, p 是协变量参数的数量, 也就是 β 的长度。这里我们假设 $\chi^2 = N - p$, $a(\phi) = \phi$, 则有

$$\phi = \frac{\chi^2}{N - p} = \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)(N - p)} \quad (8.7.3)$$

有关皮尔逊卡方统计量的细节在后续章节中会继续讨论, 这里可以先记住就可以了。需要注意的是, 利用皮尔逊卡方统计量估计 ϕ 的方法, 同样是建立在所有样本拥有相同 ϕ 的假设之上。

模型评估

当我们训练好一个模型后，我们需要知道这个模型的“好坏”。要评价一个模型的好坏，就需要找到合理的度量方法，模型好坏的度量方法有很多很多，但并不存在一个完美的度量方法能够适用于所有的场景、数据和模型。通常我们需要依据场景、数据、模型来选择合适度量方法。虽然度量方法很多，但我们可以根据度量目标的不同来对这些度量方法进行简单的归类和划分。

- 参数估计量的方差。
- 拟合值和观测值之间的差异。
- 每个协变量

9.1 拟合优度

拟合优度 (goodness of fit, GOF)，表示的模型输出的拟合值 (fitted value) 和实际观测值之间的差异程度，目前存在多种差异度量指标，比如大家熟知的平方误差 (损失)、似然值等都是 GOF 的度量指标。通常参数估计的过程就是极值化某种拟合优度的指标的过程，在 GLM 中一般是采用最大 (对数) 似然法估计模型参数。

一个模型拟合数据的过程，可以看做是用模型的输出拟合值 (fitted value) $\hat{\mu}$ 去替换数据的观测值 (observed) y ，这个模型拥有较少的参数。通常模型的拟合值 $\hat{\mu}$ 并不会和观测值 y 完全相等，接下来的问题就是两者的差异有多大。较小的差值说明模型的拟合效果好，反之，较大的差值说明模型拟合效果差。模型对数据拟合效果的评估通常称为拟合优度 (goodness of fit, GOF)，GOF 度量观察值 y 与该模型拟合值 $\hat{\mu}$ 之间的差异。

9.1.1 嵌套模型

嵌套模型 (nested model) 两个统计模型 (statistical model), 如果对其中的一个模型的参数施加约束就能得到一个模型, 则这两个模型是嵌套的 (nested)。

假如我们用相同的数据拟合两个 GLM, Model 1, Model 2。其中, 当限制 Model 2 中部分参数为零之后会变成 Model 1 时, 我们说 Model 1 是 Model 2 的嵌套模型。

例 1: 嵌套模型示例 I

模型 1 的线性预测器 (linear predictor) 方程为:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (9.1.1)$$

模型 2 和模型 1 相比, 响应变量使用相同的指数族分布, 并且使用相同的链接函数 (link function) 和尺度参数 (scale parameter, ϕ), 但是其线性预测器的方程为:

$$\eta = \beta_0 + \beta_1 x_1 \quad (9.1.2)$$

此时, 我们就说模型 2 是模型的 1 的嵌套模型 (nested model), 因为通过对模型 1 的参数施加约束 $\beta_2 = \beta_3 = \beta_4 = 0$ 就得到了模型 2。

例 2: 嵌套模型示例 II

模型 1 的线性预测器 (linear predictor) 方程为:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \quad (9.1.3)$$

模型 2 和模型 1 相比, 响应变量使用相同的指数族分布, 并且使用相同的链接函数 (link function) 和尺度参数 (scale parameter, ϕ), 但是其线性预测器的方程为:

$$\eta = \beta_0 + \beta_1 x_1 \quad (9.1.4)$$

此时, 仍然认为模型 2 是模型的 1 的嵌套模型 (nested model), 因为通过对模型 1 的参数施加约束 $\beta_2 = 0$, 模型 1 就变成了模型 2。

我们知道, 模型的参数越多对数据的拟合程度就越好, 极端情况下, 模型参数的数量和样本的数量相同, 这时就相当于对每条样本都有一个独立的参数 (模型) 去拟合它, 理论上可以完美拟合所有的样本。我们把这样的模型成为之饱和模型 (saturated model), 也可以称为完整模型 (full model) 或者最大模型 (maximal model)。饱和模型虽然能完美拟合数据集, 但它并没有从数据集中学习出任何的统计信息 (统计规律), 所不具备泛化能力, 俗称过拟合 (over-fitted)。通过为饱和模型中的参数添加约束, 比如令一些参数值为 0, 相当于去掉了一个参数, 这样就得到了简化的模型。简化模型对数据集拟合度下降了, 但是其泛化能力会得到提升, 更少的参数数量可以得到更大的泛化能力。但是参数数量变少, 会降低拟合程度, 参数数量越少拟合度就越差, 所以也不是参数越少越好。

我们把完美拟合数据的模型称之为饱和模型 (saturated model), 饱和模型为每一条样本定义一个参数, 有多少条样本就有多少个参数, 这样就能完美拟合所有的样本。同样的道理, 我们可以定义一个“最差”的模型, 参数越多的模型拟合度越好, 参数越少拟合度越差, 我们定义只有一个截距参数的模型为“最差”的模型, 通常称为零模型 (null model)。零模型的线性预测器只有截距 (β_0) 部分, 而没有预测变量。

$$\eta = \beta_0 \quad (9.1.5)$$

图 9.1.1 展示了饱和模型 (saturated model)、拟合的逻辑回归模型 (logistic regression)、空模型 (null model) 三种

模型拟合效果的对比情况。饱和模型可以拟合所有的点，而空模型对所有样本只能输出一个固定值。

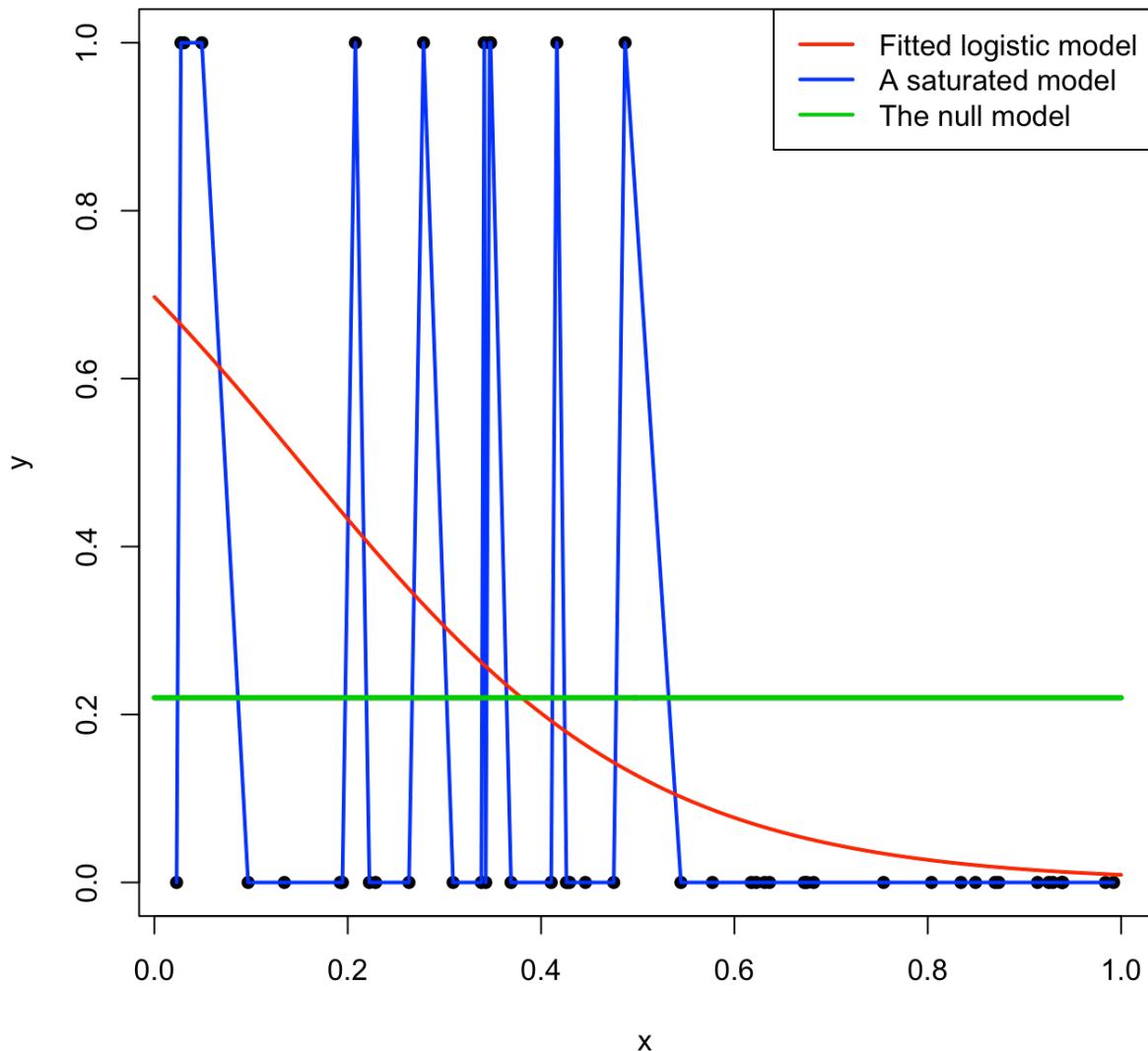


图 9.1.1: 饱和模型 (saturated model)、拟合的逻辑回归模型 (logistic regression)、空模型 (null model)

9.1.2 对数似然比 (Likelihood ratio)

我们知道似然 (Likelihood) 其实就是样本的联合概率, 似然值越大说明模型对样本的拟合程度越好, 因此我们可以通过对比两个模型的似然值来比较两个模型的好坏。我们把参数少的模型称为简单模型, 用符号 s 表示, L_s 表示模型 s 的似然; 另一个参数较多的模型称为复杂模型, 用符号 g 表示, L_g 表示模型 g 的似然。

在统计学中, 可以通过比较两个嵌套模型的似然值来评判哪个模型对数据拟合的更好。似然比 (likelihood-ratio,LR) 就是用来对比两个嵌套模型对于同一份数据集的拟合程度, LR 的计算公式如下:

$$LR = -2 \left(\frac{L_s}{L_g} \right) \quad (9.1.6)$$

其中 L_g 为复杂模型似然值, L_s 为简单模型似然值。从公式可以看出, 似然比就是两个模型的似然值比值。通常并不直接使用上述似然值的比值, 而是会加上一个对数, 变成对数似然比。

$$LLR = -2 \ln \left(\frac{L_s}{L_g} \right) = 2 \ln \frac{L_g}{L_s} = 2(\ln L_g - \ln L_s) \quad (9.1.7)$$

加上对数操作后, 就变成了两个模型对数似然值的差值, 使得计算更加方便。似然 (likelihood), 实际上也可以翻译为可能性, 表示的是样本发生的概率, 显然似然值越大的模型对数据的拟合也就越好。似然比就是直接比较两个模型的似然值大小。但是并不是任意两个模型都可以应用似然比去比较, 只有在特定条件下似然比才有意义。

1. 两个模型采用同一份数据集, 样本的数量和特征都是相同的。这很好理解, 不同数据集似然值自然是不同的, 没有比较的意义。
2. 两个模型是嵌套关系 (nested)。

对于两个嵌套模型, 其差别就是参数数据量不同。在 GLM 中, 就是协变量参数向量 β 的长度不同, 简单模型的参数向量 β_s 是复杂模型参数向量 β_g 的子集, 把 β_g 中部分元素设置为 0 就得到了 β_s 。我们知道, 在拟合效果一样的前提下, 参数数量越少的模型越“好”, 我们更期望于得到一个参数少的模型, 也就是尽量得到一个简单模型。然而理论上, 参数越多的模型对数据拟合就越好, 复杂模型的对数似然值一定是大于等于简单模型的, 因此一定有 $LLR \geq 0$ 。理论上当 $LLR = 0$ 时, 说明两个模型的拟合效果完全一样。然而实际上几乎是不可能实现的, 对数似然比的值基本上都是一个大于零的值。

LLR 的值什么范围的值意味着两个模型拟合程度接近呢? 这就需要找到一个判断的方法和标准。事实上, 对数似然比也是一个统计量, 称为似然比统计量 (Likelihood-ratio statistic), 并且其渐进服从卡方分布, 其自由度 (期望) 等于两个嵌套模型的参数数量之差。既然是一个统计量, 就表示 LLR 的值是一个随机值, 因此直接使用 LLR 值进行模型好坏的判断是不可靠的。可以通过假设检验的方法, 利用似然比统计量对两个模型进行对比检验, 这种检验方法称为似然比检验 (likelihood-ratio test,LRT) 有时也被称为似然比卡方检验 (likelihood-ratio chi-squared test,LRCT)。在统计学中, 似然比检验, 是用来比较两个嵌套模型的拟合优度 (goodness of fit,GOF) 的方法。似然比检验是基于最大似然估计的统计模型中应用广泛的一种模型对比方法, 有关假设检验的相关内容下一章在详细讨论。

9.1.3 偏差 (deviance)

上一节我们介绍了似然比统计量, 似然比检验是常用的一种嵌套模型比较的方法。似然比检验是对比两个模型的, 不是用来衡量单个模型的。本节我们介绍似然比统计量的一个衍生量-偏差统计量 (deviance,statistic), 偏差统计量本质上就是似然比统计量, 但它可以用来度量单个模型的拟合效果。

在开发一个模型时, 我们希望模型的预测值 \hat{y} 尽可能的接近数据的真实值 y , 对于一个规模为 N 的观测值样本, 我们可以考虑参数数量在 $[1, N]$ 之间的候选模型。最简单的模型是只有一个参数的模型, 但它对所有的样本的预测值都是一样的, 缺乏拟合能力, 只有一个参数的模型称为空模型 (null model)。最复杂的模型是含有 N 个参数的模型, 它可以完美拟合所有样本, 但是它缺乏泛化能力, 这样的模型称为饱和模型 (saturated model)。在实际应用中, 空模型过于简单, 而饱和模型又缺乏数据的抽象进而没有泛化能力。虽然饱和模型不能直接拿来用, 但是其却可以作为模型拟合能力评价指标的基准。

我们把训练出的模型模型称为拟合模型 (fitted model), 用符号 L_t 表示它的似然值, 同理用符号 L_f 表示对应饱和模型的似然值。则二者之间的对数似然比统计量为:

$$D = 2(\ln L_f - \ln L_t) \quad (9.1.8)$$

在 GLM 中, 我们把饱和模型和拟合模型之间的似然比统计量定义为 **偏差 (deviance)** 统计量, 常用符号 D 表示。严格来说, 偏差统计量就是似然比统计量的一个特例, 其比较的是饱和模型和训练出的模型之间的拟合度。饱和模型是完美拟合数据的模型, 其对数似然值是理论最大值, 代表了模型拟合度的最高值, 可以作为训练模型拟合度量的一个“参考线”, 训练模型的对数似然值“越接近”饱和模型的对数似然值说明训练模型拟合度越好。

模型的预测值 \hat{y}_i 就是分布 $p(y_i|x_i)$ 的期望值 $\mathbb{E}[p(y_i|x_i)] = \hat{\mu}_i$, 即 $\hat{y}_i = \hat{\mu}_i$ 。所以这里我们用 $\hat{\mu}_i$ 表示模型的预测值。现在回顾一下 GLM 中的指数族概率分布函数的形式公式 (9.1.9),

$$p(y_i|\theta_i) = \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \quad (9.1.9)$$

其中自然参数 θ_i 可以表示成期望 μ_i 的函数, 所以对于拟合模型自然参数 θ_i 可以写成 $\theta_i(\hat{\mu}_i)$, 拟合模型的对数似然函数可以写成:

$$\ln L_t = \sum_{i=1}^N \frac{y_i \theta(\hat{\mu}_i) - b(\theta(\hat{\mu}_i))}{a(\phi)} + \sum_{i=1}^N c(y_i; \phi) \quad (9.1.10)$$

至此, 我们把拟合模型的似然函数表示成了关于 $\hat{\mu}$ 的函数。同理, 对于饱和模型 (saturated model), 模型是完美拟合数据的, 所以其预测值是精确等于样本的观测值的, 即 $\hat{y}_i = y_i$, 换句话说, 对于饱和模型, 满足 $\hat{y}_i = \hat{\mu}_i = y_i$ 。因此, 饱和模型的对数似然函数为:

$$\ln L_f = \sum_{i=1}^N \frac{y_i \theta(y_i) - b(\theta(y_i))}{a(\phi)} + \sum_{i=1}^N c(y_i; \phi) \quad (9.1.11)$$

注意在 GLM 中, 分散参数 ϕ 与模型的期望 μ 无关, 模型关注的参数是协变量参数 β , 而分散参数 ϕ 是冗余参数, 并且与样本无关。因此分散参数没有下标 i , 并且无论是饱和模型还是拟合模型, 分散参数是相同的值。

现在把公式 (9.1.10) 和公式 (9.1.11) 代入到偏差统计量公式 (9.1.8), 两个对数似然函数中的项 $\sum_{i=1}^N c(y_i; \phi)$ 是相等的, 可以抵消掉。

$$D = \frac{2}{a(\phi)} \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \quad (9.1.12)$$

在 GLM 的应用中, 多数情况下 $a(\phi) = \phi$, 但有时会假设 $a(\phi) = \phi/w_i$ 此时 w_i 表示样本权重值, 意味着每条观测样本都可以有不同的权重值, 权重值 w_i 是已知的。偏差统计量 D 就变成:

$$D = \frac{2w_i}{\phi} \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \quad (9.1.13)$$

权重 w_i 并不是必要的, 只有在实际使用场景中需要为每条观测样本设置不同权重时才需要, 并且其值是事先已知的。因此在很多有关偏差得资料中并没有提及, 在本书后续的讨论中, 若无特别说明, 默认也省略掉权重, 并且假设 $a(\phi) = \phi$, 则偏差统计量的计算公式为

$$D = \frac{2}{\phi} \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \quad (9.1.14)$$

偏差统计量的计算依赖分散参数 ϕ , 在 GLM 的众多分布中, 离散分布都是不存在分散参数的, 相当于 $\phi = 1$; 连续值分布虽然存在分散参数 ϕ , 但在使用过程中一般都假设分散参数 ϕ 是已知的常量, 比如高斯模型通

常假设 $\phi = 1$ 。鉴于此, 早期很多资料默认把分散参数 ϕ 从偏差统计量的计算中去掉了(相当于 $\phi = 1$), 变成了

$$D = 2 \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \quad (9.1.15)$$

这造成了很多混淆, 为了区分二者, 把带有分散参数的称为尺度化偏差 (scaled deviance), 用符号 D^* 表示; 不带分散参数的称为偏差, 用符号 D 表示。当 $\phi = 1$ 时, 二者是等价的, 它们的关系是

$$D = \phi D^* \quad \text{或者} \quad D^* = \frac{D}{\phi} \quad (9.1.16)$$

为了表述清晰统一, 本书默认使用完整的带有分散参数的标准公式, 并统一使用名称”偏差 (deviance)” 和符号 D 表示, 即偏差统计量 (deviance statistic), D , 定义为

$$\begin{aligned} D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\ &= \frac{2}{a(\phi)} \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \\ &= \frac{2}{\phi} \sum_{i=1}^N [y_i \{\theta(y_i) - \theta(\hat{\mu}_i)\} - b\{\theta(y_i)\} + b\{\theta(\hat{\mu}_i)\}] \end{aligned} \quad (9.1.17)$$

偏差统计量是定义在整个观测样本上的, 单条样本的偏差 (deviance) 通常称为 unit deviance, 习惯上用符号 $d_i(y_i, \hat{\mu}_i)$ 表示, 整个观测样本的偏差就是所有个体 unit deviance 的求和, $D = \sum_{i=1}^N d_i(y_i, \hat{\mu}_i)$ 。

偏差统计量就是对数似然比统计量的一个特例, 比较的是拟合模型 (我们训练出来的模型) 和饱和模型的拟合度, 饱和模型的对数似然值是可以直接计算出的, 并且是当前观测样本集的似然值上限, 因此偏差统计量可以用度量模型的拟合优度。当然由于偏差统计量是对数似然比的一个特例, 其也继承了对数似然比统计量的特性, 比如偏差统计量的渐近分布也是卡方分布。同样偏差值本身也不能直接用来判断模型的好坏, 需要借助假设检验的手段才行, 在下一章会详细讨论。

最小偏差与最大似然

偏差统计量是饱和模型的对数似然值和拟合模型对数似然值差值的 2 倍, 在确定性观测样本集合模型下, 饱和模型的对数似然值是一个常量,

$$D = 2 \underbrace{[\ell(y; y) - \ell(\hat{\mu}; y)]}_{\text{常量}} \quad (9.1.18)$$

在进行参数估计时, 最小化偏差就相当于最大化拟合模型的对数似然, 因此 **参数的最大化似然估计和最小化偏差估计是等价的**。

偏差和最小二乘法的关系

对于标准连接的高斯分布模型, 也就是传统的线性回归模型, 有 $\theta = \eta = \mu, b(\theta) = \mu^2/2, a(\phi) = \sigma^2$, 因此其偏差为:

$$\begin{aligned} D &= 2 \sum_{i=1}^N [y_i \{y_i - \hat{\mu}_i\} - y_i^2/2 + \hat{\mu}_i^2/2] \\ &= 2 \sum_{i=1}^N [y_i^2/2 - y_i \hat{\mu}_i + \hat{\mu}_i^2/2] \\ &= \sum_{i=1}^N (y_i - \hat{\mu}_i)^2 \end{aligned} \quad (9.1.19)$$

可以看到, 标准连接高斯分布的偏差 (deviance) 和平方和损失是一致的, 实际上 **偏差 (deviance) 可以看做是传统线性回归模型最小二乘 (平方损失) 在 GLM 中的扩展**。

表 9.1.1: 常见 GLM 模型的偏差 (采用标准连接函数, 并且 $\phi = 1$)

分布	偏差 (deviance)
Gaussian(Normal)	$\sum_{i=1}^N (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_{i=1}^N \{y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$
Binomial	$2 \sum_{i=1}^N \{y_i \ln(y_i/\hat{\mu}_i) + (m - y_i) \ln[(m - y_i)/(m - \hat{\mu}_i)]\}$
Gamma	$2 \sum_{i=1}^N \{-\ln(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\}$
Inverse Gaussian	$\sum_{i=1}^N \{(y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)\}$

偏差统计量的抽样分布

偏差 (deviance) 是对数似然比统计量的一个特例, 因此也是一个统计量, 并且其渐近分布也是卡方分布。偏差统计量 D 是渐近服从卡方分布 χ^2 的。

$$D \sim \chi^2(N - p) \quad (9.1.20)$$

符号 $\chi^2(N - p)$ 的意思是自由度为 $N - p$ 的卡方分布, 卡方分布的自由度就是其期望参数, 自由度为 $N - p$, 也就意味着期望是 $N - p$ 。其中 N 是饱和模型的参数数量, 同时也是观测样本的数量。 p 是拟合模型的参数数量 (包含截距参数),

注意, 所谓渐近分布是指随着样本数量增加, 变量逐渐服从某个概率分布。理论当 N 无限大时, 变量才精确服从这个概率分布。实际上经验来看, 当样本数量在几百个以上时误差已经基本可以忽略了, 当然具体地还要看实际情况如何。在 GLM 中有一个特殊的情况是, 标准连接的高斯模型其偏差是精确服从卡方分布的, 而不是渐近的。

两个嵌套的模型的偏差统计量的差值, 就相当于这两个嵌套模型的对数似然比统计量。当两个嵌套模型的偏差统计相减时, 其中饱和模型的项就会抵消掉, 最后就等于两个模型的对数似然值做差值, 也就变成了对数似然比统计量。偏差统计量是 GLM 中最常用的统计量, 有关如何利用偏差统计量进行模型检验的方法下章详细介绍。

9.1.4 决定系数 R^2

另一种流行的模型拟合度的度量方法是 R^2 。通常在线性回归 (linear regression) 入门中讨论此统计量, 并给出了各种各样的解释。有很多方法可以解释这一统计数据, 并且这些解释已被推广到线性回归以外的领域。下面我们将在线性回归模型中 R^2 的几种定义方法, y_i 表示样本的观测值, \hat{y}_i 表示模型对样本的预测值, \bar{y} 表示样本的均值。

R^2 统计量

假设只有 Y 的观测数据, 还没有任何预测变量 X , 此时要预测 Y 的值, 可以使用样本的均值 (期望) \bar{y} 作为模型的预测值。显然此时, 模型只会输出一个值 $\hat{y}_i = \bar{y}$, 这样的模型是最基本的模型, 就相当于只有一个截距参数的模型 $\eta = \beta_0$, 不包含任何的预测 (特征) 变量 X , 这样的模型就是前文提到的空模型 (null model)。实际上, 空模型学习到的就是样本的均值 $\beta_0 = Y$ (这里暂时忽略连接函数的影响), 其最终就把样本均值作为模型的预测值输出。理论上, 空模型就是一个最简单的基础版模型, 它不包含任何预测变量 X 。

空模型预测值 \bar{y} 和观测值 y_i 之间的误差平方和为

$$TSS = \sum_{i=1}^N (y_i - \bar{y}_i)^2 \quad (9.1.20)$$

可以看到, 这个等式其实就是 **样本数据的方差**, 我们把它称为全部平方和 (total sum of squares, TSS), 表示数据的方差总和, 这也是数据的全部方差。

如果我们在空模型的基础上增加一个预测变量 X_1 (假设只有一维特征), 模型的线性预测器就变成了 $\beta_0 + \beta_1 x_1$, 显然如果 X_1 和变量 Y 存在一定的线性关系, 新模型相比空模型对数据的拟合能力一定会有所提升的, 用符号 \hat{y} 表示新模型的预测值。新模型和观测值的残差平方和为

$$RSS = \sum_{i=1}^N (y_i - \bar{y}_i)^2 \quad (9.1.21)$$

我们把线性模型 (包含预测变量) 的预测值 \hat{y} 和样本观测值之间的误差平方和称为残差平方和 (residual sum of squares, RSS)。

理论上, 我们为模型增加了一个预测变量 X_1 , 这会提高模型对数据的拟合能力, 也就是会减少模型预测值和观测值的误差, 即 $RSS \leq TSS$ 。显然可以通过比较 RSS 与 TSS 的大小关系来衡量增加了一个预测变量后, 模型对数据的拟合能力是否提升了。

定义如下统计量为 R^2 统计量, 英语读作 “R squared”, 中文读作 “R 方”。

$$R^2 = 1 - \frac{RSS}{TSS} \quad (9.1.22)$$

R^2 是一种独立模型拟合优度的统计量, TSS 是观测样本的数据的全部方差, RSS 是模型预测值和观测值之间的残差, 模型对数据拟合越好, RSS 的值就越小, 极限情况如果模型能完美拟合所有数据, RSS 值就是 0, 此时 R^2 的值就是 1。 R^2 越接近 1 表示模型对数据拟合的越好。反之, 如果模型对数据的拟合和空模型一样, 此时 $RSS = TSS$, 这时 $R^2 = 0$ 。 $R^2 = 0$ 就表示模型对数据的拟合能力和空模型是一样的。

有一种特殊的情况是 $R^2 < 0$, $R^2 < 0$ 说明 $RSS > TSS$, 这表示你的模型比空模型还要差, 还不如直接用样本的均值 \bar{y} 作为预测值。一般造成这种情况的原因可能是用错了模型, 也就是模型假设和观测数据严重不符, 还有可能是代码有 bug 或者数据有异常等等, 总之, 一旦遇到 $R^2 < 0$ 就需要好好排查了。 R^2 取值范围是 $(-\infty, 1]$, 越大说明模型拟合度越好, 反之越差。

R^2 又称为决定系数 (coefficient of determination), 它可以用来判断在空模型的基础上增加一个预测 (特征) 变量后, 模型对数据的拟合程度是否变得更好, 可以依此判断这个特征变量对观测变量 Y 是否具有线性解释的能力。

方差解释

R^2 的另一种解释是方差, TSS 是观测样本的全部方差, RSS 是增加了预测 (特征) 变量之后还剩下的方差, R^2 就表示预测 (特征) 变量对观测变量全部方差解释的比例。比如, 假设计算出 $R^2 = 0.7$, 这可以看做是预测 (特征) 变量 X 可以解释了观测变量 Y 的 70% 的方差, 剩下的 30% 是当前的 X 没有解释的。

用 $V(y)$ 表示观测样本的方差, $V(\hat{y})$ 表示模型的方差, $V(\hat{\epsilon}) = V(y) - V(\hat{y})$, R^2 也可以定义成如下的形式。

$$R^2 = \frac{V(\hat{y})}{V(y)} = \frac{V(\hat{y})}{V(\hat{y}) + V(\hat{\epsilon})} \quad (9.1.23)$$

$V(\hat{y})$ 是拟合模型的方差, $V(y)$ 是观测样本的方差,

偏差解释

我们知道偏差 (deviance) 统计量是误差平方和的扩展, 因此可以计算一个偏差版本的 R^2 。

用符号 L_0 表示空模型 (null model) 的似然, 并且把空模型的偏差值定义为空偏差 (null deviance), 用符号 D_0 表示。有些翻译把 “null deviance” 翻译成 “零偏差”, 个人觉得这非常容易产生混淆, “零偏差” 看上去好像是偏差值为 0, 然而这里指的是 “null model” 的偏差值, 所以我们继续沿用 “空” 来翻译 “null”。

$$D_0 = 2\phi(\ln L_f - \ln L_0) \quad (9.1.24)$$

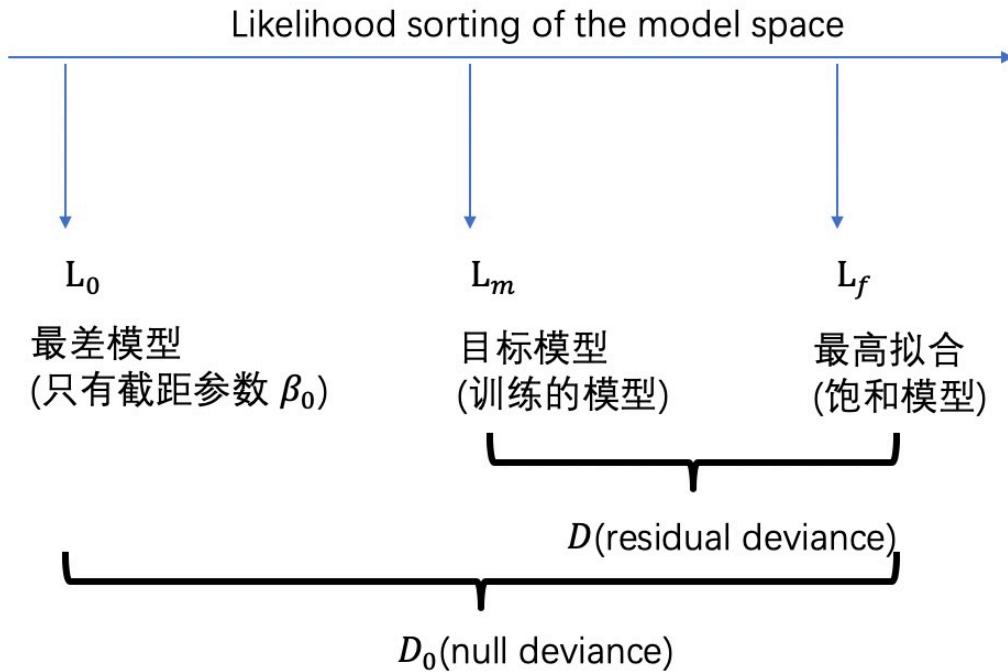


图 9.1.2: The residual deviance (D) and the null deviance (D_0).

用符号 D 表示拟合模型的残差偏差统计量, 可以用一张图来说明 D_0 和 D 之间的关系。我们拟合的模型, 可以看做是在 null model 的基础上增加预测变量 x_1, x_2, \dots, x_p 得到更小的偏差 (deviance)。

可以用 D 替代 RSS , D_0 替代 TSS , 这样就得到一个偏差版本的 R^2 统计量。

$$R^2 = 1 - \frac{D}{D_0} \quad (9.1.25)$$

https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/

校正 R^2

实际上原始版本的 R^2 统计量是存在一个缺陷的。 R^2 用来衡量添加一个预测 (特征) 变量后, 模型对数据的拟合是否变得更好, 然而实际上, 只要添加了新的预测 (特征) 变量, R^2 的值都会增加, 至少不会减少, 无论新增的这个预测变量 X 是否和 Y 相关, 也就是说即使 X 和 Y 完全不相关, 添加之后也会导致 R^2 变大, 这里我们省略证明过程。这就导致在多维 (多个特征变量) 模型中, R^2 的值不再可靠。针对这种情况, 提出了改进版本的 R^2 , 校正 (Adjusted) R^2 , 记作 \bar{R}^2 , 也可以记作 R^2_{adj} 。

$$\bar{R}^2 = 1 - (1 - R^2) \frac{N - 1}{N - p - 1} \quad (9.1.26)$$

其中 N 是观测样本的数据, p 是模型参数的数量 (不包括截距参数)。校正版的 R^2 会惩罚没有意义的特征, 添加了无意义的特征后, \bar{R}^2 的值甚至会变小, 这使得 \bar{R}^2 可以用来检验新增加的特征对模型是否有意义。在单特征的模型中, \bar{R}^2 与 R^2 的效果是一样的, 在多特征的模型中, \bar{R}^2 是更好的, 因此建议大家尽量使用 \bar{R}^2 。

9.1.5 广义皮尔逊卡方统计量

在 GLM 中, 另一个常用的拟合度统计量是广义皮尔逊卡方统计量 (generalized Pearson chi-square statistic) , 其计算公式为

$$\begin{aligned}\chi^2 &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{a(\phi)\nu(\hat{\mu}_i)} \\ &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{V(\mu_i)}\end{aligned}\tag{9.1.27}$$

其中 $V(\mu_i) = a(\phi)\nu(\hat{\mu}_i)$ 表示模型的方差。和偏差统计量类似的情况, 很多资料中会省略分散函数 $a(\phi)$, 直接定义为如下形式

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)}\tag{9.1.28}$$

然而这是不准确的, 非常容易产生误导, 只有当 $a(\phi) = 1$ 时, 才能省略掉。也有些资料把公式 (9.1.28) 称为皮尔逊卡方统计量, 而把公式 (9.1.27) 称为尺度化 (scaled) 的皮尔逊卡方统计量。为了简单清晰表达, 如无特别说明, 本书中默认使用公式 (9.1.27) 的完整定义表示皮尔逊卡方统计量。

顾名思义, 皮尔逊卡方统计量的渐近分布是卡方分布, 这和偏差统计量是一样的, 同样其自由度 (期望) 是样本数量减去模型参数数量, $N - p$ 。偏差统计量是基于最大似然估计的, 因此其再基于最大似然估计的嵌套模型比较时有很大的优势, 而皮尔逊卡方统计量胜在可解释性更强。对于高斯模型, 有 $\nu(\mu) = 1, a(\phi) = 1$, 其皮尔逊卡方统计量、偏差统计量以及平方损失都是等价的, 并且都是精确服从卡方分布的。

平方损失, 亦即残差的平方和 (residual sum of squares, RSS), 是样本观测值和模型拟合值之间误差的平方和, 这非常直观, 易于理解。但是对于不同场景的观测样本, 其取值范围差别很大, RSS 值难以直接评判大小。皮尔斯卡方统计量在 RSS 的基础上除以模型分方差, 可以看成 RSS 的归一化版本, 从误差的绝对值变成多少个标准差, 有利于对其值进行直观上的大小比较。

对于分散参数 ϕ 值未知的模型和场景, 可以利用皮尔逊卡方统计量得到分散参数 ϕ 的一个估计值。 χ^2 统计量的渐近分布是卡方分布, 并且其期望为 $N - p$, 因此有

$$\mathbb{E}[X^2] = \mathbb{E} \left[\sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{a(\phi)\nu(\hat{\mu}_i)} \right] = N - p\tag{9.1.29}$$

利用这个特点可以近似的得到 $a(\phi)$ 。

$$a(\phi) = \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{(N - p)\nu(\hat{\mu}_i)}\tag{9.1.30}$$

当 $a(\phi) = \phi$ 时, 分散参数的近似估计值为

$$\hat{\phi} = \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{(N - p)\nu(\hat{\mu}_i)}\tag{9.1.31}$$

当 $a(\phi) = \phi/w_i$ 时, 分散参数的近似估计值为

$$\hat{\phi} = \sum_{i=1}^N \frac{w_i(y_i - \hat{\mu}_i)^2}{(N - p)\nu(\hat{\mu}_i)}\tag{9.1.32}$$

9.2 残差分析 (Residual analysis)

在评估模型时, 残差 (residual) 用于衡量我们的每个样本观察值与拟合值之间的差异。一个观测值影响估计系数的程度是一种影响的度量。Pierce 和 Schafer (1986) 以及 Cox 和 Snell (1968) 对 GLM 中残差的各种定义提供了出色的调查。在以下各节中, 我们介绍为 GLM 提出的几种残差的定义。但是目前的文献中, 对 GLM 中各类残差的定义缺乏统一的术语, 导致容易产生混淆, 因此我们尽量保留使用英文术语, 以方便与其他书籍和论文进行比较。

残差可以用来探索模型拟合的充分性。

通常残差 (residual) 也用符号 r 表示, 这和响应函数 r 在符号上重复了, 所以本节在出现响应函数的地方我们用链接函数的反函数 g^{-1} 表示。另外残差 (residual) 都是定义在单条样本上的, 以下所有残差的定义中下标 i 都表示第 i 条样本。

9.2.1 Response residuals

Response residuals, 也叫作 raw residuals, 其定义十分简单直接, 就是样本的观测值 (真实值) y_i 和模型拟合值 (预测值) \hat{y}_i 之间的差值。

$$r_i^R = y_i - \hat{y}_i \quad (9.2.1)$$

在 GLM 中, 拟合值 \hat{y}_i 就是响应函数的输出值, 并且表示响应变量的期望值 $\hat{y}_i = \hat{\mu}_i = g^{-1}(\eta_i)$ 。因此在 GLM 中, 上式也可以写成:

$$r_i^R = y_i - \hat{\mu}_i \quad (9.2.2)$$

在之后的残差定义中, 我们都使用 $\hat{\mu}_i$ 表示样本的拟合值。

9.2.2 Working residuals

工作残差 (working residuals) 是在模型收敛时的残差, working response 和 linear predictor 之间的差值。

$$r_i^W = (y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \hat{\mu}_i} = (y_i - \hat{\mu}_i) g'(\hat{\mu}_i) \quad (9.2.3)$$

其中 $\frac{\partial \eta}{\partial \mu}$ 是链接函数的导数 $g'(\mu_i)$ 。这里我们回归一下 IRLS 算法迭代过程中 Z 项的计算公式, working residuals 就是 Z 的一部分。

$$Z^{(t)} = \left\{ (y - \mu) \left(\frac{\partial \eta}{\partial \mu} \right) + \eta^{(t)} \right\}_{(n \times 1)} \quad (9.2.4)$$

9.2.3 Partial residuals

Partial residuals 用于评估每个预测变量 (predictor) 的, 并因此针对每个预测变量进行计算。O’ Hara Hines 和 Carter (1993) 讨论了这些残差在评估模型拟合中的图形使用。

$$r_{ij}^T = r_i^W + x_{ij} \beta_j \quad (9.2.5)$$

上式中 r_i^W 表示的是样本 i 的 Working residuals, $x_{ij} \beta_j$ 仅是第 j 维输入特征的线性预测器。Partial residuals 评估的是单一维度特征的预测器的残差。

9.2.4 Pearson residuals

皮尔逊残差 (Pearson residuals) 是工作残差 (working residuals) 的重新缩放版本。皮尔逊残差平方的总和等于皮尔逊卡平方统计量 (Pearson chi-squared statistic)。

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\nu(\hat{\mu}_i)}} \quad (9.2.6)$$

分母是方差函数的平方根, 缩放将残差置于相似的方差尺度上。残差的绝对值较大, 表明该模型无法满足特定的观察要求。检测异常值的常见诊断方法是绘制标准化的皮尔逊残差 (standardized Pearson residuals) 与观察值的关系。表 A.9 中列出了常见族的皮尔逊残差公式。

9.2.5 Deviance residuals

偏差 (deviance) 在推导 GLM 和结果推断中起着关键作用。偏差残差 (deviance residual) 是每个观察值相对于总体偏差 (overall deviance) 的增量。这些残差很常见, 通常是标准化的 (standardized), 学生化的 (studentized) 或两者兼而有之。偏差残差 (deviance residual) 是基于卡方分布的, 其公式如下。

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{\hat{d}_i^2} \quad (9.2.7)$$

GLM 中不同分布族:math:hat{d}_i^2` 计算法不同, 表 xxxx 给出了各分布族的计算方法。通常, 在模型检查中, 偏差残差 (标准化或非标准化残差) 优于 Pearson 残差, 因为其分布特性更接近于线性回归模型中产生的残差。

9.2.6 Adjusted deviance residuals

可以调整 (校正) 偏差残差以使收敛到极限正态分布更快。调整后术语。

9.2.7 Likelihood residuals

9.2.8 Score residuals

这些是用于计算方差三明治估计的分数。分数与已优化的分数函数 (估计方程) 有关。

$$r_i^S = \frac{y_i - \hat{\mu}_i}{\nu(\hat{\mu}_i)g'(\hat{\mu}_i)} \quad (9.2.8)$$

9.2.9 Anscombe residuals

Anscombe residuals 是以弗兰克 · 安斯科姆 (Frank Anscombe) 的名字命名的。弗兰克 · 安斯科姆 (Frank Anscombe) 作为 Hotelling (1953) 讨论中的回应者之一, 贡献了以下残差定义。首先, 让

9.3 模型选择 (model selection)

模型选择 (model selection) 是指从给定数据的一组候选统计模型中选择出最佳模型的过程。模型选择是一个既可以应用于不同类型的模型 (例如逻辑回归, SVM, KNN 等), 又可以应用于配置了不同超参数 (例如 SVM 中的不同内核) 的相同类型模型的过程。

在深入探讨不同的模型选择 (model selection) 方法以及何时使用它们之前, 我们需要弄清楚模型选择 (model selection) 与模型评估 (model evaluation) 之间的区别。模型选择 (model selection) 关注的是 **模型训练阶段** 的效果, 在给定的数据集下, 评价模型的训练误差, 即哪个候选模型拟合的更好。模型评估 (model evaluation) 旨在评估所选模型的泛化误差 (generalization error), 即所选模型在未知数据上的表现如何。

但是为什么我们需要区分模型选择和模型评估? 原因是过度拟合 (overfitting)。一个模型可以在训练阶段表现的非常好, 比如饱和模型 (saturated model) 可完美的拟合每一条训练集样本, 但是其在未知数据上很可能表现的一塌糊涂。显然, 一个好的机器学习模型, 它不仅可以在训练过程中表现出色, 而且可以在未知数据上表现出色。因此, 在将模型交付生产之前, 我们应该相当确定, 当面对新数据时, 模型的性能不会降低。

训练一个模型是相对简单的事情, 但选择一个“合适”的模型却是一件有挑战的事情。首先, 我们需要克服“最佳”模型的想法。考虑到数据中的统计噪声, 数据样本的不完整以及每种不同模型类型的局限性, 所有模型都具有一定的预测误差。因此, 完美或最佳模型的概念没有用。相反, 我们必须寻求一个“足够好”的模型。

选择最终模型时我们关心什么? 不同的应用场景可能会有不同的要求, 例如可维护性、有限的模型复杂性、较强的解释性, 等等, 有时, 具有较低性能但更容易理解的模型可能是优选的。而有时更倾向于效果好的模型, 无需关注计算复杂度。因此, “足够好”的模型可能涉及很多东西, 并且特定于您的项目。

通常有三种方法来选择模型。

- Train, Validation, and Test datasets.
- 概率测度 (probabilistic measures): 通过样本误差和复杂度选择模型。
- 重采样方法 (resampling methods): 通过估计的样本外误差选择模型。

在理想情况下, 我们拥有足够多的数据, 最简单可靠的模型选择方法, 将数据分成训练集 (training dataset)、验证集 (validation dataset)、测试集 (test dataset)。在训练集上拟合候选模型, 在验证数据集上进行调整和优化, 最后根据所选度量 (例如准确性或误差) 指标在测试数据集上选择表现最佳的模型。这种方法的致命问题是, 它需要大量数据。鉴于我们很少有足够的数据, 甚至无法判断什么将是足够的, 因此对于大多数预测性建模问题而言, 这是不切实际的。在数据不足的情况下, 经常使用后两种方法: 概率测度 (probabilistic measures) 和重采样方法 (resampling methods)。

概率测度 (probabilistic measures)

概率测度 (probabilistic measures) 依据候选模型在训练数据集上的 **模型表现 (Model Performance)** 和 **模型的复杂性 (Model Complexity)** 对候选模型进行分析评分。模型复杂性的概念可用于创建有助于模型选择的度量。

众所周知, 训练误差偏向乐观, 因此不是选择模型的良好基础。可以根据认为训练错误的乐观程度来惩罚表现。通常使用特定于算法的方法 (通常为线性方法) 来实现此目的, 该方法会根据模型的复杂性对分数进行惩罚。历史上已经提出了各种“信息标准 (Information Criterion)”, 试图通过增加惩罚项来补偿最大似然性的偏差, 以补偿更复杂模型的过度拟合。

根据奥卡姆剃刀 (Occam's razor) 的原理, 给定具有相似预测或解释能力的候选模型, 最简单的模型很可能是最佳选择。具有较少参数的模型复杂性更低, 因此, 首选简单模型, 因为它平均而言可能会更好地泛化。四种常用的概率模型选择度量包括:

- 赤池信息准则 (Akaike Information Criterion,AIC)。
- 贝叶斯信息准则 (Bayesian Information Criterion,BIC)。
- 最小描述长度 (Minimum Description Length,MDL)。
- 结构风险最小化 (Structural Risk Minimization,SRM)。

当使用更简单的线性模型（例如线性回归或逻辑回归）时，概率度量是适当的，其中模型复杂度损失的计算（例如样本偏差）是已知的并且易于处理的。

例如 赤池信息量准则 (*Akaike information criterion, AIC*) 和 贝叶斯信息准则 (*Bayesian information criterion, BIC*)，两者都会惩罚模型参数的数量，但会奖励训练集的拟合优度，因此，最佳模型是 AIC / BIC 最低的模型。BIC 会更严厉地惩罚模型复杂性，因此 BIC 倾向于“错误得多”但更简单的模型。虽然这允许在不使用验证集的情况下进行模型选择，但它只能严格应用于参数线性的模型，即使它通常也适用于更一般的情况，例如适用于广义线性模型，例如逻辑回归，等等。

重采样方法 (resampling methods)

重采样方法旨在估计训练样本外数据的模型性能。这是通过将训练数据集分为子训练集和测试集，在子训练集上拟合模型并在测试集上对其进行评估来实现的。然后可以重复此过程多次，并报告每个试验的平均性能。

这是对样本外数据的模型性能进行的蒙特卡洛估计，尽管每个试验并非严格独立，这取决于所选择的重采样方法，但是同一数据可能会在不同的训练数据集或测试数据集中多次出现。

三种常见的重采样模型选择方法包括：

- 随机训练/测试分组 (Random train/test splits)。
- 交叉验证 (Cross-Validation), k-fold, 留一法 (LOOCV) 等。
- Bootstrap。

在概率测度 (probabilistic measures) 不可用的时候，可以大使用重采样方法。到目前为止，使用最广的是交叉验证方法系列。

9.3.1 Criterion measures

首先，我们定义一些符号

$$\begin{aligned}
 p &= \text{模型的参数数量} \\
 N &= \text{观测样本的数量} \\
 L &= \text{模型的似然} \\
 \ell &= \text{模型的对数似然} \\
 D &= \text{模型的偏差 deviance} \\
 G^2 &= \text{模型的似然比检验}
 \end{aligned} \tag{9.3.1}$$

接下来，我们提供用于模型比较的 AIC 和 BIC 的量度的公式。这些度量指标试图找到在模型拟合优度和模型负责度之间的平衡点。

AIC

赤池信息准则 (*Akaike information criterion, AIC*) 是样本外预测误差的估计值，度量的时在给定数据集下统计模型的相对质量。在给定数据集的候选模型集合中，AIC 估计每个模型相对于其他模型的质量。因此，AIC 提供了一种模型选择的方法。

AIC 建立在信息论 (information theory) 的基础上。当使用统计模型来表征数据的生成过程时，几乎永远不会是精确的；统计模型一定会丢失一些信息。AIC 估计模型丢失的相对信息量：模型丢失的信息越少，该模型的质量越高。在估算模型丢失的信息量时，AIC 会在模型拟合优度 (goodness of fit) 和模型复杂性之间进行权衡。换句话说，AIC 同时处理过度拟合和欠拟合的风险。*Akaike information criterion* 是由制定该标准的统计学家 Hirotugu Akaike 命名的。现在，它构成了统计基础范例的基础，并且广泛用于统计推断 (statistical inference)。

AIC 可用于比较嵌套模型或非嵌套模型。信息准则对目标模型丢失的信息的度量，目的是找到信息丢失最少的模型。

$$AIC = 2p - 2\ell \quad (9.3.2)$$

$\ell(M_k)$ 代表了模型拟合能力， p 代表了模型的复杂程度。注意，AIC 的绝对值是没有意义的，模型之间的相对大小才有意义。当比较的两个模型拟合能力(似然)相差较大时，AIC 受到似然值的影响更大一些；当两个模型拟合能力(似然)相当时，AIC 受到模型参数数量 p 的影响更大一些。

参数数量 p 的项是对较大的参数变量列表的一种惩罚，AIC 特别适合比较具有相同链接和方差函数但具有不同参数变量列表的 GLM。当模型嵌套时，我们将惩罚项视为从模型中消除候选预测变量所需的精度。

我们需要注意如何计算 AIC，上面的定义包括模型对数似然。在 GLM 中，参数估计过程通常不是基于似然的，而是基于偏差 (deviance) 的，不能使用偏差值去算 AIC 的值，因为偏差的计算过程中不包括归一化项 $c(y_i, \phi)$ ，而在 GLM 中不同的分布拥有不同的归一化项。

后来又衍生出了以几种 AIC 的变种版本，这里给出两种替代方法。第一种是 Sugiura (1978) 和 Hurvich and Tsai (1989) 提出的校正或有限样本 (finite-sample) AIC。第二个是 Hannan 和 Quinn (1979) 描述的 AIChq。这些版本的公式 (未缩放) 为：

$$\begin{aligned} AICc &= 2 \frac{p(p+1)}{N-p-1} + 2p - 2\ell \\ AIChq &= 2p \ln\{\ln(N)\} - 2\ell \end{aligned} \quad (9.3.3)$$

我们如何确定两个 AIC 统计数据之间的差异是否足够大，足以使我们得出结论，一个模型比另一个模型更合适？特别是，具有较低 AIC 统计量的模型是否比其他模型更受青睐？尽管我们知道具有较小 AIC 的模型是更可取的，但尚无可用于计算 p-值的特定统计检验。Hilbe (2009) 根据模拟研究设计了一个主观表，可用于做出比较无标度 AIC 度量的决策。

Difference: Model A and B	Decision (assuming $A < B$)
$0.00 < \text{difference} \leq 2.50$	No difference in models
$2.50 < \text{difference} \leq 6.00$	Prefer A if $n > 256$
$6.00 < \text{difference} \leq 9.00$	Prefer A if $n > 64$
$10.00 < \text{difference}$	Prefer A

BIC

在统计中，贝叶斯信息准则 (BIC) 或 Schwarz 信息准则 (也称为 SIC, SBC, SBIC) 是用于在有限的一组模型中选择模型的标准，BIC 最低的模型是首选。它部分基于似然函数，并且与 AIC 密切相关。

$$BIC_\ell = p \ln(N) - 2\ell \quad (9.3.4)$$

与 AIC 相反，BIC 包含的惩罚项随着样本数量的增加而变得更加严格。该特征反映了可用于检测重要性的能力。Raftery (1995) 提出了一个基于偏差版本的 BIC，Raftery 的目的是使用替代版本的 BIC 在 GLM 模型之间进行选择。

$$BIC_D = D - p \ln(N) \quad (9.3.5)$$

当比较非嵌套模型时，可以根据两个模型的 BIC 统计数据之间的差的绝对值来评估模型的偏好程度。Raftery (1995) 给出的用于确定相对偏好的量表。

Difference	Degree of preference
0–2	Weak
2–6	Positive
6–10	Strong
>10	Very strong

Minimum Description Length

<https://machinelearningmastery.com/probabilistic-model-selection-measures/>

G

模型检验

我们基于样本训练模型，基于样本计算模型拟合优度指标，并给出模型好坏的结论。然而，这一切都是建立随机样本的基础上，模型拟合优度指标也是一个随机量，我们的结论是根据样本推断 (influence) 得出的，推断得出结论不是百分百准确的，这就需要同时给出这个结论的可靠程度，而这就是统计推断 (statistical inference) 所做的事情。

上一章我们介绍了 GLM 中评价模型拟合好坏程度的常见指标，以及这些指标的定义和计算方法，但是没有说明如何根据指标值得出结论，本章我们探讨如何根据拟合优度指标的值得出模型优劣的结论，以及结论的可靠程度。再讨论 GLM 推断方法前，先简单讲解一下统计学中的推断和检验的理论，GLM 的推断过程就是统计学推断理论的一个应用。

10.1 GLM 中的抽样分布

无论是置信区间还是假设检验都需要知道统计量的抽样分布，因此要对 GLM 拟合优度进度推断和检验，就要知道各个拟合优度指标的抽样分布。本节我们推导一下 GLM 中一些关键统计量的抽样分布。

如果响应变量是正态分布，则通常可以准确确定一些统计量的抽样分布。反之，如果响应变量不是正态分布，就需要依赖中心极限定理，找到其大样本下的近似分布。注意，这些结论的成立都是有一些前提条件的，对于来自属于指数族分布的观测数据，特别是对于广义线性模型，确实满足了必要条件。在本节我们只给出统计量抽样分布的一些关键步骤，Fahrmeir 和 Kaufmann (1985) 给出了广义线性模型抽样分布理论的详细信息。

如果一个统计量 S ，其渐近服从正态分布 $S \sim \mathcal{N}(\mathbb{E}[S], Var(S))$ ，其中 $\mathbb{E}[S]$ 和 $Var(S)$ 分别是 S 的期望和方差，则近似的有：

$$\frac{S - \mathbb{E}[S]}{\sqrt{Var(S)}} \sim \mathcal{N}(0, 1) \quad (10.1.1)$$

根据卡方分布的定义，等价有

$$\frac{(S - \mathbb{E}[S])^2}{Var(S)} \sim \chi^2(1) \quad (10.1.2)$$

如果 S 是一个向量 $\mathbf{S}^T = [S_1, \dots, S_p]$ ，上述结论可以写成向量的模式。

$$(\mathbf{S} - \mathbb{E}[\mathbf{S}])^T \mathbf{V}^{-1} (\mathbf{S} - \mathbb{E}[\mathbf{S}]) \sim \chi^2(p) \quad (10.1.3)$$

其中 \mathbf{V} 是协方差矩阵，并且必须是非奇异矩阵。

10.1.1 得分统计量 (score statistic)

我们已经知道似然函数及其一阶导数都是一个关于样本的函数，所以似然函数及其一阶导数都是统计量 (statistic)。似然函数的一阶导数又叫做得分函数 (score function)，也称为得分统计量 (score statics)。假设 Y_1, \dots, Y_N 是相互独立的 GLM 样本变量，这里我们强调 Y_i 是一个随机变量，所以用大写符号表示。其中有 $\mathbb{E}[Y_i] = \mu_i$ ， $g(\mu_i) = \beta^T x_i = \eta_i$ ，自然参数 θ_i 是一个关于 μ_i 函数。GLM 模型的对数似然函数为

$$\ell = \sum_{i=1}^N \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad (10.1.4)$$

根据 章节 8.1 的内容，GLM 对数似然函数的一阶导数，得分统计量为：

$$\begin{aligned} U_j &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \left(\frac{\partial \ell_i}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{\partial \eta_i}{\partial \beta_j} \right) \\ &= \sum_{i=1}^N \left\{ \frac{Y_i - b'(\theta_i)}{a(\phi)} \right\} \left\{ \frac{1}{\nu(\mu_i)} \right\} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= \sum_{i=1}^N \frac{Y_i - \mu_i}{a(\phi)\nu(\mu_i)} \left(\frac{\partial \mu}{\partial \eta} \right)_i x_{ij} \\ &= \sum_{i=1}^N \frac{Y_i - \mu_i}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_{ij} \end{aligned} \quad (10.1.5)$$

注意下标 j 表示的参数向量的下标， U_j 是 β_j 的一阶偏导数。对于任意的样本 Y_i 都有 $\mathbb{E}[Y_i] = \mu_i$ ，因此有：

$$\mathbb{E}_{Y_i}[U_j] = 0 \quad (10.1.6)$$

U 的协方差矩阵就是信息矩阵 \mathcal{J} 。

$$\mathcal{J}_{jk} = \mathbb{E}[U_j U_k] \quad (10.1.7)$$

在 章节 3 参数估计 我们讲过，信息矩阵 \mathcal{J} 又等于对数似然函数二阶偏导数的期望的负数，

$$\mathcal{J} = -\mathbb{E}[\ell''] = -\mathbb{E}[U'] \quad (10.1.8)$$

如果只有一个参数 β ，得分统计量 U 是一个标量，其渐近服从正态分布。

$$U \sim \mathcal{N}(0, \mathcal{J}) \text{ 或者 } \frac{U}{\sqrt{\mathcal{J}}} \sim \mathcal{N}(0, 1) \quad (10.1.9)$$

根据卡方分布的定义，也可以写成

$$\frac{U^2}{\mathcal{J}} \sim \chi^2(1) \quad (10.1.10)$$

如果 β 是一个参数向量， $\beta^T = [\beta_1, \dots, \beta_p]$ ，则 U 是得分向量 $\mathbf{U}^T = [U_1, \dots, U_p]$ ，此时 \mathbf{U} 渐近服从多维正态分布 (multivariate Normal distribution, MVN)。

$$\mathbf{U} \sim MVN(\mathbf{0}, \mathcal{J}) \quad (10.1.11)$$

在大样本下有

$$\mathbf{U}^T \mathcal{J}^{-1} \mathbf{U} \sim \chi^2(p) \quad (10.1.12)$$

高斯分布的得分统计量

令 Y_1, \dots, Y_N 是独立同分布的高斯随机变量, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, 其中 μ 是未知的, σ^2 是已知的常量, 并且所有变量 Y_i 都是拥有同样的均值参数 μ 和常量方差 σ^2 。其对数似然函数为:

$$\ell(\mu; Y, \sigma) = -\frac{1}{2\sigma^2} \sum_{i=1}^N (Y_i - \mu)^2 - N \ln(\sigma\sqrt{2\pi}) \quad (10.1.13)$$

其得分统计量 (一阶导数) 为:

$$U = \frac{d\ell}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^N (Y_i - \mu) \quad (10.1.14)$$

可以把样本均值统计量 $\sum_{i=1}^N Y_i = N\bar{Y}$ 代入到 U ,

$$U = \frac{1}{\sigma^2} (N\bar{Y} - N\mu) = \frac{N}{\sigma^2} (\bar{Y} - \mu) \quad (10.1.15)$$

通过令 $U = 0$ 可以得到参数 μ 的最大似然估计量 $\hat{\mu} = \bar{Y}$ 。

现在看下统计量 U 的期望和方差, 其期望是

$$\mathbb{E}[U] = \frac{N}{\sigma^2} (\mathbb{E}[Y_i] - \mu) = \frac{N}{\sigma^2} (\mu - \mu) = 0 \quad (10.1.16)$$

统计量 U 的方差是

$$\begin{aligned} Var(U) &= Var\left[\frac{1}{\sigma^2} \sum_{i=1}^N (Y_i - \mu)\right] \\ &= \frac{1}{\sigma^4} Var\left[\sum_{i=1}^N (Y_i - \mu)\right] \\ &= \frac{1}{\sigma^4} \sum_{i=1}^N Var(Y_i) \\ &= \frac{N}{\sigma^2} \\ &= \mathcal{J} \end{aligned} \quad (10.1.17)$$

结合公式 (10.1.14) 和公式 (10.1.17) 有

$$\frac{U}{\sqrt{\mathcal{J}}} = \frac{\sqrt{N}(\bar{Y} - \mu)}{\sigma} \quad (10.1.18)$$

根据结论公式 (10.1.9), 这个统计量拥有渐近分布 $\mathcal{N}(0, 1)$, 同理有

$$U^T \mathcal{J}^{-1} U = \frac{U^2}{\mathcal{J}} = \frac{N(\bar{Y} - \mu)^2}{\sigma^2} \sim \chi^2(1) \quad (10.1.19)$$

二项分布的得分统计量

现在假设 $Y_i \sim Bin(n, \pi)$, 对数似然函数为

$$\ell(\pi; y) = \sum_{i=1}^N \left[Y_i \ln \pi + (n - Y_i) \ln(1 - \pi) + \ln \binom{n}{Y_i} \right] \quad (10.1.20)$$

得分统计量是

$$\begin{aligned}
 U &= \frac{d\ell}{d\pi} \\
 &= \sum_{i=1}^N \left[\frac{Y_i}{\pi} - \frac{n - Y_i}{1 - \pi} \right] \\
 &= \sum_{i=1}^N \frac{Y_i - n\pi}{\pi(1 - \pi)} \\
 &= \frac{1}{\pi(1 - \pi)} \sum_{i=1}^N (Y_i - n\pi)
 \end{aligned} \tag{10.1.21}$$

然后代入样本均值统计量, $\sum_{i=1}^N Y_i = N\bar{Y}$, 可以把 U 改写成

$$U = \frac{N(\bar{Y} - n\pi)}{\pi(1 - \pi)} \tag{10.1.22}$$

因为 $\mathbb{E}[Y_i] = n\pi$, 所以 $\mathbb{E}[U] = 0$ 。又因为 $Var(Y_i) = n\pi(1 - \pi)$, 所以

$$\begin{aligned}
 Var(U) &= Var \left[\sum_{i=1}^N \frac{Y_i - n\pi}{\pi(1 - \pi)} \right] \\
 &= \frac{1}{\pi^2(1 - \pi)^2} Var \left[\sum_{i=1}^N (Y_i - n\pi) \right] \\
 &= \frac{1}{\pi^2(1 - \pi)^2} \sum_{i=1}^N Var(Y_i - n\pi) \\
 &= \frac{1}{\pi^2(1 - \pi)^2} \sum_{i=1}^N Var(Y_i) \\
 &= \frac{1}{\pi^2(1 - \pi)^2} \sum_{i=1}^N n\pi(1 - \pi) \\
 &= \frac{Nn}{\pi(1 - \pi)} \\
 &= \mathcal{J}
 \end{aligned} \tag{10.1.23}$$

因此有

$$\frac{U}{\sqrt{\mathcal{J}}} = \frac{\sqrt{N}(\bar{Y} - n\pi)}{\sqrt{n\pi(1 - \pi)}} \sim \mathcal{N}(0, 1) \tag{10.1.24}$$

10.1.2 最大似然估计量

在继续讨论其他统计量的抽样分布之前, 我们先回顾一下泰勒级数近似 (Taylor series approximation), 后续统计量抽样分布的推导依赖泰勒级数。

定义一个单变量的函数 $f(x)$, 对于函数上的某个点 $x = t$ 的附近有如下近似成立:

$$f(x) = f(t) + (x - t) \left[\frac{df}{dx} \right]_{x=t} + \frac{1}{2}(x - t)^2 \left[\frac{d^2f}{dx^2} \right]_{x=t} + \dots \tag{10.1.25}$$

单参数模型

对于一个拥有单个参数 β 的对数似然函数 $\ell(\beta)$, 假设其估计值是 b , 在估计值 b 的附近用泰勒级数展开为

$$\ell(\beta) = \ell(b) + (\beta - b)U(b) + \frac{1}{2}(\beta - b)^2 U'(b) \tag{10.1.26}$$

其中我们用 $U(b)$ 表示 $\ell(\beta = b)$ 的一阶导数, 用 $U'(b)$ 表示 $\ell(\beta = b)$ 的二阶导数。我们知道 $\mathbb{E}[U'] = -\mathcal{J}$, 现在我们用 $U'(b)$ 的期望值代替其自身,

$$\ell(\beta) = \ell(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^2\mathcal{J}(b) \quad (10.1.27)$$

现在我们把得分函数 U 也在 $\beta = b$ 的附近展开, 但这里我们只要前两项, 忽略二阶以及更高阶的项。

$$U(\beta) = U(b) + (\beta - b)U'(b) \quad (10.1.28)$$

同样, 可以用 $-\mathcal{J}$ 代替 $U'(b)$,

$$U(\beta) = U(b) - (\beta - b)\mathcal{J}(b) \quad (10.1.29)$$

多参数

如果 β 是一个向量, 表达式仍然适用, 只需要把参数改成向量即可。对数似然函数的近似展开式,

$$\ell(\beta) = \ell(\mathbf{b}) + (\beta - \mathbf{b})^T \mathbf{U}(\mathbf{b}) - \frac{1}{2}(\beta - \mathbf{b})^T \mathcal{J}(\mathbf{b})(\beta - \mathbf{b}) \quad (10.1.30)$$

得分函数的近似展开式,

$$\mathbf{U}(\beta) = \mathbf{U}(\mathbf{b}) - (\beta - \mathbf{b})\mathcal{J}(\mathbf{b}) \quad (10.1.31)$$

标量和向量在公式以及推导上没有本质区别, 所以后续不再显示的区分向量形式。

我们假设似然估计值 $\hat{\beta} = b$ 和参数的真实值应该是很接近的, 处在 b 附近的 β_{true} 是参数的真实值。注意参数的似然估计量 b 是一个统计量 (随机变量), 而参数真实值 β_{true} 是一个数值。公式 (10.1.31) 可以用来推导最大似然估计量 b 的抽样分布, 按照定义, b 是通过极大化对数似然函数 $\ell(\beta)$ 得到的估计量, 并且有 $U(b) = 0$ 。公式 (10.1.31) 简化为

$$U(\beta_{true}) = -\mathcal{J}(b)(\beta_{true} - b) \quad (10.1.32)$$

或者等价的,

$$(b - \beta_{true}) = \mathcal{J}(\beta_{true})^{-1}U(\beta_{true}) \quad (10.1.33)$$

我们知道 $\mathbb{E}[U] = 0$, 如果把 \mathcal{J} 看做一个常量, 则有

$$\mathbb{E}[(b - \beta_{true})] = \mathbb{E}[\mathcal{J}^{-1}U] = \mathcal{J}^{-1}\mathbb{E}[U] = 0 \quad (10.1.34)$$

因此有

$$\mathbb{E}[b] = \beta_{true} \quad (10.1.35)$$

最大似然估计量 b 的期望就是参数的真实值 β_{true} 。现在来看下方差 $Var(b)$ 。

$$\begin{aligned} Var(b) &= \mathbb{E}[(b - \mathbb{E}[b])(b - \mathbb{E}[b])^T] \\ &= \mathbb{E}[(b - \beta_{true})^T(b - \beta_{true})] \\ &= \mathbb{E}[\mathcal{J}^{-1}UU^T\mathcal{J}^{-1}] \\ &= \mathcal{J}^{-1}\mathbb{E}[UU^T]\mathcal{J}^{-1} \\ &= \mathcal{J}^{-1} \end{aligned} \quad (10.1.36)$$

注意, 上述推导过程中 $\mathbb{E}[UU^T] = Var(U) = \mathcal{J}$ 。似然估计量 b 的渐近分布是

$$b \sim \mathcal{N}(\beta_{true}, \mathcal{J}^{-1}) \quad (10.1.37)$$

如果总体分布是正态分布, 似然估计量 b 就是精确服从正态分布, 而不是渐近了。如果总体分布是非正态分布, 似然估计量 b 就是渐近服从正态分布。

参考本节开始时公式 (10.1.3) 的理论, 公式 (10.1.37) 另一个等价的表示是

$$(b - \beta_{true})^T \mathcal{J}(b)(b - \beta_{true}) \sim \chi^2(p) \quad (10.1.38)$$

公式 (10.1.38) 又叫做 Wald 统计量。

10.1.3 偏差 (deviance) 统计量

我们继续用符号 b 表示拟合模型参数的最大似然估计量, 回顾一下泰勒展开式公式 (10.1.30), 其中满足 $U(b) = 0$, 变化一下公式, 则近似有如下等式成立。

$$\ell(\beta) - \ell(b) = -\frac{1}{2}(\beta - b)^T \mathcal{J}(b)(\beta - b) \quad (10.1.39)$$

继续移项, 可得到如下统计量

$$2[\ell(b) - \ell(\beta)] = (b - \beta)^T \mathcal{J}(b)(b - \beta) \quad (10.1.40)$$

依据公式 (10.1.38) 这个统计量是服从自由度为 p 的卡方分布, p 是参数数量。

$$2[\ell(b) - \ell(\beta)] \sim \chi^2(p) \quad (10.1.41)$$

仔细观察下这个统计量, 其和偏差的定义基本是一致的。我知道偏差 (deviance) D 和对数似然比统计量 (log-likelihood ratio statistic) 是等价的, 其计算公式为

$$D = 2[\ell(b_s; y) - \ell(b_f; y)] \quad (10.1.42)$$

其中, 符号 b_s 表示饱和 (saturated) 模型参数的最大似然估计量, $\ell(b_s; y)$ 表示饱和模型的似然统计量。符号 b_f 表示我们目标拟合 (fitted) 模型参数的最大似然估计量, $\ell(b_f; y)$ 表示拟合模型的似然统计量。注意二者是统计量 (随机变量), 不是数值量。参数向量 b_s 和 b_f 的长度是不同的, 饱和模型的参数数量就等于样本容量 N , 假设拟合模型的参数向量 b_f 的长度是 p , $p < N$ 。现在我们把 D 变换一下。

$$\begin{aligned} D &= 2[\ell(b_s; y) - \ell(b_f; y)] + 2\ell(\beta_s; y) - 2\ell(\beta_s; y) + 2\ell(\beta_f; y) - 2\ell(\beta_f; y) \\ &= \underbrace{2[\ell(b_s; y) - \ell(\beta_s; y)]}_{\chi^2(N)} - \underbrace{2[\ell(b_f; y) - \ell(\beta_f; y)]}_{\chi^2(p)} + \underbrace{2[\ell(\beta_s; y) - \ell(\beta_f; y)]}_{\text{数值 } v} \end{aligned} \quad (10.1.43)$$

其中符号 $\ell(\beta_s; y)$ 表示饱和模型真实参数值的似然值 (模型的理论最大似然值), 是一个数值, 不是统计量。同理 $\ell(\beta_f; y)$ 是拟合模型的理论最大似然值, 也是一个数值。最终统计量 D 可以看做是由三部分组成, 自由度为 N 卡方分布减去自由度为 p 的卡方分布, 再加上一个数值 v 。

根据卡方分布的特性, 统计量 D 渐近服从 **非中心卡方分布**, 其自由度是 $N - p$ 。

$$D \sim \chi^2(N - p, v) \quad (10.1.44)$$

注意偏差统计量 D 是一个 **非中心卡方分布**, 这和之前介绍的统计量不同, v 是非中心参数。 D 的期望值是 $\mathbb{E}[D] = N - p + v$ 。现在来重点看一下 v 的值,

$$v = 2[\ell(\beta_s; y) - \ell(\beta_f; y)] \quad (10.1.45)$$

v 的值是饱和模型的理论最大似然值和拟合模型的理论最大似然值的差, 前者 $\ell(\beta_s; y)$ 的值是固定不变的, 后者 $\ell(\beta_f; y)$ 是你的拟合模型的理论上限, 拟合模型的拟合效果越好, $\ell(\beta_f; y)$ 就越接近前者饱和模型, v 的值也就越小。极限情况下, 拟合模型对样本的拟合能力和饱和模型一样好, 此时 $v = 0$ 。这时偏差 D 就是渐进服从 **中心卡方分布** $\chi^2(N - p)$ 。本节的内容是下一节的理论基础, 对于理解检验过程非常重要。如果难以理解本节的推导过程, 可以先记住以下结论。

重要结论

模型对数据拟合的越好 (越接近饱和模型), 其偏差 D 就越接近中心卡方分布 $\chi^2(N - p)$, 此时偏差统计量 D 的期望就越接近 $N - p$ 。反之如果模型拟合的不好, 偏差统计量 D 就是非中心卡方分布 $\chi^2(N - p, v)$, 其期望值就是 $v + N - p$, 相比于好的模型期望值会偏离 $N - p$ 。后续的比较两个模型效果的假设检验过程就利用这个特性。

如果响应变量 Y 是高斯分布, 则偏差统计量 D 就是确切服从 (非中心) 卡方分布的, 而不是渐近的。如果响应变量 Y 不是高斯分布, 则偏差统计量 D 是渐近服从 (非中心) 卡方分布的。这个特性我们已经多次强调过。

提示: 统计量 D 的计算是需要根据对数似然值计算, 而对数似然值的计算又需要计算 $Var(Y_i) = a(\phi)\nu(\mu_i)$ 。显然要计算对数似然值就需要知道模型的分散参数 ϕ 的值。指数族中部分分布是没有分散参数 ϕ 的, 比如二项分布、多项分布、泊松分布等, 这些模型可以直接计算出统计量 D 的值。然而, 有些指数族分布, 比如高斯分布, 就存在分散参数 $\phi = \sigma^2$, 此时理论上是无法直接计算出 D 的值。这时有两种解决方法, 第一种方法是假设 ϕ 为一个常量值, 传统线性回归模型就是这么干的, 其假设 $\phi = \sigma^2 = 1$ 。第二种方法就是利用其它估计方法得到 ϕ 的一个近似值, 关于 ϕ 估计方法在本书后面再讨论。

10.1.4 参数估计量

10.2 GLM 中的模型检验

我们已经知道偏差统计量是饱和模型的对数似然值和拟合模型的对数似然值的差,

$$D = 2[\ell(b_s; y) - \ell(b_f; y)] \quad (10.2.1)$$

饱和模型的对数似然值 $\ell(b_s; y)$ 代表了模型似然值的理论最大值, 偏差的含义就是拟合模型的似然值和这个上限值差了多少, 偏差越小说明拟合模型对数据的拟合度越好。理论上偏差 D 的取值范围是 $[0, +\infty]$, 然而实际上偏差 D 是不大可能得到一个接近 0 的值。饱和模型虽然似然值最大, 但其是一种极端过拟合 (overfitted) 的状态, 没有学习到任何关于总体的特征, 不具备丝毫泛化能力, **似然值最大并不意味着模型就一定是最好的**。

为了保障模型能从样本数据中学习到总体特征, 拟合模型的参数数量 p 必然是远小于样本容量 N 的, 拟合模型的似然值 $\ell(b_f; y)$ 也必然是远小于饱和模型的似然值 $\ell(b_s; y)$, 因此偏差 D 通常会得到一个比较大的值。并且不同的样本、不同的模型必然会得到不同的 D 值, 通常差异也会比较大。那么当你计算出一个 D 值时, 如何判断模型是好还是坏呢? 以及你的结论可靠吗? 毕竟 D 是一个统计量 (随机量), 仅根据一个值得出结论可信度有多高? 如果你彻底理解了第 4 章 的内容, 那么此时你的脑海中应该已经有答案了。

10.2.1 模型检验

卡方检验

F 检验

我们知道偏差 D 是一个统计量, 并且其抽样分布是卡方分布 $\chi^2(N - p, v)$, 期望值是

$$\mathbb{E}[D] = N - p + v \quad (10.2.2)$$

注意, 样本容量 N 和模型参数数量 p 的值是已知的, 而 v 的值我们是无法计算出的。根据之前的结论, 模型对数据拟合的越好, v 的值就越小, D 的期望值就越接近 $N - p$ 。

$$\mathbb{E}[D] = N - p \quad (10.2.3)$$

那么我们可以基于这个假设对偏差统计量 D 进行推断和检验, 如果模型对数据拟合的足够好, 则统计量 D 的期望值就是 $N - p$, 反之期望值就是 $N - p + v$, 基于此零假设 H_0 和备择假设 H_1 分别是

$$\begin{aligned} H_0 : \mathbb{E}[D] &= N - p \\ H_1 : \mathbb{E}[D] &\neq N - p \end{aligned} \quad (10.2.4)$$

假设显著水平为 $\alpha = 0.05$ ，然后根据统计量 D 的值 $D = d$ 计算出 $P - Value$ ， P 值就是 $D \geq d$ 的概率，可以通过查卡方检验表直接得到。通过比较 $P - Value$ 和 α 得出检验结论。

10.2.2 参数检验

置信区间

Z 检验

T 检验

10.2.3 模型比较

有些时候我们需要比较两个模型，利用偏差统计量和假设检验可以做到，但是这种方法只适用于嵌套模型。在 GLM 中，要求两个模型具有相同的指数族分布，以及同样的链接函数，被比较的两个模型只有线性预测器是不同的，一个参数多，一个参数少，换句话说一个使用的特征多，另一个使用的特征少。这种模型比较通常可以用来判断某些特征是否有价值，对模型是否有足够的贡献。显然理论上，两个模型参数不同，对数据的拟合度必然会略有不同，两个模型的偏差统计量的值也必然会有一些差异。通常情况下，参数少的模型偏差会稍大一些。假设两个模型之间偏差的差值为 ΔD ，那么这个 ΔD 能否证明两个模型对数据的拟合能力有本质的差别，还是由于随机性导致？假设检验，又或者叫显著性检验，就是用来回答这个问题的。显著性检验用于说明 ΔD 能否证明两个模型的拟合能力有“显著性”的差异，当然假设检验并不能给出百分百准确的结论，其只能依概率给出结论。

在 GLM 中，检验两个模型拟合能力是否有显著差异的一般性步骤是：

1. 定义模型 M_0 对应着零假设 H_0 ，定义另一个更一般（参数更多）的模型 M_1 对应着备择假设 H_a 。零假设 H_0 表示模型 M_0 和 M_1 拟合度一样好，反之，备择假设 H_a 表示 M_0 比 M_1 拟合度差。
2. 训练模型 M_0 ，然后计算一个拟合优度 (goodness of fit, GOF) 指标统计量 G_0 。同样训练模型 M_1 并计算拟合优度指标 G_1 。
3. 计算两个模型拟合度的差异，通常可以是 $\Delta G = G_1 - G_0$ ，或者是 $\Delta G = G_1/G_0$ 。
4. 使用差值统计量 ΔG 的抽样分布检验接受假设 $G_1 = G_0$ 还是 $G_1 \neq G_0$ 。
5. 如果假设 $G_1 = G_0$ 没有被拒绝，则接受 H_0 。反之，如果假设 $G_1 = G_0$ 被拒绝，则接受备择假设 H_a ， M_1 模型在统计学上显著更优。

现在我们以偏差统计量为例，详细介绍一下比较两个模型的检验过程。首先我们设定零假设代表模型 M_0 ，模型参数是 β_0 ，参数数量为 q 。备择假设代表模型 M_1 ，模型参数是 β_1 ，参数数量为 p ，有 $q < p$ 。

$$\begin{aligned} H_0 : G_0 &= G_1 \text{ 两个模型拟合效果一样} \\ H_1 : G_0 &\neq G_1 \text{ 两个模型拟合效果具有统计学上的显著差异} \end{aligned} \quad (10.2.5)$$

我们用 D_0 表示模型 M_0 的偏差，用符号 D_1 表示模型 M_1 的偏差，两个模型偏差统计量的差值为

$$\begin{aligned} \Delta D &= D_0 - D_1 \\ &= 2[\ell(b_s; y) - \ell(b_0; y)] - 2[\ell(b_s; y) - \ell(b_1; y)] \\ &= 2[\ell(b_1; y) - \ell(b_0; y)] \end{aligned} \quad (10.2.6)$$

我们发现 ΔD 的计算方法和 D 的计算方法是一致的，都是两个模型对数似然值的差。根据节 10.1.3 的理论，如果两个模型拟合效果接近，则 ΔD 就渐近服从自由度为 $q - p$ 中心卡方分布

$$\Delta D \sim \chi^2(p - q) \quad (10.2.7)$$

此时 ΔD 的期望值是 $p - q$ 。然而, 如果两个模型拟合效果相差较大, 则 ΔD 漐近服从非中心卡方分布

$$\Delta D \sim \chi^2(p - q, v) \quad (10.2.8)$$

此时 ΔD 的期望值是 $p - q + v$, 将会明显大于 $p - q$, 这个结论将用于对 H_0 进行显著性检验。

根据假设检验的过程, 我们计算出 ΔD 的值, 然后看这个值是否落在分布 $\chi^2(p - q)$ 的拒绝域 (比如是否落在图形两端 $100 * \alpha\%$ 的区域内)。如果落在拒绝域内, 则拒绝 H_0 假设, 接受 H_1 假设。

重要: 通常如果两个模型拟合能力相差巨大, ΔD 直观上就很大了, 此时也没有进行假设检验的必要了。当两个模型的拟合能力比较接近, 从经验上 (直观上) 无法判断 ΔD 是否显著时, 才有假设检验的必要。此外, 相比于直接使用偏差 D 做检验, 使用统计量 ΔD 进行假设检验更好一些。因为 ΔD 通常比单独的偏差 D 更加接近中心卡方分布。这是因为计算 D 的两个模型 (饱和模型和拟合模型) 的拟合能力差别更大, 实际中 D 更接近非中心卡方分布。

然而要使用统计量 ΔD 作为检验统计量, 这就需要能计算出 ΔD 的值。前文我们讲过, 部分指数族分布存在分散参数 ϕ , 比如高斯分布, 对于这些模型必须知道分散参数 ϕ 值才能计算出真实的偏差值 D , 进而计算出统计量 ΔD 。虽然我们可以通过一些前提假设解决这个问题, 比如传统线性回归 (高斯) 模型假设 $a(\phi) = \sigma^2 = 1$, 但这样做会增大 D 的计算误差, 很可能导致得出错误的检验结论。针对这个问题, 我们可以采用另一个检验统计量, F 检验统计量, 又称 F 检验。

在 GLM 中, 我们把分散参数 ϕ 看做是冗余参数, 冗余参数的意思是其不再 GLM 最大似然参数估计的范畴内, 在进行最大似然估计时认为其值是已知。这就需求通弄其它方法来确定冗余参数的值, 一般是根据经验进行假设, 也可以单独从数据中估计。回顾下 GLM 模型一般形式的定义, 在定义中, 分散参数 ϕ 与线性预测器 $\eta = \beta^T x$ 是独立无关的, 换句话说, 两个嵌套模型, 拥有同样的 ϕ , 再结合偏差和尺度化偏差的关系, 可得

$$\begin{aligned} D_0 &= \frac{d_0}{\phi} \\ D_1 &= \frac{d_1}{\phi} \\ \Delta D &= D_0 - D_1 = \frac{d_0 - d_1}{\phi} \end{aligned} \quad (10.2.9)$$

现在回顾下三大抽样分布中的 F 分布, 根据 F 分布的定义, 以下统计量服从 F 分布。

$$F = \frac{\Delta D}{p - q} \Big/ \frac{D_1}{N - p} = \frac{d_0 - d_1}{p - q} \Big/ \frac{d_1}{N - p} \sim F(p - q, N - P) \quad (10.2.10)$$

F 检验统计量可以消除分散参数 ϕ 的影响。利用 F 统计量检验过程和 ΔD 是一样的, 如果 H_0 成立, 两个模型拟合效果接近, 则 F 统计量渐近服从中心分布 $F(p - q, N - P)$ 。计算出 F 的值, 检验其是否落在拒绝域内。

重要: 按照 F 分布的定义, 两个独立的 **中心卡方**随机变量各自除以自由度后, 再相除得到 **中心 F 分布**。一个 **非中心卡方**随机变量除以一个 **中心卡方**随机变量得到 **非中心 F 分布**。这里都要求第二个卡方变量必须是 **中心卡方变量**, 所以要应用 F 检验统计量前提是模型 M_1 是一个“好的”模型, 其偏差统计量 D_1 是一个中心卡方分布。

10.2.4 正态性检验

10.3 案例

10.3.1 线性回归

10.3.2 GLM

10.4 笔记

在统计学中，我们需要通过数据的表现去证明假设的成立与否。如果原假设是成立的，那么其一定会影响到数据的表现，也就是数据一定会受到原假设的影响。因此最直接的方法就是，找到一个和原假设相关的数据统计量 (statistic)，通过这个统计量的值去验证原假设是否成立。

然而，在统计的世界中，通常我们只能得到一些采样数据，背后隐藏的“真理”是不可知的，此时只能通过局部采样数据去“猜测”背后的“真理”。通常我们会用一个随机变量去表示背后的“真理”，采样数据就是这个随机变量的观测样本 (observations)。样本的统计量 (statistic) 是随机变量样本的函数，不同的观测样本得到不同的统计量值，因此样本统计量也是一个随机变量，统计量的概率分布是受到样本所属概率分布的影响的。

我们需要根据样本统计量的值去验证原假设是否成立，但是统计量也是一个随机变量，它的值也就是随机值，通常只有统计量取得的某些值时才能证明原假设的成立，取得其它值时原假设就是不成立的。既然统计量是个随机变量，我们就需要用概率去描述它，

我们的样本的数据是随机变量的采样值，样本的统计量作为样本的函数也是一个随机量，

假设一个事件为真，作为零假设，它的相反事件

通常这个统计量抽样分布是 (近似、渐近) 正态分布或者卡方分布

score statistic 服从卡方分布。在 GLM 中，标准化的 score 统计量是服从正态分布的，其平方是服从卡方分布的似然估计量是服从正态分布的

偏差是服从卡方分布的

Wald statistic 统计量服从卡方分布

GLM 中假设检验的方法

1. score statistic
2. Wald statistic
3. 偏差统计量 (compare the goodness of fit of two models.)

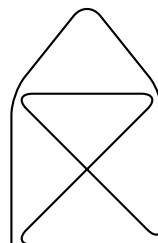


图 10.4.1: An Example Directive with Caption

An example role A  B

高斯模型

1800 年代，约翰·卡尔·弗里德里希·高斯 (Johann Carl Friedrich Gauss) 提出了最小二乘拟合方法和以他的名字命名的分布，高斯分布。高斯分布的概率密度函数具有对称的钟形形状，通常又被称为正态分布，高斯分布是统计学中应用最广泛的概率分布之一。传统的线性回归模型就是建立在高斯分布假设的基础上，传统线下回归模型也被称普通最小二乘 (ordinary least-squares, OLS) 模型。在最初时，GLM 的理论被认为是对普通最小二乘 (OLS) 模型的扩展，因此我们先讨论 OLS 如何适合 GLM 框架。

11.1 传统线性回归

高斯分布是实数域的连续分布，其输入域是整个实数域 $R = (-\infty, +\infty)$ 。用均值 μ 进行参数化的高斯概率密度函数可以表示为：

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\} \quad (11.1.1)$$

其中 $f(\cdot)$ 表示在给定参数 μ 和 σ^2 时，变量 y 的概率密度函数的一般形式。 y 表示输出变量， μ 表示均值参数， σ^2 表示尺度参数 (scale parameter)。

基于高斯分布的回归模型通常称为普通最小二乘 (ordinary least-squares, OLS) 模型，该回归模型通常是统计学模型的入门模型。在传统线性回归模型中，我们定义响应变量和特征变量之间的关系是：

$$y = \beta^T x + \epsilon \quad (11.1.2)$$

其中 $\beta^T x$ 是输入变量的线性预测器， ϵ 一个误差项，并且模型假设这个误差项服从均值为 0 的高斯分布， $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 。因此，输出变量 y 是一个服从高斯分布的变量 $y \sim \mathcal{N}(\beta^T x, \sigma^2)$ 。然后模型利用最大似然估计，估计出模型的参数，现在我们用 GLM 框架来解释 OLS。

11.2 高斯分布

11.3 高斯回归模型

当响应变量 Y 数值范围是实数值时, 并且其数据分布是近似高斯分布时, 就可以为响应变量 Y 建立高斯分布假设, 本节我们讨论高斯模型在 GLM 框架下的解释。我们首先把高斯分布的概率密度函数转化成指数族的形式。

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \\ &= \exp\left\{-\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \end{aligned} \quad (11.3.1)$$

和公式 (7.1.3) 对比下, 可以直接得到:

$$\begin{aligned} \theta &= \mu \\ b(\theta) &= \mu^2/2 \\ a(\phi) &= \sigma^2 \end{aligned} \quad (11.3.2)$$

θ 是自然参数, $b(\theta)$ 是累积函数 (cumulant function), $a(\phi)$ 是分散函数 (dispersion function)。可以看到对于高斯分布, 自然参数 θ 和期望值 μ 之间是恒等函数的关系。此外, 分布的期望值 μ 和方差函数 $\nu(\mu)$ 是可以通过累积函数的导数求得的。期望 μ 可以通过累积函数 $b(\theta)$ 的一阶导数求得:

$$\begin{aligned} b'(\theta) &= \frac{\partial b}{\partial \theta} \\ &= \frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \\ &= (\mu)(1) = \mu \end{aligned} \quad (11.3.3)$$

方差函数 $\nu(\mu)$ 可以通过累积函数 $b(\theta)$ 的二阶导数得到。

$$\begin{aligned} b''(\theta) &= \frac{\partial^2 b}{\partial \theta^2} \\ &= \frac{\partial}{\partial \theta} \left(\frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \right) \\ &= \frac{\partial}{\partial \theta} (\mu) \\ &= \frac{\partial}{\partial \mu} \mu \frac{\partial \mu}{\partial \theta} \\ &= (1)(1) = 1 = \nu(\mu) \end{aligned} \quad (11.3.4)$$

分布方差可以通过分散函数 $a(\phi)$ 和方差函数 $\nu(\mu)$ 的乘积求得:

$$\text{Var}(Y) = a(\phi)\nu(\mu) = \sigma^2 \quad (11.3.5)$$

可以看到, 对于高斯分布, 其方差是与期望无关的, 只和尺度参数相关。此外, 通过公式 (11.3.2) 可以看到, 自然参数 θ 和期望值 μ 之间是恒等函数的关系, 因此高斯分布的标准链接函数就是恒等函数。对于高斯分布来说, 标准链接函数 (canonical link) g 是恒等函数, 即 $\eta = \mu$ 。

$$\eta = g(\mu) = \mu = \theta \quad (11.3.6)$$

由于响应函数 r 是链接函数 g 的反函数, 所以响应函数 r 也是恒等函数。因此, 高斯模型的预测值为:

$$\hat{y} = \mathbb{E}[Y] = \mu = r(\eta) = \eta = \beta^T x \quad (11.3.7)$$

显然, 这和传统线性回归模型的定义方式是等价的。

11.4 参数估计

对于采用标准链接函数的高斯模型, 满足 $\theta = \mu = \eta = \beta^T x$, 此时模型的概率函数表达式是简单的, 对数似然函数的求导也比较方便, 可以直接利用梯度下降法进行求解。在 GLM 框架, 我们有一个适用于 GLM 框架下所有模型的参数估计算法, 迭代重加权最小二乘法 (Iteratively Reweighted Least Square, IRLS)。

采用何种参数估计算法, 直接影响着模型的概率函数的参数化方法。比如, 如果采用 IRLS 算法, 就用均值 μ 参数化模型; 如果采用牛顿法 N-R, 就用线性预测器 $\beta^T x$ 参数化模型。

11.4.1 似然函数

根据公式 (11.3.1), 可以方便的写出高斯模型的对数似然函数

$$\begin{aligned} \ell(\mu, \sigma^2; y) &= \sum_{i=1}^N \left\{ \frac{y_i \mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \mu_i)^2 - \frac{N}{2} \ln(2\pi\sigma^2) \end{aligned} \quad (11.4.1)$$

得分统计量 (score statistic) 是对数似然函数的一阶偏导数,

$$U = \frac{\partial \ell}{\partial \mu_i} = \frac{1}{\sigma^2} \sum_{i=1}^N (Y_i - \mu_i) \quad (11.4.2)$$

上述对数似然函数中是以 μ 参数化的模型, 然而采用梯度下降法进行参数求解时, 需要求数参数 β 的导数, 并使用 β 导数作为每次更新迭代的“梯度”, 所以我们需要用 $\beta^T x$ 对概率分布函数进行参数化。均值参数 μ 和线性预测器 $\eta = \beta^T x$ 之间可以通过响应函数 (也可以叫做激活函数, 链接函数的反函数) 进行连接, 我们用符号 r 表示响应函数。高斯模型的标准链接函数是恒等函数, 因此其标准响应函数也是恒等函数。

$$\mu = r(\eta) = r(\beta^T x) = \beta^T x \quad (11.4.3)$$

现在我们用 β 重新参数化对数似然函数。

$$\begin{aligned} \ell(\beta, \sigma^2; y) &= -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \mu_i)^2 - \frac{N}{2} \ln(2\pi\sigma^2) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \beta^T x_i)^2 - \frac{N}{2} \ln(2\pi\sigma^2) \end{aligned} \quad (11.4.4)$$

对于采用标准链接函数的高斯模型的对数似然函数的一阶偏导数为:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^N \frac{1}{\sigma^2} (y_i - \beta^T x_i) x_{jn} \\ \frac{\partial \ell}{\partial \sigma} &= \sum_{i=1}^N \frac{1}{\sigma} \left\{ \left(\frac{y_i - \beta^T x_i}{\sigma} \right)^2 - 1 \right\} \end{aligned} \quad (11.4.5)$$

二阶偏导数为：

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} &= - \sum_{i=1}^N \frac{1}{\sigma^2} x_{jn} x_{kn} \\ \frac{\partial^2 \ell}{\partial \beta_j \partial \sigma} &= - \sum_{i=1}^N \frac{2}{\sigma^3} (y_i - \beta^T x_i) x_{jn} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \sigma} &= - \sum_{i=1}^N \frac{1}{\sigma^2} \left\{ 3 \left(\frac{y_i - \beta^T x_i}{\sigma} \right)^2 - 1 \right\}\end{aligned}\quad (11.4.6)$$

有了参数的偏导数，就可以利用梯度下降法或者牛顿法迭代求解。这里我们虽然同时给出了 σ 的偏导计算公式，但实际上在进行最大似然估计的过程中，通常是假设 σ 是已知的，标准的最大似然估计算法是无法同时估计 β 和 σ 的。

11.4.2 IRLS

梯度下降法或者牛顿法，需要把对数似然函数用 β 进行参数化表示，并且求解参数的偏导数，在更换链接函数或者指数族分布时，都需要重新求一遍，比较麻烦。而 IRLS 算法提供了一种适用于 GLM 中全部模型和链接函数的统一的参数迭代求解法，使用起来比较简单。现在我们讨论下如何利用 IRLS 算法求解高斯模型的参数。首先回顾一下 IRLS 算法的过程，然后再给出采用标准链接函数的高斯模型的 IRLS 算法过程。IRLS 算法参数更新公式为

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)} \quad (11.4.7)$$

其中 X 是样本特征矩阵， W 是权重矩阵，它是一个对角矩阵，计算方法如下：

$$W^{(t)} = \text{diag} \left\{ \frac{1}{a(\phi)\nu(\mu)(g')^2} \right\}_{(N \times N)} \quad (11.4.8)$$

其中， $a(\phi)$ 是分散函数， $\nu(\mu)$ 是方差函数， $\frac{\partial \mu}{\partial \eta}$ 是响应函数 r 的导数，等价于链接函数 g 的导数的倒数（反函数的导数是原函数导数的倒数）。公式 (11.4.7) 中 Z 的计算方法如下：

$$Z^{(t)} = \left\{ (y - \hat{\mu})g' + \eta^{(t)} \right\}_{(N \times 1)} \quad (11.4.9)$$

$\frac{\partial \eta}{\partial \mu}$ 是链接函数 g 对 μ 的导数。在 IRLS 算法的过程中，不需要把模型用 β 进行参数化，只需要根据具体的分布和链接函数计算出 W 和 Z 即可。

在采用标准链接函数的高斯模型中，连接函数 g 和响应函数 r 都是恒等函数，链接函数的导数是常数值。

$$g' = g'(\mu) = \mu' = 1 \quad (11.4.10)$$

此外，标准链接的高斯模型的方差函数也是常量， $\nu(\mu) = 1, a(\phi) = \sigma^2$ 。所以在采用标准链接，以及方差为常量 1 的假设下，高斯模型 W 和 Z 可以通过下式计算。

$$\begin{aligned}W^{(t)} &= \text{diag} \left\{ \frac{1}{a(\phi)\nu(\mu)(g')^2} \right\}_{(N \times N)} = 1 \\ Z^{(t)} &= \left\{ (y - \hat{\mu})g' + \eta^{(t)} \right\}_{(n \times 1)} = y\end{aligned}\quad (11.4.11)$$

参数 β 的更新过程也就简化成：

$$\beta^{(t+1)} = (X^T X)^{-1} X^T y \quad (11.4.12)$$

我们看到这和最小二乘法的结果是一致的。

11.4.3 拟合优度

偏差统计量

在传统线下回归模型中, 是通过普通最小二乘法 (OLS) 定义出模型的损失函数, 并极小化损失进行参数估计的。最小二乘定义的损失称为残差平方和 (Residual Sum of Squares, RSS), 也叫平方残差和 (sum of squared residuals, SSR), 又叫平方损失和 (sum of squared estimate of errors, SSE)。然而在 GLM 中, 我们通过最大化对数似然或者最小化偏差来估计模型的参数, 最大化对数似然和最小化偏差二者其实等价的。

高斯分布的概率分布函数为

$$f(y; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\} \quad (11.4.13)$$

我们知道似然函数就是所有样本的联合概率, 模型对样本的预测值的对数似然函数称之为观测对数似然函数 (observed log-likelihood function), 样本的真实值的对数似然函数称为饱和对数似然函数 (saturated log-likelihood function)。我们首先写出样本的预测值的观测对数似然函数。

$$\ell(\hat{\mu}, \sigma^2; y) = \sum_{i=1}^N \left\{ \frac{y_i \hat{\mu}_i - \hat{\mu}_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \quad (11.4.14)$$

其中 y_i 是第 i 条样本的真实值, $\hat{\mu}_i$ 是模型的期望参数, 也是模型的拟合值 (预测值)。 σ^2 是尺度参数, 这里我们假设所有样本是共享尺度参数的, 并且暂时认为是常量值, 因此, 上式的似然函数可以继续简化。

$$\begin{aligned} \ell(\hat{\mu}, \sigma^2; y) &= \sum_{i=1}^N \left\{ \frac{y_i \hat{\mu}_i - \hat{\mu}_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} \right\} - \frac{N}{2} \ln(2\pi\sigma^2) \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^N \{2y_i \hat{\mu}_i - \hat{\mu}_i^2 - y_i^2\} - \frac{N}{2} \ln(2\pi\sigma^2) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \hat{\mu}_i)^2 - \frac{N}{2} \ln(2\pi\sigma^2) \end{aligned} \quad (11.4.15)$$

我们发现上式就是 OLS 定义的平方损失的负数, 显然殊途同归, 在 σ^2 是常量的假设下, 最小化平方损失与最大化对数似然等价的。

现在我们来看下饱和模型的对数似然函数 (saturated log likelihood), 在饱和模型中, μ_i 不再是模型的预测值, 而是样本的真实值, 因此对数似然函数是如下的形式。

$$\begin{aligned} \ell(y, \sigma^2; y) &= \sum_{i=1}^N \left\{ \frac{y_i^2 - y_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right\} \\ &= \sum_{i=1}^N \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) \right\} \end{aligned} \quad (11.4.16)$$

尺度偏差的计算方法是饱和对数似然和观测对数似然的差值的 2 倍, 因此高斯模型的偏差 D 为:

$$\begin{aligned}
 D &= 2\{\ell(y, \sigma^2; y) - \ell(\hat{\mu}, \sigma^2; y)\} \\
 &= 2 \sum_{i=1}^N \left\{ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{y_i \hat{\mu}_i - \hat{\mu}_i^2/2}{\sigma^2} + \frac{y_i^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) \right\} \\
 &= 2 \sum_{i=1}^N \left\{ \frac{y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2}{2\sigma^2} \right\} \\
 &= \frac{2}{2\sigma^2} \sum_{i=1}^N \{y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2\} \\
 &= \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - \hat{\mu}_i)^2
 \end{aligned} \tag{11.4.17}$$

我们发现, 在假设 $\sigma^2 = 1$ 的条件下, 高斯模型偏差 $D = \sum_{i=1}^N (y_i - \hat{\mu}_i)^2$ 就是平方和损失, 在高斯模型中, 最小化平方和损失与最小化偏差 D 是等价的, 最小化偏差 D 又和最大化似然是等价的。这里三者的计算公式中用的 μ , 所以与链接函数无关的, 因此任意链接函数都是成立的。注意, 这不是所有 GLM 模型都有的特性, 仅是高斯模型独有的属性。在高斯模型中, 最小二乘法定义的平方损失是模型偏差的一个特例, 偏差是 RSS 在 GLM 中的扩展。

我们用符号 $\hat{\beta}$ 表示参数向量 β 的估计值, 偏差统计量 D 就可以写成

$$\begin{aligned}
 D &= \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i \hat{\beta})^2 \\
 &= \frac{1}{\sigma^2} (y - X \hat{\beta})^T (y - X \hat{\beta})
 \end{aligned} \tag{11.4.18}$$

符号 X 表示输入数据矩阵, 也叫设计矩阵 (design matrix), 其中 $y - X \hat{\beta}$ 可以进行改写, 用 $\hat{\beta}$ 的解析解 $\hat{\beta} = (X^T X)^{-1} X^T y$ 替代。

$$\begin{aligned}
 y - X \hat{\beta} &= y - X(X^T X)^{-1} X^T y \\
 &= [I - X(X^T X)^{-1} X^T] y \\
 &= [I - H] y
 \end{aligned} \tag{11.4.19}$$

其中定义矩阵 $H = X(X^T X)^{-1} X^T$, 矩阵 H 被称为 帽子矩阵 (hat matrix)。因此偏差 D 中的二次项就可以写成

$$(y - X \hat{\beta})^T (y - X \hat{\beta}) = \{[I - H] y\}^T [I - H] y = y^T [I - H] y \tag{11.4.20}$$

帽子矩阵 H 是幂等的 ($H = H^T, HH = H$), 利用帽子矩阵可以推导出偏差统计量 D 是精确服从卡方分布的, 而不是渐近服从, 这里不介绍推导过程了, 有兴趣的可以查看 Graybill 1976。

偏差统计量公式 (11.4.17) 移项可得

$$\sigma^2 D = \sum (y_i - \hat{\mu}_i)^2 \tag{11.4.21}$$

如果模型对数据拟合的比较好, 则偏差统计量 D 服从中心卡方分布 $\chi^2(N - p)$, 其期望值是 $N - p$, 我们用 D 的期望值 $N - p$ 代替 D 就可以得到方差 σ^2 的一个估计。

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{N - p} \tag{11.4.22}$$

在之前的章节中我们多次提到可以用样本的方差作为总体方差的一估计值

$$\hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{N - 1} \tag{11.4.23}$$

其中 \bar{y} 表示样本的平均值, 公式 (11.4.23) 其实是公式 (11.4.22) 的一个特例。当模型只有一个参数时 (线性预测器 $\eta = \beta_0$ 只有截距参数), 也就是空模型, 此时模型对所有样本的输出值都是一样的, 并且就等于训练样本的平均值 \bar{y} , 相当于

$$\forall i, \exists \hat{\mu}_i = \bar{y} \quad (11.4.24)$$

当模型参数数量 $p = 1$ 时, 公式 (11.4.22) 就变成了公式 (11.4.23)。

皮尔逊卡方统计量

高斯模型的方差统计量是常量, $\nu(\mu) = 1$, 根据皮尔逊卡方统计量的定义, 高斯模型的皮尔逊卡方统计量为

$$\chi^2 = \sum_{i=1}^N (y_i - \hat{\mu}_i)^2 \quad (11.4.25)$$

对于高斯模型来说, 皮尔逊卡方统计量, 偏差, 平方损失和, 三者其实是等价的。

11.5 其它链接函数

使用 GLM 作为模型构建框架的重要原因是能够轻松调整模型以适合特定的响应数据情况。标准链接 (canonical-link) 高斯模型假定响应数据为正态分布。尽管正态分布模型对于克服此假设具有一定的鲁棒性, 但仍然有许多数据情况不适合正态分布。不幸的是, 许多研究人员已将标准链接高斯模型用于不符合高斯模型假设的数据上。当然, 许多研究人员很少接受非正态分布模型建模方面的培训。

对数高斯 (log-Gaussian) 模型

线性预测器 $\eta = \beta^T x$ 的取值范围是整个实数域 $R = (-\infty, +\infty)$, 链接函数的主要作用就是将实数域的 η 映射到响应数据的值域。当响应数据仍然服从高斯分布, 但是其范围不再是整个实数域, 而是大于 0 的实数域时, 恒等函数的标准链接 (canonical-link) 函数将不再适用, 这时就需要选取一个合适的链接函数, 将 η 从 $R = (-\infty, +\infty)$ 映射到 $R = (0, +\infty)$ 。log-Gaussian 模型仍然是基于高斯分布, 但它的链接函数不再是标准链接 (canonical-link), 而是对数链接函数, 对数链接通常用于响应数据是非负值的情况。

$$\begin{aligned} \eta &= \ln \mu \\ \mu &= \exp\{\eta\} \end{aligned} \quad (11.5.1)$$

在提出 GLM 框架之前, 研究人员在遇到响应数据都是正数的情形时, 采用的方法是把响应数据进行对数转化, $y = \ln(y)$, 通过这种转换令数据符合整个实数域上的高斯模型假设, 然后再利用标准高斯模型进行建模。然而事实证明, 这种方法在易用性和效果上都不如采用对数链接函数的对数高斯模型。对数高斯模型的实现非常简单, 只需要把标准高斯模型 (采用恒等函数, 标准链接函数) 的链接函数替换成对数链接函数即可。

$$\begin{aligned} f(y; \beta, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \exp(\beta^T x))^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{(y - \exp(\beta^T x))^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right\} \\ &= \exp\left\{-\frac{y \exp(\beta^T x) - \exp(\beta^T x)^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right\} \end{aligned} \quad (11.5.2)$$

对数高斯模型的对数似然函数 (log-likelihood) 为:

$$\ell(\mu; y, \sigma^2) = \sum_{i=1}^N \left[\frac{y_i \exp(\beta^T x_i) - \{\exp(\beta^T x_i)\}^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right] \quad (11.5.3)$$

对数高斯模型就是在标准高斯模型的基础上, 把链接函数由 $\eta = \mu$ 改成了 $\eta = \ln(\mu)$, 相应的响应函数由 $\mu = \eta$ 变成 $\mu = \exp(\eta)$, 对数链接函数及其响应函数的导数分别是

$$\begin{aligned} g' &= 1/\mu \\ r' &= \exp(\eta) \end{aligned} \tag{11.5.4}$$

标准链接函数的导数为 1, 然而对数链接函数的导数为 $1/\mu$ 。注意对公式 (8.1.11) 相应位置替换即可得到似然函数的梯度。同理, 我们可以在不改变 IRLS 算法流程的情况下, 仅需替换掉链接函数 g , 就能使用 IRLS 算法估计对数高斯模型的参数。

标准或正态线性模型、最小二乘线性回归模型, 使用恒等链接函数 (identity link), 这是高斯分布模型的标准形式, 它对于大多数较小的违反高斯分布假设的数据上都具有较强的鲁棒性。但是实际上, 它用于对二值、比例和离散计数数据进行建模时, 并不适用。

从理论上讲, 没有理由将高斯族模型的链接函数限制为恒等链接函数和对数链接函数。倒数链接函数, $1/\mu$, 已经被用于对比例数据进行建模。此外, 幂链接函数也可能与高斯模型一起使用。

逆高斯模型

逆高斯 (inverse Gaussian) 模型是所有传统 GLM 中最不常用的模型，虽然在 GLM 家族谱中总能看到逆高斯模型，但是实际当中却很少使用和讨论。尽管如此，本书还是单独列出一张讨论逆高斯模型，帮助读者学习和研究。

12.1 逆高斯分布

在统计学中，逆高斯分布 (inverse Gaussian distribution)，又叫 Wald distribution，是拥有两个参数的连续值分布，其支持域是 $(0, +\infty)$ 。通常其概率密度函数写成：

$$f(y; \mu, \lambda) = \left(\frac{\lambda}{2\pi y^3} \right)^{1/2} \exp\left\{ -\frac{\lambda(y - \mu)^2}{2\mu^2 y} \right\} \quad (12.1.1)$$

其中 $\mu > 0$ 是分布的均值参数， $\lambda > 0$ 是分布的形状参数 (shape parameter)。当 $\lambda \rightarrow \infty$ 时，逆高斯分布就接近正态分布。逆高斯分布具有多个与高斯分布相似的属性。

为了直观的了解到逆高斯分布的形状和特点，我们看下在不同参数值情况下，逆高斯分布图形的差异，首先我们假设 $\mu = 5.0, \lambda = 2.0$ 。

我们看到随着 μ 的增大，

现在我们固定 $\mu = 1.0$ ，观察下不同的 λ 值图形的差异

尽管分析师在对数据建模时很少使用此逆高斯模型，但有时它比其他连续模型更适合数据。

它特别适合于拟合正值连续数据，这些数据包含低值数据且右偏较长。与 Poisson 分布混合以创建稍后讨论的 Poisson 逆高斯混合模型时，此功能也将非常有用。see section 14.11.

为了说明未经调整的逆高斯密度函数的形状，我们创建了一组简单的 Stata 命令，以针对指定的均值和标度参数生成概率密度函数的值。各种参数值的概率密度函数图显示了灵活性。

高斯分布的两个参数是 μ 和 σ^2 ，而上面给出逆高斯分布的参数是 μ 和 λ 。实际上，逆高斯分布也可以用 σ^2 表示形状参数，二者的是倒数的关系， $\lambda = 1/\sigma^2$

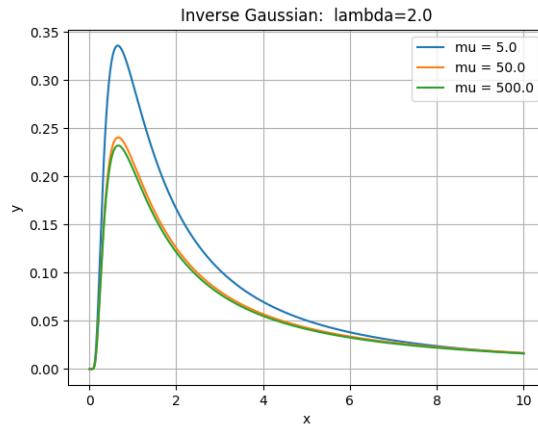


图 12.1.1: $\lambda = 2.0$ 的逆高斯分布

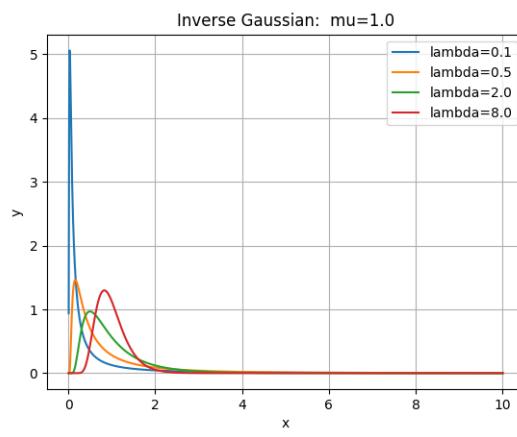


图 12.1.2: $\mu = 1.0$ 的逆高斯分布

12.2 逆高斯回归模型

。在 GLM 中, 用 σ^2 会更方便一些, 所以这里用 σ^2 重新参数化逆高斯分布的概率密度函数。

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi y^3 \sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}\right\} \quad (12.2.1)$$

现在把上式转化成指数族的形式。

$$\begin{aligned} f(y; \mu, \sigma^2) &= \exp\left\{-\frac{(y-\mu)^2}{2y(\mu\sigma)^2} - \frac{1}{2} \ln(2\pi y^3 \sigma^2)\right\} \\ &= \exp\left\{\frac{y/(2\mu^2) - 1/\mu}{-\sigma^2} - \frac{1}{2y\sigma^2} - \frac{1}{2} \ln(2\pi y^3 \sigma^2)\right\} \end{aligned} \quad (12.2.2)$$

和 GLM 中指数族的标准形式对比下, 不难得到各个组件的内容。

$$\begin{aligned} \theta &= \frac{1}{2\mu^2} \\ b(\theta) &= \frac{1}{\mu} \\ a(\phi) &= -\sigma^2 \end{aligned} \quad (12.2.3)$$

现在来看下逆高斯分布的期望和方差。

$$\begin{aligned} b'(\theta) &= \frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \\ &= \left(\frac{-1}{\mu^2}\right) (-\mu^3) = \mu \\ b''(\theta) &= \frac{\partial^2 b}{\partial \mu^2} \left(\frac{\partial \mu}{\partial \theta}\right)^2 + \frac{\partial b}{\partial \mu} \frac{\partial^2 \mu}{\partial \theta^2} \\ &= \left(\frac{2}{\mu^3}\right) (\mu^6) + \left(\frac{-1}{\mu^2}\right) (3\mu^5) \\ &= 2\mu^3 - 3\mu^3 \\ &= -\mu^3 \end{aligned} \quad (12.2.4)$$

逆高斯分布的方差为:

$$Var(Y) = a(\phi)b''(\theta) = -\sigma^2(-\mu^3) = \sigma^2\mu^3 \quad (12.2.5)$$

显然逆高斯分布的方差是和其期望相关的。

根据标准链接函数的定义, 逆高斯分布的标准链接函数为:

$$\eta = g(\mu) = \frac{1}{2\mu^2} \quad (12.2.6)$$

链接函数的导数为:

$$g'(\mu) = -\mu^{-3} \quad (12.2.7)$$

响应函数 $r(\eta)$ 为链接函数的反函数。

$$\mu = r(\eta) = g^{-1}(\eta) = \frac{1}{\sqrt{2\eta}} \quad (12.2.8)$$

总结一下逆高斯模型的关键部分。

$$\begin{aligned}
 \text{标准链接函数: } \eta &= g(\mu) = \frac{1}{2\mu^2} \\
 \text{反链接(响应)函数: } \mu &= r(\eta) = \frac{1}{\sqrt{2\eta}} \\
 \text{方差函数: } \nu &= -\mu^3 \\
 \text{分散函数: } a(\phi) &= -\sigma^2 \\
 \text{链接函数导数: } g' &= -\mu^{-3}
 \end{aligned} \tag{12.2.9}$$

12.3 参数估计

12.3.1 似然函数

逆高斯分布的指数形式去掉底数就得到了对数似然函数。

$$\ell = \sum_{i=1}^N \left\{ \frac{y_i/(2\mu_i^2) - 1/\mu_i}{-\sigma^2} - \frac{1}{2y_i\sigma^2} - \frac{1}{2} \ln(2\pi y_i^3 \sigma^2) \right\} \tag{12.3.1}$$

根据公式 (8.1.12)，标准链接函数的 Gamma 模型的似然函数的一阶偏导为

$$\begin{aligned}
 U_j &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)g(\mu_i)'} x_{ij} \\
 &= -\sum_{i=1}^N \frac{y_i - \mu_i}{\sigma^2} x_{ij}
 \end{aligned} \tag{12.3.2}$$

我们发现逆高斯模型和高斯模型的得分统计量只差了一个负号。

12.3.2 IRLS

逆高斯模型的 W 和 Z 分别为

$$\begin{aligned}
 W &= \text{diag} \left\{ \frac{1}{a(\phi)\nu(\hat{\mu})(g')^2} \right\}_{(N \times N)} \\
 &= \text{diag} \left\{ \frac{\hat{\mu}^3}{\sigma^2} \right\}_{(N \times N)}
 \end{aligned} \tag{12.3.3}$$

$$\begin{aligned}
 Z &= \{(y - \hat{\mu})g' + \eta\}_{(N \times 1)} \\
 &= \left\{ \frac{-(y - \hat{\mu})}{\hat{\mu}^3} + \eta \right\}_{(N \times 1)}
 \end{aligned} \tag{12.3.4}$$

12.3.3 拟合优度

逆高斯模型的饱和模型的对数似然函数为

$$\ell(y, \sigma^2; y) = \sum_{i=1}^n \left\{ -\frac{1}{2} \ln (2\pi y_i^3 \sigma^2) \right\} \quad (12.3.5)$$

逆高斯模型的偏差统计量为

$$\begin{aligned} D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\ &= 2 \sum_{i=1}^N \left\{ -\frac{1}{2} \ln (2\pi y_i^3 \sigma^2) \right\} - 2 \sum_{i=1}^N \left\{ \frac{y_i/(2\hat{\mu}_i^2) - 1/\hat{\mu}_i}{-\sigma^2} - \frac{1}{2y_i \sigma^2} - \frac{1}{2} \ln (2\pi y_i^3 \sigma^2) \right\} \\ &= 2 \sum_{i=1}^N \left\{ \frac{y_i/(2\hat{\mu}_i^2) - 1/\hat{\mu}_i}{\sigma^2} + \frac{1}{2y_i \sigma^2} \right\} \\ &= \frac{2}{\sigma^2} \sum_{i=1}^N \left\{ y_i/(2\hat{\mu}_i^2) - 1/\hat{\mu}_i + \frac{1}{2y_i} \right\} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^N \left\{ \frac{y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2}{\hat{\mu}_i^2 y_i} \right\} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^N \left\{ \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i} \right\} \end{aligned} \quad (12.3.6)$$

逆高斯模型的皮尔逊卡方统计量为

$$\begin{aligned} \chi^2 &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)} \\ &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{-\hat{\mu}_i^3} \end{aligned} \quad (12.3.7)$$

12.4 其它链接函数

类似于伽玛模型, 除了标准链接函数, 对数 (log) 链接函数和恒等 (identity) 链接函数是也是逆高斯分布经常使用的链接函数。

对于逆高斯, 当对持续时间类型的数据进行建模时, 恒等链接函数是另一个合适的选择。

二项式模型

在机器学习领域，应用最广的两个模型，一个是线性回归模型另一个就是逻辑回归模型。线性回归模型就是采用标准连接函数的高斯模型，高斯模型是处理连续值数据的基本模型。而逻辑回归模型是处理二分类数据的基本模型，逻辑回归模型就是标准连接函数的二项式回归模型。二项式回归模型对应的是指数族中的二项式分布，二项式分布是统计学中最常见的概率分布之一，应用十分广泛。本章我们讨论 GLM 框架下的二项式回归模型。

13.1 伯努利分布

如果一个随机变量只有两种可能状态，就可以认为这个随机变量服从伯努利分布 (Bernoulli distribution)。比如，在广告场景中，用户点击广告的行为可以分成点击和不点击两个状态；投掷一枚硬币，只能是正面向上或者反面向上（排除硬币站立的情况）。服从伯努利分布的随机变量通常称为伯努利变量，伯努利变量只有两个不同的状态，因此它是离散随机变量，伯努利分布属于离散概率分布。

通常会用数字 0 和 1 分别表示伯努利变量的两种状态，假设状态为 1 的概率是 π ，那么状态为 0 的概率就是 $1 - \pi$ 。伯努利概率分布的概率分布函数通常可以写成

$$f(y; \pi) = \pi^y (1 - \pi)^{1-y} \quad (13.1.1)$$

$Y = 1$ 的概率和 $Y = 0$ 的概率分别为

$$\begin{aligned} P(Y = 1) &= f(y = 1; \pi) = \pi^1 (1 - \pi)^{1-1} = \pi \\ P(Y = 0) &= f(y = 0; \pi) = \pi^0 (1 - \pi)^{1-0} = 1 - \pi \end{aligned} \quad (13.1.2)$$

伯努利分布的期望和方差分别是

$$\begin{aligned} \mathbb{E}[Y] &= \mu = \pi \\ V(Y) &= \pi(1 - \pi) = \mu(1 - \mu) \end{aligned} \quad (13.1.3)$$

可以看出 π 其实就是分布的期望参数 μ ，并且伯努利分布的方差是期望的一个二次函数。伯努利分布仅有一个期望参数，因此它是一个单参数的概率分布。

伯努利分布是离散变量的概率分布，离散分布的概率分布函数称为 **概率质量函数**，概率质量函数的值直接就是概率值。这一点和连续值分布是不同的，连续值分布的概率分布函数叫做 **概率密度函数**，概率密度函数的值并不是概率值，需要积分才能得到概率值。

13.2 逻辑回归模型

13.2.1 模型定义

假设响应变量 Y 是一个伯努利变量, 它的概率分布函数如公式 (13.1.1) 所示, 现在把它转化成指数族的形式

$$f(y; \pi) = \exp \left\{ y \ln \left(\frac{\pi}{1 - \pi} \right) + \ln(1 - \pi) \right\} \quad (13.2.1)$$

由于参数 π 就是分布的期望 μ , 因此我们直接用符号 μ 替换 π 。

$$f(y; \mu) = \exp \left\{ y \ln \left(\frac{\mu}{1 - \mu} \right) + \ln(1 - \mu) \right\} \quad (13.2.2)$$

和指数族的自然形式对比下, 可以直接给出各项的内容。

$$\begin{aligned} \text{自然参数 } \theta &= \ln \left(\frac{\mu}{1 - \mu} \right) \\ \text{累积函数 } b(\theta) &= -\ln(1 - \mu) \\ \text{分散函数 } a(\phi) &= \phi = 1 \end{aligned} \quad (13.2.3)$$

它的期望可以通过累积函数的一阶导数求得

$$\mathbb{E}[Y] = b'(\theta) = \mu \quad (13.2.4)$$

方差函数通过累积函数的二阶导数得到

$$\nu(\mu) = b''(\theta) = \mu(1 - \mu) \quad (13.2.5)$$

分散函数和方差函数的乘积就是分布的方差

$$V(Y) = a(\phi)\nu(\mu) = \mu(1 - \mu) \quad (13.2.6)$$

可以看到伯努利分布的方差不是常量, 而是关于期望参数 μ 的二次函数, 显然伯努利分布的方差会受到期望 μ 的影响。

根据标准连接函数的定义, 标准连接函数是使得线性预测器 η 等于自然参数 θ 的连接函数, 所以对于伯努利分布, 其标准连接函数就是

$$\eta = \theta = g(\mu) = \ln \left(\frac{\mu}{1 - \mu} \right) \quad (13.2.7)$$

在统计学中, 这个函数称为 `logit` (/ləʊdʒɪt/ LOH-jit) 函数, 逻辑回归模型的标准连接函数就是 `logit` 函数, 它的一阶导数为

$$g'(\mu) = \frac{1}{\mu(1 - \mu)} \quad (13.2.8)$$

响应函数是连接函数的反函数, `logit` 函数的反函数为

$$\text{logit}(\mu)^{-1} = \frac{e^\eta}{1 + e^\eta} = \text{logistic}(\eta) \quad (13.2.9)$$

`logit` 函数的反函数就是我们熟知 `logistic` 函数, `logistic` 函数中文叫做 **逻辑函数**, 它是标准连接的伯努利回归模型的响应函数, 因此一般把伯努利回归模型叫做 **逻辑回归模型** (**logistic regression model**)。

$$\text{响应函数 } \hat{y} = \hat{\mu} = r(\eta) = \frac{e^\eta}{1 + e^\eta} \quad (13.2.10)$$

注解: 很多人把 logistic 函数称为 sigmoid, 这是不准确的。sigmoid 定义是: 拥有 S 形状的一类函数。sigmoid 是一类函数的统称, 并不是特指某一个函数, logistic 函数是 sigmoid 中的一例, 其它的还有 Arctangent 函数、Hyperbolic tangent 函数、Gudermannian 函数等等。

最后整理下逻辑回归模型的关键组件

$$\begin{aligned}
 \text{标准连接函数: } \eta &= g(\mu) = \text{logit}(\mu) = \ln \frac{\mu}{1-\mu} = \ln(\mu) - \ln(1-\mu) \\
 \text{响应函数: } \mu &= r(\eta) = \frac{e^\eta}{1+e^\eta} \\
 \text{方差函数: } \nu(\mu) &= \mu(1-\mu) \\
 \text{分散函数: } a(\phi) &= 1 \\
 \text{连接函数导数: } g' &= \frac{1}{\mu(1-\mu)}
 \end{aligned} \tag{13.2.11}$$

13.2.2 参数估计

大部分有关逻辑回归模型的资料中, 都是采用完全最大似然法估计模型的参数, 比如梯度下降法、牛顿法等等。然而逻辑回归模型是可以纳入到 GLM 框架中的, 因此逻辑回归模型也是可以用 IRLS 算法进行的参数估计的。

逻辑模型的对数似然函数为

$$\begin{aligned}
 \ell(\mu; y) &= \sum_{i=1}^N \left\{ y_i \ln \left(\frac{\mu_i}{1-\mu_i} \right) + \ln(1-\mu_i) \right\} \\
 &= \sum_{i=1}^N \{ y_i \ln \mu_i + (1-y_i) \ln(1-\mu_i) \}
 \end{aligned} \tag{13.2.12}$$

IRLS 算法中权重矩阵 W 和工作响应矩阵 Z 的计算公式分别为:

$$\begin{aligned}
 W_{ii} &= \frac{1}{a(\phi)\nu(\hat{\mu}_i)(g'_i)^2} \\
 &= \hat{\mu}_i(1-\hat{\mu}_i)
 \end{aligned} \tag{13.2.13}$$

$$\begin{aligned}
 Z_i &= (y_i - \hat{\mu}_i)g'_i + \eta_i \\
 &= \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i(1-\hat{\mu}_i)} + \eta_i
 \end{aligned} \tag{13.2.14}$$

偏差统计量为:

$$\begin{aligned}
 D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\
 &= 2 \sum_{i=1}^N \left\{ y_i \ln \frac{y_i}{\hat{\mu}_i} + (1-y_i) \ln \left(\frac{1-y_i}{1-\hat{\mu}_i} \right) \right\}
 \end{aligned} \tag{13.2.15}$$

13.2.3 odds 与 logit

在 GLM 中连接函数的作用是把线性预测器 η 和响应变量的期望值 μ 连接在一起，本质上就是把 η 的空间和 μ 的空间进行映射，并且这种映射关系必须是 **双射**，也就是对于任意一个 η 都可以得到一个唯一的 μ 与之对应，反过来也成立，对于任意一个 μ 都可以得到一个唯一的 η 与之对应，这就要求连接函数必须是单调可逆的。在很多有关 logistic 回归模型的资料中都会提到一个概念，`odds`，为了令读者对二项式回归模型理解的更透传，这里我们介绍下 `odds` 与标准连接函数 `logit` 的关系。

学过基础数学技能的人都知道，概率 (probability) 是用来描述事件发生的可能性的。概率一般是通过频次来计算的，比如投掷一枚骰子 n 次，其中点数 1 的次数是 a ，那么点数 1 概率为

$$p(1) = \frac{a}{n} \quad (13.2.16)$$

点数不是 1 的概率为

$$p(-1) = 1 - \frac{a}{n} = \frac{n-a}{n} \quad (13.2.17)$$

用概率来描述事件发生可能性是符合人的直觉的，因此在日常生活中概率的应用是广泛的。然而在统计学中，除了概率以外，还可以用 **几率** (`odds`) 来描述事件发生的可能性。在英语里，`odds` 的意思就是指几率、可能性。`odds` 指的是 **事件发生的概率与不发生的概率之比**。

$$odds = \frac{\text{probability of event}}{\text{probability of no event}} = \frac{p}{1-p} \quad (13.2.18)$$

在上面的例子中，点数为 1 的 `odds` 为

$$odds(1) = \frac{a/n}{(n-a)/n} = \frac{a}{n-a} = \frac{\text{frequency of event}}{\text{frequency of no event}} \quad (13.2.19)$$

可以看到事件总次数 n 是可以被抵消掉的，因此 `odds` 也可以看做是频次之比。由于 `odds` 是概率或者频次的比值，显然 `odds` 的取值范围是 $[0, \infty)$ ，`odds` 的值越大，事件发生的可能性就越大。概率的值域范围是 $[0, 1]$ ，而 `odds` 的值域范围是 $[0, \infty)$ ，从概率到 `odds` 的转变，实现了值域的改变。

如果对 `odds` 取自然对数，就得到了 `logit`

$$logit(odds) = \ln(odds) = \ln \frac{p}{1-p} \quad (13.2.20)$$

`odds` 的自然对数就称为 `logit`，`logit` 是 `log-it` 的简写。`odds` 取自然对数后，输出值的范围就变成了 $(-\infty, \infty)$ ，正好和线性预测器 η 的值域范围变得一致了。从概率到 `logit` 值域范围的演变过程为

$$\text{probability} : [0, 1] \implies \text{odds} : [0, \infty) \implies \text{logit} : (-\infty, \infty) \quad (13.2.21)$$

逻辑回归模型的期望值 μ 就表示一个概率值，其取值范围是 $[0, 1]$ 。线性预测器 η 的取值范围是 $(-\infty, \infty)$ 。`logit` 作为逻辑 (二项式) 回归模型的标准连接函数，其作用就是实现 μ 到 η 的映射。

13.3 二项式分布

在英语语境中，会把随机变量的单次采样称为一次实验，连续多次独立实验的结果形成的序列，称为一次 `trial`。如果把伯努利变量进行多次独立取样，就得到一个伯努利状态序列，如果把这个序列中状态为 1 的次数看做一个随机变量，这个变量就是一个二项式 (binomial) 变量，二项式变量的概率分布称为二项式分布 (binomial distribution)。二项式分布表示进行 n 次伯努利实验，状态 1 的次数的概率分布。假设响应变量 Y 服从二项式分布，则概率分布函数为

$$f(y; n, \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \quad (13.3.1)$$

其中符号 π 表示单次伯努利实验状态为 1 的概率, 也就是伯努利分布的期望参数。符号 n 是进行的实验次数, 通常是已知的常量。可以看出二项式分布的概率分布函数就是在伯努利概率分布函数的基础上加了组合数 $\binom{n}{y}$, 之所以是组合数, 而不是排列数, 是因为二项式随机变量 Y 表示的次数, 与顺序无关, 因此是组合数。

二项式分布的期望和方差分别是

$$\begin{aligned}\mathbb{E}[Y] &= \mu = n\pi \\ V(Y) &= n\pi(1 - \pi)\end{aligned}\tag{13.3.2}$$

通常实验次数 n 是已知的常量, 并不是未知参数, 二项式分布的未知参数只有 π , 这和伯努利分布是同一个参数。事实上, 伯努利分布是二项式分布的一个特例, 二项式分布中, 当 $n = 1$ 时, 就是退化成了伯努利分布。因此, 有些资料中, 把伯努利分布也称作二项式分布, 这是合理的。

13.4 二项式回归模型

13.4.1 模型定义

假设响应变量 Y 服从二项式分布, 其概率分布函数为

$$f(y; n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}\tag{13.4.1}$$

其中 π 是单次实验成功的概率, n 是实验的总次数, y 是成功的次数, 这个分布函数中 π 是唯一的参数。转化成自然指数族的形式为

$$\begin{aligned}f(y; n, \pi) &= \exp \left\{ y \ln(\pi) + n \ln(1 - \pi) - y \ln(1 - \pi) + \ln \binom{n}{y} \right\} \\ &= \exp \left\{ y \ln \left(\frac{\pi}{1 - \pi} \right) + n \ln(1 - \pi) + \ln \binom{n}{y} \right\}\end{aligned}\tag{13.4.2}$$

和指数族的自然形式对比下, 可以直接给出各项的内容。

$$\begin{aligned}\text{自然参数 } \theta &= \ln \left(\frac{\pi}{1 - \pi} \right) \\ \text{累积函数 } b(\theta) &= -n \ln(1 - \pi) \\ \text{分散函数 } a(\phi) &= \phi = 1\end{aligned}\tag{13.4.3}$$

累积函数的一阶导数和二阶导数分别为

$$\begin{aligned}b'(\theta) &= \frac{\partial b}{\partial \pi} \frac{\partial \pi}{\partial \theta} = \frac{n}{1 - \pi} \pi(1 - \pi) = n\pi \\ b''(\theta) &= \frac{\partial^2 b}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \theta} \right)^2 + \frac{\partial b}{\partial \pi} \frac{\partial^2 \pi}{\partial \theta^2} \\ &= \frac{n}{(1 - \pi)^2} (1 - \pi)^2 \pi^2 + \frac{n}{1 - \pi} (1 - \pi) \pi(1 - 2\pi) \\ &= n\pi^2 + n\pi(1 - 2\pi) \\ &= n\pi(1 - \pi)\end{aligned}\tag{13.4.4}$$

通过累积函数的导数可以分别得到分布的期望和方差,

$$\begin{aligned}\mathbb{E}[Y] &= \mu = b'(\theta) = n\pi \\ Var(Y) &= a(\phi)b''(\theta) = a(\phi)\nu(\mu) = n\pi(1 - \pi) = \mu(1 - \frac{\mu}{n})\end{aligned}\tag{13.4.5}$$

同样, 二项式分布的方差是关于期望的一个函数, 方差会受到期望的影响。

二项式分布的自然参数和伯努利分布的自然参数是完全一样, 因此二项式模型的标准连接函数也是 `logit` 函数。

$$\eta = \theta = g(\mu) = \ln \left(\frac{\pi}{1 - \pi} \right) = \ln \left(\frac{\mu}{n - \mu} \right) \quad (13.4.6)$$

连接函数的导数为

$$g'(\mu) = \frac{n}{\mu(n - \mu)} \quad (13.4.7)$$

二项式模型的响应函数同样也是 `logistic` 函数。

$$\hat{y} = \hat{\mu} = r(\eta) = \frac{ne^\eta}{1 + e^\eta} = n\hat{\pi} = n \text{logistic}(\eta) \quad (13.4.8)$$

对比下伯努利回归模型与二项式回归模型, 可以看出, 无论是连接函数还是响应函数, 仅仅只是差了一个常量 n 而已, 而 n 是一个已知的常量, 并且当 $n = 1$ 时, 二者就完全一样了。连接函数都是映射的线性预测器 η 与 π 的关系, 这和 n 无关。因此可以认为 **伯努利模型和二项式模型是同一个模型, 二项式回归模型也是逻辑回归模型**。

虽然严格来说二项式分布的期望是 $\mu = n\pi$, 但是由于 n 是已知常量, 仅仅起到一个倍数的作用, 所以在强调 **期望参数**时, 可以只考虑 π , 而忽略 n 。这一点请铭记, 否则在看某些资料时会迷惑!

最后, 我们汇总下二项式回归 (逻辑回归) 模型的一些关键组件。

$$\begin{aligned} \text{标准连接函数: } \eta &= g(\mu) = \text{logit}(\mu) = \ln \frac{\mu}{n - \mu} = \ln(\mu) - \ln(n - \mu) \\ \text{响应函数: } \mu &= r(\eta) = \frac{ne^\eta}{1 + e^\eta} = n\pi \\ \text{方差函数: } \nu(\mu) &= \mu(1 - \frac{\mu}{n}) \\ \text{分散函数: } a(\phi) &= 1 \\ \text{连接函数导数: } g' &= \frac{n}{\mu(n - \mu)} \end{aligned} \quad (13.4.9)$$

13.4.2 参数估计

二项式回归模型的对数似然函数为

$$\begin{aligned} \ell(\hat{\mu}; y) &= \sum_{i=1}^N \left\{ y_i \ln \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + n_i \ln(1 - \hat{\pi}_i) + \ln \left(\frac{n_i}{y_i} \right) \right\} \\ &= \sum_{i=1}^N \left\{ y_i \ln \left(\frac{\hat{\mu}_i}{n - \hat{\mu}_i} \right) + n_i \ln(n_i - \hat{\mu}_i) - n_i \ln(n_i) + \ln \left(\frac{n_i}{y_i} \right) \right\} \end{aligned} \quad (13.4.10)$$

IRLS 算法中 W 和 Z 的计算公式分别为

$$\begin{aligned} W_{ii} &= \frac{1}{a(\phi)\nu(\hat{\mu}_i)(g'_i)^2} \\ &= \frac{n_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} \end{aligned} \quad (13.4.11)$$

$$\begin{aligned} Z_i &= (y_i - \hat{\mu}_i)g'_i + \eta_i \\ &= \frac{n_i(y_i - \hat{\mu}_i)}{\hat{\mu}_i(n_i - \hat{\mu}_i)} + \eta_i \end{aligned} \quad (13.4.12)$$

偏差统计量为

$$\begin{aligned}
 D &= 2\{\ell(y; y)_f - \ell(\hat{\mu}; y)_m\} \\
 &= 2 \sum_{i=1}^N \left\{ y_i \ln \left(\frac{y_i}{n_i - y_i} \right) + n_i \ln(n_i - y_i) - y_i \ln \left(\frac{\hat{\mu}_i}{n_i - \hat{\mu}_i} \right) - n_i \ln(n_i - \hat{\mu}_i) \right\} \\
 &= 2 \sum_{i=1}^N \{y_i \ln y_i - y_i \ln(n_i - y_i) + n_i \ln(n_i - y_i) - y_i \ln \hat{\mu}_i + y_i \ln(n_i - \hat{\mu}_i) - n_i \ln(n_i - \hat{\mu}_i)\} \\
 &= 2 \sum_{i=1}^N \{[y_i \ln y_i - y_i \ln \hat{\mu}_i] - [y_i \ln(n_i - y_i) - y_i \ln(n_i - \hat{\mu}_i)] + [n_i \ln(n_i - y_i) - n_i \ln(n_i - \hat{\mu}_i)]\} \\
 &= 2 \sum_{i=1}^N \left\{ y_i \ln \frac{y_i}{\hat{\mu}_i} + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right\}
 \end{aligned} \tag{13.4.13}$$

我们用符号 o_{ik} 表示从样本中 **观测 (observed)** 到的 k 个状态的次数, 比如 o_{i0} 表示失败的次数 $n_i - y_i$, o_{i1} 表示成功的次数 y_i 。用符号 e_{ik} 表示模型 **拟合 (fitted)** 的结果, 比如 e_{i0} 表示模型预测的失败的次数 $n_i - \hat{\mu}_i$, e_{i1} 表示模型预测的成功的次数 $\hat{\mu}_i$ 。则偏差统计量可以简写为

$$D = 2 \sum_{i=1}^N \sum_k o_{ik} \ln \frac{o_{ik}}{e_{ik}} \tag{13.4.14}$$

二项式模型的偏差统计量是不包含冗余参数的, 比如分散参数 ϕ , 所以可以直接用它的渐近分布进行假设检验。

$$D \sim \chi^2(N - p) \tag{13.4.15}$$

注意对于 n_i 比较小的数据, 这个近似的效果会比较差。

二项式模型的皮尔逊卡方统计量为

$$\begin{aligned}
 \chi^2 &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1 - \frac{\hat{\mu}_i}{n_i})} \\
 &= \sum_{i=1}^N \frac{(y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i(1 - \hat{\pi}_i)}
 \end{aligned} \tag{13.4.16}$$

二项式模型的卡方统计量 χ^2 和偏差统计量 D 是近似相等, 可以用泰勒级数进行证明。这里省略证明。当 n_i 比较小, 卡方统计量会比偏差更准确一些, 但是当 n_i 非常小时, 无论偏差统计量还是卡方统计量都不在准确。

13.5 其它连接函数

我们已经很清楚连接函数的作用了, 它的作用就是把线性预测器 η 和模型的期望参数 μ 进行可逆的映射。二项式模型的期望参数 $\mu = \pi$ (注意, 这里我们忽略 n) 是一个概率值, 它的合理取值范围是 $[0, 1]$ 。因此, 能用来做二项式模型响应函数 (连接函数的反函数) 的函数, 它的输出域就必须是 $[0, 1]$ 。而在统计学中, 有一类函数是符合这个特点的, 那就是概率分布的 **累积分布函数**。概率分布的累积分布函数, 满足单调性, 并且输出域是 $[0, 1]$, 因此累积分布函数的反函数是可以作为二项式模型的连接函数的。

假设函数 $f(s)$ 是一个概率密度 (质量) 函数, 其累积分布函数 $F(s)$ 就是对 $f(s)$ 的积分。

$$F(s) = \int_{-\infty}^t f(s) ds \tag{13.5.1}$$

其中 $f(s) \geq 0, \int_{-\infty}^{\infty} f(s) ds = 1$, 累计分布函数表示的是随机变量 s 大于等于 t 的概率, $P(s \geq t)$, 显然, $F(s)$ 的输出范围是 $[0, 1]$ 。

13.5.1 恒等连接函数

首先我们看下均匀分布的累积分布函数, 假设概率分布函数 $f(s)$ 是均匀分布的概率密度函数, 随机变量 s 的范围是 $[c_1, c_2]$, 均匀分布 $f(s)$ 的概率密度函数为

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1} & c_1 \leq s \leq c_2 \\ 0 & \text{otherwise} \end{cases} \quad (13.5.2)$$

均匀分布的累积概率分布函数 $F(x)$ 是

$$\begin{aligned} F(x) &= \int_{c_1}^x f(s) ds \\ &= \frac{x - c_1}{c_2 - c_1} \quad \text{for } c_1 \leq x \leq c_2 \end{aligned} \quad (13.5.3)$$

令 $\beta_1 = \frac{-c_1}{c_2 - c_1}$, $\beta_2 = \frac{1}{c_2 - c_1}$, 此时 $F(x)$ 等价于

$$F(x) = \beta_1 + \beta_2 x \quad (13.5.4)$$

$F(x)$ 就是响应变量的期望 μ , 线性部分就是 η , 连接函数 $g(\mu)$ 就是恒等函数, 恒等连接函数就相当于是均匀分布的累积概率分布函数。

但是恒等连接函数有个限制, 就是只有 x 在区间 $[c_1, c_2]$ 才有意义, 此时参数 β 也被约束在某个区间内。然而, GLM 的通用参数估计算法 IRLS 并不能解决带约束的参数估计问题, 因此恒等连接函数在二项式模型中并不常用。

13.5.2 probit 回归

现在我们看下正态分布的累积分布函数。假设概率分布 $f(s)$ 是标准正态分布 $\mathcal{N}(0, 1)$, 其 CDF 又称为累积正态分布函数 (cumulative normal distribution function, CNDF), 累积正态分布函数是对标准正态分布 $\mathcal{N}(0, 1)$ 概率密度函数的积分。习惯上用符号 $\phi(x)$ 表示标准正态分布的概率密度函数, 用符号 Φ 表示累积正态分布函数, 注意不要和指数族的分散参数搞混, 在这里不是尺度参数。累积正态分布函数为

$$\begin{aligned} \Phi(x) &= \int_{-\infty}^x \phi(x) dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned} \quad (13.5.5)$$

它的反函数通常称为 probit 函数, 用符号 $\Phi^{-1}(x)$ 表示。

$$\text{probit} = \Phi^{-1}(x) \quad (13.5.6)$$

采用 probit 函数作为连接函数二项式回归模型称为 probit 回归模型, probit 和 logit 非常的类似, 二者的 (反) 函数图像都是 S 型曲线, 并且以点 $(0, 0.5)$ 为对称点呈现对称结构, 二者只是在曲率上稍微有些差别。

对二值或分组的二项式数据使用 probit 回归模型通常会产生与逻辑回归相似的输出。但是 logit 模型可以解释为胜率比 (odds ratio), 而 probit 没有这样的解释。但是, 如果线性关系中涉及正态性 (通常在生物学领域中就是如此), 则 probit 可能是合适的模型。当研究人员对赔率不感兴趣而对预测或分类感兴趣时, 也可以使用它。然后, 如果 probit 模型的偏差显着低于相应的 logit 模型的偏差, 则首选前者。当然, 比较二项式家族中的任何连接函数时, 偏差小的模型永远是最优的选择。

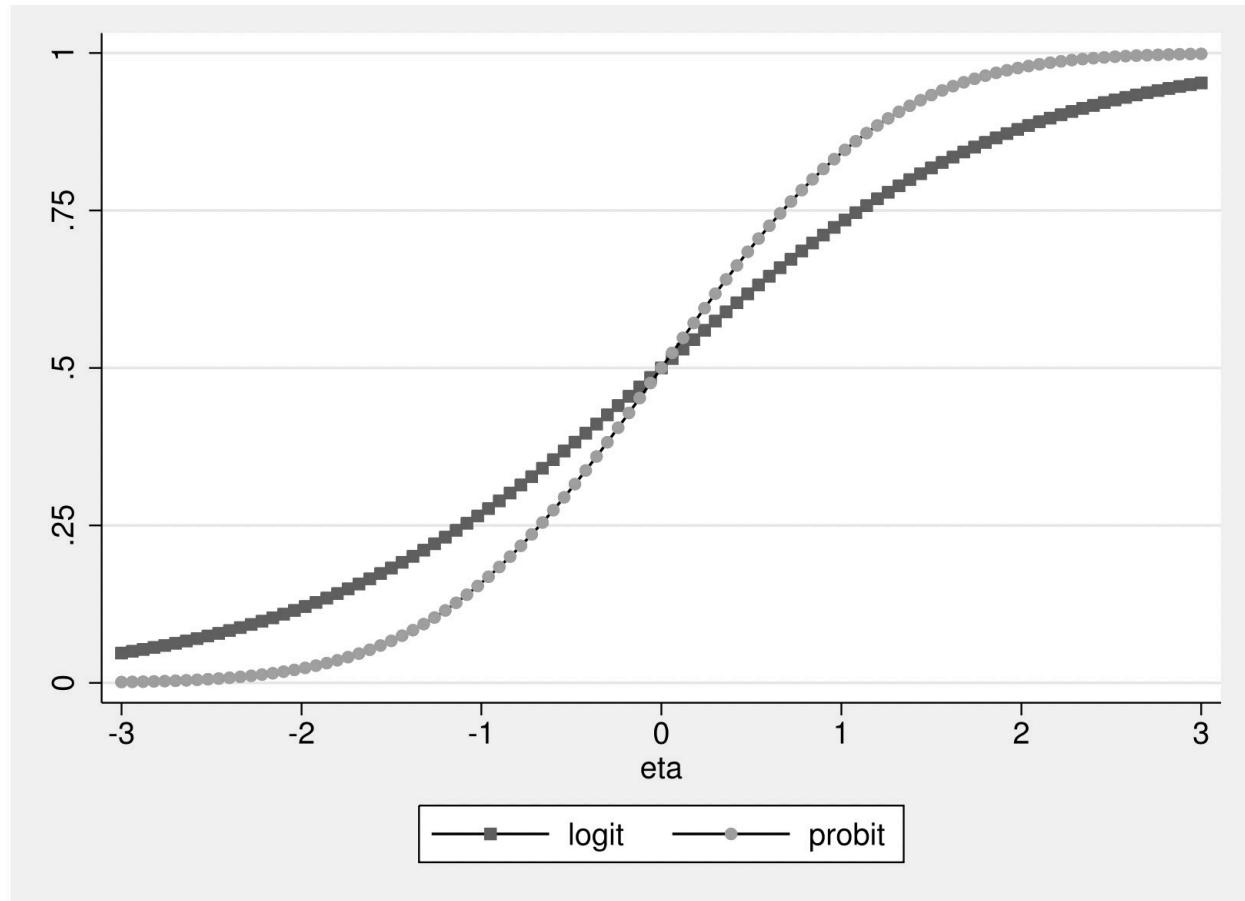


图 13.5.1: logit 和 probit 函数曲线对比。

我们可以非常容易的应用 IRLS 算法对 probit 模型进行参数估计, 只需要替换算法中的连接函数 g 、响应函数 r , 以及连接函数的导数 g' 。

$$\begin{aligned} \text{连接函数 } \eta &= g(\mu) = \Phi^{-1}(\mu) \\ \text{响应函数 } \mu &= g^{-1}(\eta) = \Phi(\eta) \\ \text{连接函数一阶导 } g'(\mu) &= \phi(\eta) \end{aligned} \quad (13.5.7)$$

probit 连接函数相比 logit 连接函数复杂了很多, 它不如 logit 更加的流行。但是, 当特征数据存在正态性时, probit 可能更合适。

13.5.3 log-log 和 clog-log

logit 和 probit 的曲线都是以点 $(0, 0.5)$ 对称的”S”型曲线, 所以, 二分类 logit 和 probit 模型假定响应数据中为 0 和 1 的比例是相同的。然而, 当响应数据中 0 和 1 的比例相差巨大时, clog-log 或 log-log 可能会提供更好的模型效果, 因为它们具有非对称性, 这些不对称的连接函数有时会更适合特殊的数据情况。

clog-log 和 log-log 是不对称的”S”型, 对于 clog-log 模型, S 曲线的上部比 logit 或 probit 更大或更长; 而 log-log 模型恰好相反, S 曲线的底部向左拉长或倾斜。

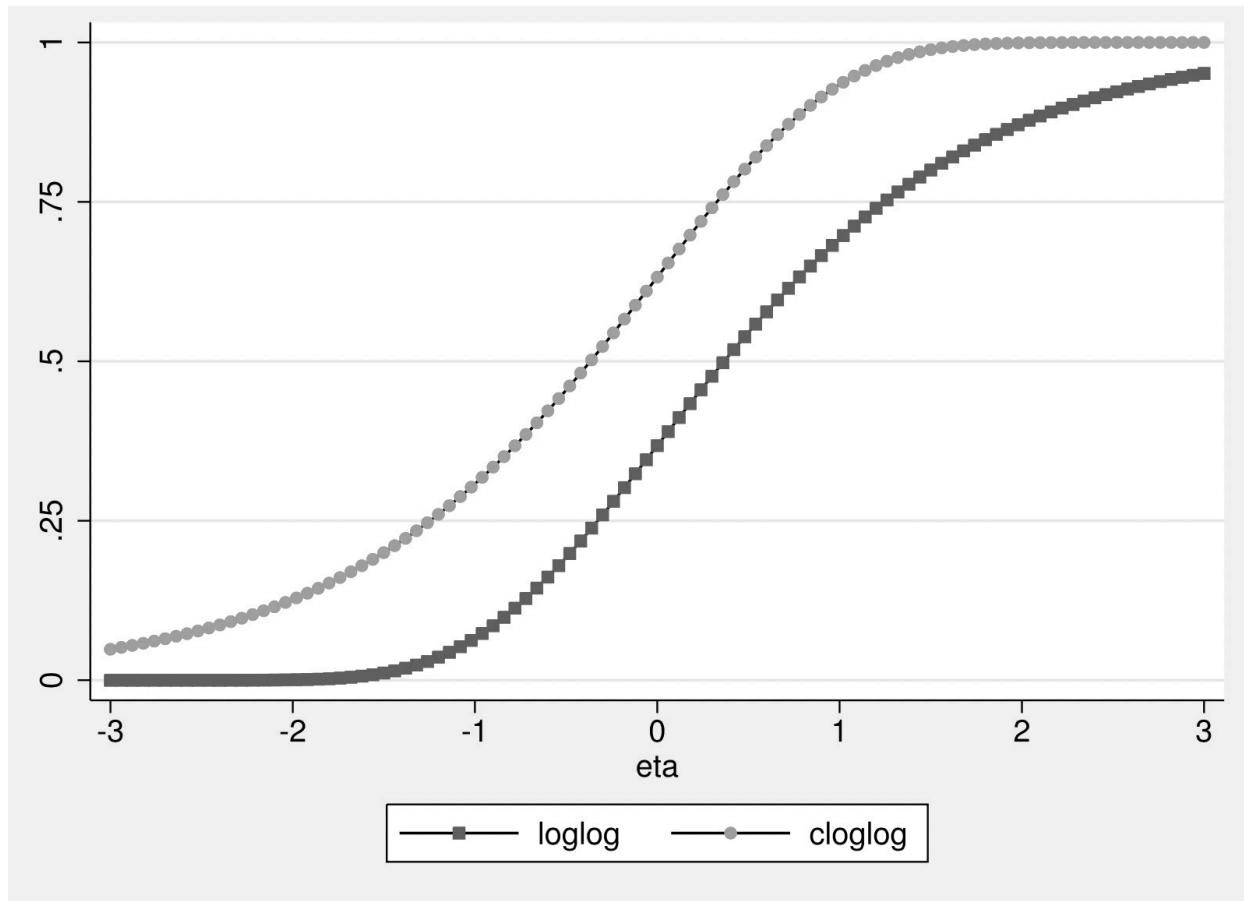


图 13.5.2: clog-log 和 log-log 函数

上图展示了 clog-log 和 log-log 连接函数的非对称性, 非对称结构使得我们可以更关注其中的一类。假设在二分类场景中, y 表示样本是正类的概率, $\bar{y} = 1 - y$ 表示样本是负类的概率。如果我们使用对称的

连接函数, 则可以拟合一个同时适用于两个类别的模型, 两个类别的模型仅仅是系数的正负号不同。此时, $probit(\bar{y})$ 和 $probit(\bar{y})$ 是互补的, $logit(y)$ 和 $logit(\bar{y})$ 是互补的。但这在非对称的链接模型中, 将不再适用, 在非对称连接函数 log-log 和 clog-log 的情况下, $loglog(y)$ 不再和 $loglog(\bar{y})$ 互补, $cloglog(y)$ 不再和 $cloglog(\bar{y})$ 互补, 而是 $loglog(y)$ 与 $cloglog(\bar{y})$ 互补。下面用公式展示出各自的互补关系。

$$\begin{aligned} probit(\bar{y}) &= \Phi^{-1}\bar{y} = \Phi^{-1}(1-y) = -\Phi^{-1}(y) = -probit(y) \\ logit(\bar{y}) &= \ln\left(\frac{\bar{y}}{1-\bar{y}}\right) = \ln\left(\frac{1-y}{1-(1-y)}\right) = \ln\left(\frac{1-y}{y}\right) = -\ln\left(\frac{y}{1-y}\right) = -logit(y) \\ loglog(\bar{y}) &= -\ln\{-\ln(\bar{y})\} = -\ln\{-\ln(1-y)\} = -[\ln\{-\ln(1-y)\}] = -cloglog(y) \\ cloglog(\bar{y}) &= \ln\{-\ln(1-\bar{y})\} = \ln[-\ln\{1-(1-y)\}] = -[-\ln\{-\ln(y)\}] = -loglog(y) \end{aligned} \quad (13.5.8)$$

直观的来讲, 就是在对称链接的情况下, 有 $r(\eta) + r(-\eta) = 1$, 而在非对称链接的情况下 $r(\eta) + r(-\eta) \neq 1$, clog-log 和 log-log 的关系是对称的, $r_{loglog}(\eta) + r_{cloglog}(-\eta) = 1$ 。

log-log

$$\begin{aligned} \text{连接函数: } \eta &= g(\mu) = -\ln[-\ln(\mu)] \\ \text{响应函数: } \mu &= r(\eta) = \exp[-\exp(-\eta)] \\ \text{连接函数导数: } g' &= -\mu \ln \mu \end{aligned} \quad (13.5.9)$$

Clog-log

$$\begin{aligned} \text{连接函数: } \eta &= g(\mu) = \ln[-\ln(1-\mu)] \\ \text{响应函数: } \mu &= r(\eta) = 1 - \exp[-\exp(\eta)] \\ \text{连接函数导数: } g' &= (\mu - 1) \ln(1 - \mu) \end{aligned} \quad (13.5.10)$$

13.6 分组数据与比例数据

伯努利模型对应着二分类的场景, 也就是一条数据可以有 0 或 1 两个类别, 此时响应变量的值 y_i 就是类别, 模型预测每条数据所属的类别。而二项式模型中, 一条数据就表示一组实验结果, 相比于伯努利数据, 每条数据中多了一个表示实验次数的 n_i , 响应变量的值 y_i 不再是 0 或 1 的二分类值, 而是表示 n_i 次实验中成功的次数, 其取值范围是 $0 \leq y_i \leq n_i$ 。

举个实际的例子说明下, 假设有两个赌徒想预测一个篮球运动员的投篮命中率。一个赌徒的做法是, 收集了这个球员 **每次投篮时** 的身体状态信息、天气状态、队员状态、对手状态等等信息, 以及本次投篮行为的结果, 进球还是没进球。然后训练了一个 **伯努利模型 (二分类模型)** 预测球员 **单次投篮进球的概率**。另一个赌徒的做法是, 收集这个球员 **每场比赛** 的信息, 这个球员在每场比赛中投了几次, 进了几次, 以及其它一些信息。然后训练了一个 **二项式模型** 预测球员 **一场比赛中进球率**, 即在投篮 n 次的情况下进球几次。

在二项式模型下, 一条数据表示一组实验的结果, 并且多了一个表示试验次数的 n_i , (x_i, n_i, y_i) , 此时样本数据称为分组数据 (grouped data), 一条数据相当于一个组。然而有时数据中的 n_i 是缺失的, 并且 y_i 不再是成功次数, 而是成功的比例 $y'_i = y_i/n_i$, 此时数据样本变成 (x_i, y'_i) , 这时称为比例 (proportional) 数据。

有些时候数据中可能缺少实验次数、成功次数这样的数据, 而仅有一个比例数据, 即成功率, 响应变量 Y 的值是一个 $[0, 1]$ 的比例值。此时, 可以使用线性回归模型拟合这个比例值, 但是, 线性回归模型的输出值可能会超出 $[0, 1]$ 的区间范围。这时, 我们可以先把 y 值转换一下。

$$y_{new} = logit(y) = \ln\left(\frac{y}{1-y}\right) \quad (13.6.1)$$

经过 `logit` 函数转换后, 新的 y_{new} 值就是实数域范围了, 然后再应用线性回归模型。但是注意 `logit` 函数不能处理 0 和 1 的样本值。

泊松模型

前文我们讨论了离散数据模型中的二项式模型，二项式分布描述的是二值离散变量。在离散数据中还有另一种数据形式，计数数据。计数是对某一事件的简单计数，其取值是 $0, 1, 2, \dots$ 等大于等于 0 的整数，通常用于描述单位时间内某个事件的发生次数。在指数族中，用于表示计数变量的概率分布是泊松分布 (Poisson distribution)，计数变量也被称为泊松变量，该模型以 Poisson (1837) 提出的研究命名。

泊松分布和二项式分布是存在关联的，泊松分布可以看做是二项分布的极限情况，二项式分布表示进行 N 次伯努利实验成功的次数，而泊松分布表示单位时间或者空间内事件发生的次数，二者很相似，泊松分布就是二项式分布中 N 趋近于无穷时的情况。本章我们先讨论如何从二项式分布推导出泊松分布，然后再讨论 GLM 家族中泊松模型的特性。

14.1 泊松 (Poisson) 分布

泊松 (Poisson) 分布的直观理解是，在一个单位时间或者空间间隔内，随机事件发生次数的概率。比如：

- 每个小时出生的婴儿数量
- 每分钟人类心脏的跳动次数
- 空气中每立方米中氧气分子的数量
- 高速公路上每公里汽车的数量

泊松模型是最基本的计数模型，本章我们重点讨论泊松模型，再后续的章节中再讨论其它的计数模型。

14.1.1 推导过程

泊松分布实际上是二项分布的试验 (trials) 次数 N 趋近于无穷时的场景, 我们用一个例子说明。假设一个交通观察员需要对某个路口的车流量进行建模, 然后用模型预测未来一个小时从这个路口通过的车次。为了简化问题, 我们假设路口的交通量不存在高峰期低峰期, 即交通量不会随着时间的变化而变化, 并且每个时间片段内通过的车辆是互不影响的, 即前一小时内车辆通过与否不影响下一个小时内车辆。观察员首先根据这个路口历史上车辆通过情况, 计算出平均每小时通过车辆的数量为 λ 。我们把一个小时从路口通过的车次数看做一个随机变量, 用符号 X 表示, 那么 λ 就是变量 X 的数学期望。

$$\mathbb{E}[X] = \lambda \quad (14.1.1)$$

我们把一辆车通过与否看做一个伯努利变量, 类似投硬币实验, 1 表示车辆通过, 0 表示车辆不通过。把一个小时的时间区间均分成 N 个时间片段, 比如每分钟作为一个片段, 这时 $N = 60$ 。每个时间片段有车辆通过就是一次成功的实验 (类似于投硬币正面向上), 没有车辆通过就是一次失败的实验 (类似于投硬币反面向上), 这样就把一小时内车辆通过问题转化成一个二项分布问题, 在 N 次实验中有 k 次成功 (车辆通过) 概率分布函数可以写成:

$$p(X = k) = \binom{N}{k} p^k (1 - p)^{N-k} \quad (14.1.2)$$

其中 p 是一次实验的成功概率, 我们已经通过历史数据知道平均每小时 (N 次实验) 中通过的车次数为 λ , 意味着 n 次实验中有 λ 次成功, 单次实验成功的概率 (平均一分钟内通过车辆数) 为:

$$p = \frac{\lambda}{N} \quad (14.1.3)$$

但是我们并不能保证每分钟只有一辆车通过, 我们需要保证一个时间片段内只有一辆车通过 (一次实验) 以上的二项分布的假设才有意义。理论上, 我们只要把一小时的时间区间拆的足够小, 比如拆成每秒, 甚至是每毫秒为一个时间片段, 这样就能尽量保证每个时间片段内只会有一辆车通过。 N 越大时间片段就越小, 极限情况, 我们可以把一小时分割成每个车辆通过的“瞬间”。换句话说, 只要 $N \rightarrow \infty$ 上述假设就是成立的, 因此我们为公式 (14.1.2) 加上极限操作。

$$p(X = k) = \lim_{N \rightarrow \infty} \binom{N}{k} p^k (1 - p)^{N-k} \quad (14.1.4)$$

我们发现公式 (14.1.4) 就是二项分布的极限情况, 表示的是路口未来一小时内通过的车辆数的概率分布, $p(X = k)$ 表示在一小时内通过车辆数为 k 的概率。 λ 表示这个时间区间内通过车辆数的期望值, 至于这个时间区间是一小时还是两小时并不重要, 只要是一个固定的时间区间就行, 所以可以看成是单位时间区间内, 或者 t 时间区间内。

公式 (14.1.4) 带有极限操作, 事实上可以通过一些变换去掉极限符号, 现在我们尝试对其进行一些变换。

$$\begin{aligned} p(X = k) &= \lim_{N \rightarrow \infty} \binom{N}{k} p^k (1 - p)^{N-k} \\ &= \lim_{N \rightarrow \infty} \frac{N!}{(N - k)! k!} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \lim_{N \rightarrow \infty} \frac{N!}{(N - k)! k!} \frac{\lambda^k}{N^k} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{\lambda}{N}\right)^{-k} \end{aligned} \quad (14.1.5)$$

结合如下两个等式,

$$\frac{N!}{(N - k)!} = \frac{N(N - 1) \cdots 2 \times 1}{(N - k)(N - k - 1) \cdots 2 \times 1} = \underbrace{N(N - 1) \cdots (N - k + 1)}_{k \uparrow} \quad (14.1.6)$$

$$\lim_{x \rightarrow a} f(x)g(x) = \lim_{x \rightarrow a} f(x) \lim_{x \rightarrow a} g(x) \quad (14.1.7)$$

公式 (14.1.5) 变成：

$$\begin{aligned} p(X = k) &= \lim_{N \rightarrow \infty} \frac{N(N-1) \cdots (N-k+1)}{N^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{\lambda}{N}\right)^{-k} \\ &= \frac{\lambda^k}{k!} \lim_{N \rightarrow \infty} \left[\frac{N(N-1) \cdots (N-k+1)}{N^k} \right] \lim_{N \rightarrow \infty} \left[\left(1 - \frac{\lambda}{N}\right)^N \right] \lim_{N \rightarrow \infty} \left[\left(1 - \frac{\lambda}{N}\right)^{-k} \right] \end{aligned} \quad (14.1.8)$$

其中各个极限都可以有近似表示。

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{N(N-1) \cdots (N-k+1)}{N^k} &= 1 \\ \lim_{N \rightarrow \infty} \left(1 - \frac{\lambda}{N}\right)^N &= \lim_{N \rightarrow \infty} \left(1 + \frac{-\lambda}{N}\right)^N = e^{-\lambda} \\ \lim_{N \rightarrow \infty} \left[\left(1 - \frac{\lambda}{N}\right)^{-k} \right] &= 1 \end{aligned} \quad (14.1.9)$$

$$p(X = k|N) = \frac{\lambda^k}{k!} \times 1 \times e^{-\lambda} \times 1 = \frac{\lambda^k}{k!} e^{-\lambda} \quad (14.1.10)$$

上式就表示在单位(固定)时间区间内, 随机事件发生 k 次的概率, 这就是泊松分布。上式稍微整理下, 就得到泊松分布的概率质量函数。

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (14.1.11)$$

其中变量 x 表示在单位时间内事件发生的次数, 显然 x 是一个离散变量, 因此泊松分布是一个离散变量分布。 λ 是变量 x 的期望值, 表示在单位时间内事件发生的平均次数, 因此通常也可以用 μ 代替 λ 。

$$p(x) = \frac{\mu^x}{x!} e^{-\mu} \quad (14.1.12)$$

二项式分布 $Binomial(k, n)$ 表示进行 n 实验成功 k 次的概率, 需要知道 n 的值才行, 并且没有时间区间的概念。而泊松分布 $Poisson(\lambda)$ 表示单位时间内事件发生 x 次的概率, 其用单位时间的概念替代了 n 的作用, 并且这个单位时间具体多长并不重要, 只是把整体时间分成相同长度的小片段。

注意, 在泊松分布中, **各个时间区间之间是相互独立的**, 互不影响, 也就是不会因为当前时间区间内有车辆通过, 而导致下一个时间区间内通过的车辆受到影响。泊松分布的应用并不是仅限于固定的时间区间, 理论上只要是固定的区间 (fixed interval) 即可, 比如固定大小的时间、长度、空间、面积、体积等等。

14.1.2 泊松分布的特性

通过泊松分布的概率质量函数公式 (14.1.12), 可以看到泊松分布是一个单参数的分布, 其唯一的参数就是分布的期望值 μ 。理论上, 泊松变量 X 可以取 0 值, 但泊松分布的期望值 μ 一定是大于 0 的, 现在我们看下不同的 μ 值下分布的差异。

图 14.1.1 是不同的 μ 下泊松分布的概率分布曲线。从中可以看出当 μ 比较小时, 图形是偏态的, 随着 μ 的增大, 图形逐渐接近正态分布。

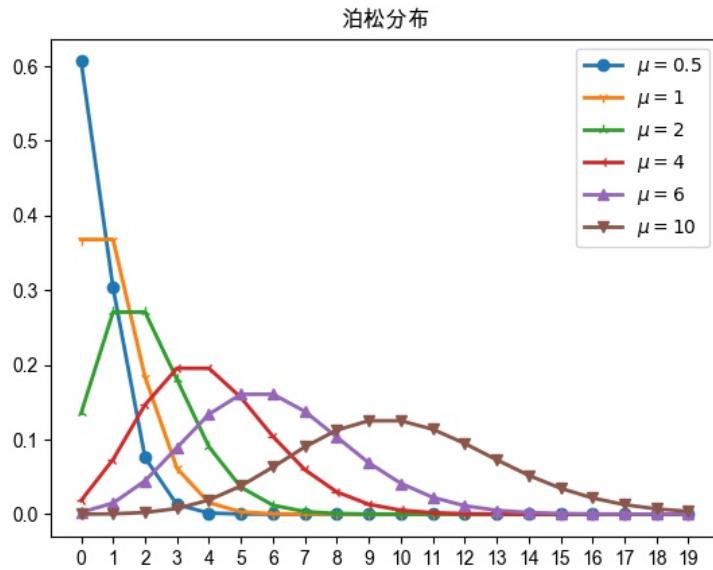


图 14.1.1: 不同均值参数下泊松分布的概率质量函数

分布的矩

14.2 泊松回归模型

泊松分布的概率分布函数通常写成：

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} \quad (14.2.1)$$

其中阶乘部分可以用 $\Gamma(y + 1)$ 函数代替。

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{\Gamma(y + 1)} \quad (14.2.2)$$

现在转化成指数族的形式：

$$f(y; \mu) = \exp\{y \ln(\mu) - \mu - \ln \Gamma(y + 1)\} \quad (14.2.3)$$

泊松分布的规范连接函数和累积函数为：

$$\begin{aligned} g(\mu) &= \theta = \ln(\mu) \\ b(\theta) &= \mu \\ a(\phi) &= 1 \end{aligned} \quad (14.2.4)$$

泊松分布的规范连接函数是对数 (log) 函数，因此响应 (反链接) 函数就是指数函数， $\mu = \exp(\eta)$ 。连接函数和响应函数的导数分别为：

$$\begin{aligned} g'(\mu) &= \frac{1}{\mu} \\ r'(\eta) &= \exp(\eta) \end{aligned} \quad (14.2.5)$$

对 $b(\theta)$ 求导得到其均值和方差函数。

$$\begin{aligned}
 b'(\theta) &= \frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \\
 &= (1)(\mu) \\
 &= \mu \\
 b''(\theta) &= \frac{\partial^2 b}{\partial \mu^2} \left(\frac{\partial \mu}{\partial \theta} \right)^2 + \frac{\partial b}{\partial \mu} \frac{\partial^2 \mu}{\partial \theta^2} \\
 &= (0)(1)^2 + (\mu)(1) \\
 &= \mu
 \end{aligned} \tag{14.2.6}$$

泊松分布的方差为 $V(y) = a(\phi)b''(\theta) = \mu$ ，我们发现，**泊松分布的方差和均值是相同的**，牢记泊松分布的这个特点，后续我们会详细讨论这个特点带来的一些影响。因为泊松分布的方差和均值相同，所以泊松分布的变异系数是 $c_\nu = \frac{\sqrt{\mu}}{\mu} = 1/\sqrt{\mu}$ 。

注解：变异系数 (Coefficient of Variation)，又称“离散系数”、“变差系数”，是概率分布离散程度的一个归一化量度，其定义为标准差与平均值之比。当需要比较两组数据离散程度大小的时候，如果两组数据的测量尺度相差太大，或者数据量纲的不同，直接使用标准差来进行比较不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点，它是原始数据标准差与原始数据平均数的比。变异系数没有量纲，这样就可以进行客观比较了。事实上，可以认为变异系数和极差、标准差和方差一样，都是反映数据离散程度的绝对值。其数据大小不仅受变量值离散程度的影响，而且还受变量值平均水平大小的影响。

最后总结一下泊松模型的关键部分。

$$\begin{aligned}
 \text{标准连接函数: } &\eta = g(\mu) = \ln(\mu) \\
 \text{反连接 (响应) 函数: } &\mu = r(\eta) = \exp(\eta) \\
 \text{方差函数: } &\nu = \mu \\
 \text{分散函数: } &a(\phi) = 1 \\
 \text{连接函数导数: } &g' = \frac{1}{\mu}
 \end{aligned} \tag{14.2.7}$$

14.3 参数估计

GLM 中泊松模型的参数估计，同样可以应用 IRLS 算法解决，按照前文讨论的 IRLS 算法的过程，我们只需要求出泊松模型对应的 W 矩阵和 Z 矩阵即可，先从泊松模型的对数似然函数开始。泊松模型的对数似然函数可以直接写出。

$$\ell(\hat{\mu}; y) = \sum_{i=1}^N \{y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln \Gamma(y_i + 1)\} \tag{14.3.1}$$

根据公式 (8.1.12)，泊松模型的得分统计量为

$$\begin{aligned}
 U_j &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \hat{\mu}_i}{a(\phi)\nu(\hat{\mu}_i)g(\hat{\mu}_i)'} x_{ij} \\
 &= \sum_{i=1}^N (y_i - \hat{\mu}_i) x_{ij}
 \end{aligned} \tag{14.3.2}$$

泊松模型的 W 和 Z 分别为

$$\begin{aligned} W_{ii} &= \frac{1}{a(\phi)\nu(\hat{\mu}_i)(g'_i)^2} \\ &= \hat{\mu}_i \end{aligned} \quad (14.3.3)$$

$$\begin{aligned} Z_i &= (y_i - \hat{\mu}_i)g'_i + \eta_i \\ &= \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} + \eta_i \end{aligned} \quad (14.3.4)$$

14.4 拟合统计量

我们知道, 在 GLM 中评估模型优劣的方法一般有三种, 拟合优度统计量 (goodness-of-fit statistic)、残差统计量 (residual statistic)、以及 AIC、BIC 等信息量准则。在拟合优度统计量中, 最常用的就是偏差统计量, 泊松模型的偏差统计量为:

$$\begin{aligned} D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\ &= 2 \sum_{i=1}^n \{y_i \ln(y_i) - y_i - y_i \ln(\hat{\mu}_i) + \hat{\mu}_i\} \\ &= 2 \sum_{i=1}^n \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right\} \end{aligned} \quad (14.4.1)$$

上述偏差统计量的计算公式有个问题, 就是当响应数据 $y_i = 0$ 时, $\ln(y_i/\hat{\mu}_i)$ 是没有意义的, 所以需要单独处理 0 值的数据。当 $y_i = 0$ 时, 其预测模型的对数似然函数简化为:

$$\ell_i(\hat{\mu}_i; 0) = -\hat{\mu}_i \quad (14.4.2)$$

此时, 饱和模型的对数似然函数为:

$$\ell_i(0; 0) = 0 \quad (14.4.3)$$

因此, 对于响应数据 $y_i = 0$ 的样本, 其偏差为:

$$D_i(y_i = 0) = 2\hat{\mu}_i \quad (14.4.4)$$

泊松模型的皮尔逊卡方统计量为

$$\begin{aligned} \chi^2 &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)} \\ &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}} \end{aligned} \quad (14.4.5)$$

14.5 频率模型

单位时间内发生的次数就是频次或者频率, 因此泊松分布也可以看做是对频率数据进行建模, 通常泊松分布可以引入一个表示时间或者空间的常量系数 t 。

$$f(y; \mu) = \frac{e^{-t\mu}(t\mu)^y}{y!} \quad (14.5.1)$$

常数系数 t 表示时间长度或者空间大小, μ 表示频率参数, $t\mu$ 就表示在长度为 t 的时间或者空间窗口内事件发生的次数的期望值。当 $t = 1$ 是就退化成泊松分布的标准形式。

14.6 泊松模型的局限性

指数模型

15.1 指数 (exponential) 分布

泊松 (Poisson) 分布是预测在一个固定时间间隔内，随机事件发生 n 次的概率。而指数分布是预测下一次时间发生需要等待多长的时间。比如，下面几种情况：

- 直到客户下次购买商品为止（成功）的时间。
- 机器下次发生故障的时间。
- 下次公交车到达需要等待的时间。

15.1.1 推导过程

在泊松分布中有一个参数 λ ，其表示在单位时间区间内事件发生次数的平均值，其是一个单位时间的比例 (rate) 值。那么 λ 倒数 $1/\lambda$ 是什么含义呢？ $1/\lambda$ 就表示 事件发生一次需要的时间的平均值。比如，当 $\lambda = 0.25$ 时，表示在单位时间内平均发生了 0.25 次，倒过来 ($1/\lambda = 1/0.25 = 4$) 就是发生一次需要 4 个单位时间。

指数概率分布表示，在一个泊松过程中，两次事件发生的间隔时间的概率分布。用泊松分布的表达就是在等待的之间范围内一次事件都没发生，这意味着有 $Poisson(x = 0)$ 。

$$p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$p(x = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda} \quad (15.1.1)$$

需要注意的是，在泊松分布的概率函数中时间间隔仅仅是 1 个单位时间 (unit time)。如果想要建立一个在任意时间区间 t (而不是一个单位时间) 无事件发生的概率分布，我们需要做些什么呢？泊松分布假设事件之间是相互独立的，计算 t 个单位时间内 0 次发生的概率可以是每个单位时间内 0 次发生概率的连乘。

$$\underbrace{p(T > t)}_{\text{在 } t \text{ 个单位时间内没有发生}} = \underbrace{p(x = 0)}_{\text{第 1 个单位时间}} \times \underbrace{p(x = 0)}_{\text{第 2 个单位时间}} \dots \underbrace{p(x = 0)}_{\text{第 } t \text{ 个单位时间}} = e^{-\lambda t} \quad (15.1.2)$$

T 表示下次事件发生的时间, $p(T > t)$ 就表示下次事件发生的时间晚于 t 的概率, 换句话说, 就是 t 时间内没有发生的概率。那么 t 时间内发生的概率就是:

$$p(T \leq t) = 1 - p(T > t) = 1 - e^{-\lambda t} \quad (15.1.3)$$

上式是累计分布函数 (cumulative distribution function, CDF), 通过对 CDF 进行微分 (求导) 就得到了分布的概率密度函数 (probability distribution function, PDF)。

$$f(t) = \frac{dp(T \leq t)}{dt} = \lambda e^{-\lambda t} \quad (15.1.4)$$

如果单位时间内事件发生次数符合泊松分布, 那么事件发生的间隔时间就服从指数分布。

15.1.2 分布的特性

曲线图

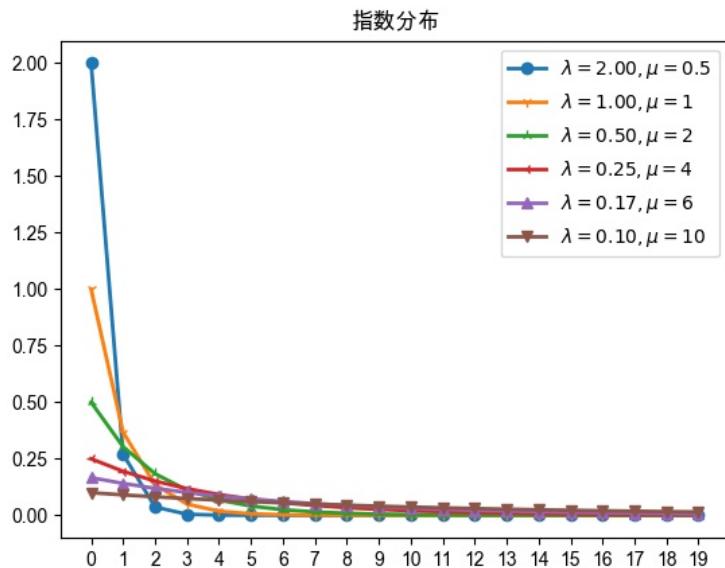


图 15.1.1: 指数分布的概率密度函数

分布的矩

15.2 指数回归模型

指数分布的概率分布函数通常写成如下的形式:

$$f(y; \lambda) = \begin{cases} \lambda e^{-\lambda y} & y \geq 0 \\ 0 & y < 0 \end{cases} \quad (15.2.1)$$

其中 $\lambda > 0$ 是分布的一个参数, 常被称为率参数 (rate parameter), 即每单位时间发生该事件的次数。指数分布的区间是 $[0, \infty)$ 只有 $y \geq 0$ 才有意义。如果一个随机变量 Y 呈指数分布, 则可以写作: $Y \sim \text{Exponential}(\lambda)$ 。

指数分布的响应变量的值是大于等于 0 的, 从概率密度函数的图形上可以看出, 0 的概率是最大的, 并且值越大概率越小, 是一个递减的非对称结构, 这和高斯模型的对称结构是完全一样的。

指数分布的期望为:

$$\mathbb{E}[Y] = \frac{1}{\lambda} = \mu \quad (15.2.2)$$

方差为:

$$Var(Y) = \frac{1}{\lambda^2} \quad (15.2.3)$$

指数分布的期望 μ 和比率参数 λ 互为倒数的关系, 二者是一一映射的, 所以可以用 μ 参数化概率密度函数。

$$f(y; \mu) = \frac{1}{\mu} e^{-\frac{y}{\mu}} \quad (15.2.4)$$

现在我们把公式 (15.2.4) 转化成 GLM 的形式。

$$f(y; \mu) = \exp\left\{-\frac{y}{\mu} + \ln\left(\frac{1}{\mu}\right)\right\} \quad (15.2.5)$$

GLM 中指数族分布标准形式为:

$$p(y|\theta) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (15.2.6)$$

对比下公式 (15.2.4) 和公式 (15.2.6), 可以直接给出各个组件的内容。

$$\begin{aligned} \theta &= -\frac{1}{\mu} \\ b(\theta) &= -\ln\left(\frac{1}{\mu}\right) \\ a(\phi) &= \phi = 1 \end{aligned} \quad (15.2.7)$$

现在我们来看下 $b(\theta)$ 的导数。

$$\begin{aligned} b'(\theta) &= \frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \\ &= \left(-1 \times \mu \frac{-1}{\mu^2}\right) (\mu^2) \\ &= \mu \\ b''(\theta) &= \frac{\partial^2 b}{\partial \theta^2} \\ &= \frac{\partial}{\partial \theta} \left(\frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \right) \\ &= \frac{\partial}{\partial \theta}(\mu) \\ &= \frac{\partial}{\partial \mu} \mu \frac{\partial \mu}{\partial \theta} \\ &= (1)(\mu^2) = \mu^2 = \nu(\mu) \end{aligned} \quad (15.2.8)$$

指数分布的方差函数为 $\nu(\mu) = \mu^2$, 分散函数为 $a(\phi) = 1$, 因此指数分布的方差为:

$$Var(y) = a(\phi)\nu(\mu) = \mu^2 \quad (15.2.9)$$

显然指数分布的方差与期望是平方关系, 这与高斯分布是不一样的, 高斯分布的方差与期望是无关的。指数分布的 $a(\phi) = 1$ 使其不再需要一个额外的分散参数 ϕ , 只需要一个期望参数即可, 这使得指数分布是一个单参数模型。

根据规范链接函数的定义, 指数分布的标准链接函数是负倒数函数。

$$\begin{aligned} \text{标准链接函数: } \eta &= \theta = g(\mu) = -\frac{1}{\mu} \\ \text{标准响应函数: } \mu &= -\frac{1}{\eta} \end{aligned} \quad (15.2.10)$$

链接函数的导数可以简单得到。

$$g'(\mu) = \frac{1}{\mu^2} \quad (15.2.11)$$

采用标准链接的指数分布模型, 通常又被称为指数回归模型 (exponential regression), 经常用来处理正数响应数据。

15.3 参数估计

15.3.1 似然函数

指数分布的对数似然函数形式比较简单, 因为其没有 $c(y, \phi)$ 这一项。

$$\ell(\hat{\mu}; y) = \sum_{n=1}^N \left\{ -\frac{y_i}{\mu_i} + \ln \left(\frac{1}{\mu_i} \right) \right\} \quad (15.3.1)$$

对数似然函数的一阶导数又叫做得分统计量, 或者得分函数 (score function), 用符号 U 表示, 似然估计目标就是求解方程 $U = 0$ 。

$$\begin{aligned} U &= \frac{\partial \ell}{\partial \mu_i} \\ &= \sum_{i=1}^N \left\{ \frac{y_i}{\mu_i^2} - \frac{1}{\mu_i} \right\} \end{aligned} \quad (15.3.2)$$

链接函数是标准链接函数的情况下, 用 β 重新参数化似然函数。

$$\begin{aligned} \ell(\beta; y) &= \sum_{n=1}^N \left\{ -\frac{y_i}{\mu_i} + \ln \left(\frac{1}{\mu_i} \right) \right\} \\ &= \sum_{i=1}^N \{ \eta_i y_i + \ln(-\eta_i) \} \\ &= \sum_{i=1}^N \{ (x_i \beta) y_i + \ln(-x_i \beta) \} \end{aligned} \quad (15.3.3)$$

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N y_i x_{ij} + \frac{x_{ij}}{x_i \beta} = \sum_{i=1}^N \left(y_i - \frac{1}{x_i \beta} \right) x_{ij} \quad (15.3.4)$$

x_i 是行向量, β 是参数列向量, 二者 \mathbb{F} 积是一个标量。

15.3.2 拟合优度

现在我们来看下模型的偏差 (Deviance)，首先写出饱和模型的似然函数，只需要把公式 (15.3.1) 中的 $\hat{\mu}_i$ 替换成 y_i 即可。

$$\begin{aligned}\ell(y; y) &= \sum_{n=1}^N \left\{ -\frac{y_i}{y_i} + \ln\left(\frac{1}{y_i}\right) \right\} \\ &= \sum_{n=1}^N \{-1 - \ln(y_i)\}\end{aligned}\tag{15.3.5}$$

模型的偏差为：

$$\begin{aligned}D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\ &= 2 \sum_{n=1}^N \left\{ -1 - \ln(y_i) + \frac{y_i}{\hat{\mu}_i} - \ln\left(\frac{1}{\hat{\mu}_i}\right) \right\} \\ &= 2 \sum_{n=1}^N \left\{ \frac{y_i}{\hat{\mu}_i} - \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - 1 \right\}\end{aligned}\tag{15.3.6}$$

高斯分布存在分散参数 $a(\phi) = \sigma^2$ 。指数分布是没有分散参数 ϕ 的，因此指数分布模型的偏差和尺度化偏差是一样的。

高斯分布的方差函数是常量 $\nu(\mu) = 1$ ，然而指数分布的方差函数是 $\nu(\mu) = \mu^2$ ，指数分布的皮尔逊卡方统计量是

$$\chi^2 = \frac{\sum(y_i - \hat{\mu}_i)^2}{\hat{\mu}^2}\tag{15.3.7}$$

15.3.3 IRLS

IRLS 算法是 GLM 模型参数估计统一框架，适用于所有 GLM 模型。牛顿法使用的是观测信息矩阵 (OIM)，IRLS 使用的是期望信息矩阵 (EIM)。IRLS 算法参数更新公式为：

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)}\tag{15.3.8}$$

我们只需要根据具体的模型去计算出相应的 W 和 Z 即可。权重矩阵 W 是一个对角矩阵，计算方法如下：

$$W^{(t)} = \text{diag} \left\{ \frac{1}{a(\phi)\nu(\mu)(g')^2} \right\}_{(n \times n)}\tag{15.3.9}$$

Z 是一个列向量，其计算方法如下：

$$Z^{(t)} = \left\{ (y - \hat{\mu})g' + \eta^{(t)} \right\}_{(n \times 1)}\tag{15.3.10}$$

不同的分布拥有不同的 $a(\phi), \nu(\mu), g$ ，只需按照特定分布提供即可。这里我们再次给出指数分布的各项内容。

$$\begin{aligned}a(\phi) &= 1 \\ \nu(\mu) &= \mu^2 \\ g'(\mu) &= \frac{1}{\mu^2}\end{aligned}\tag{15.3.11}$$

对于规范链接函数的指数分布，其 W 和 Z 分别为：

$$W^{(t)} = \text{diag} \left\{ \frac{1}{a(\phi)\nu(\mu)(g')^2} \right\}_{(n \times n)} = \text{diag} \left\{ \hat{\mu}^2 \right\}_{(n \times n)}\tag{15.3.12}$$

$$Z^{(t)} = \left\{ (y - \hat{\mu})g' + \eta^{(t)} \right\}_{(n \times 1)} = \left\{ \frac{(y - \hat{\mu})}{\hat{\mu}^2} + \eta^{(t)} \right\}_{(n \times 1)}\tag{15.3.13}$$

CHAPTER 16

Gamma 模型

指数分布是预测下一次时间发生等待的时间，更进一步，如果需要预测第 k 次事件发生需要等待的时间呢？这就是 Gamma 分布。在介绍 Gamma 分布之前，先简单的介绍一下 Gamma 函数。

16.1 Gamma 函数

https://www.probabilitycourse.com/chapter4/4_2_4_Gamma_distribution.php

16.2 Gamma 分布

在之前的章节我们已经讨论了如何从泊松过程推导出指数分布，Gamma 分布的推导的过程是类似的。不同的地方在于指数分布是等待下一次事件发生，而 Gamma 分布是等待第 k 次事件的发生。

我们用符号 k 表示事件发生的次数，符号 T 表示直到 k 次事件发生等待的时间，也就是目标随机变量。 λ 表示泊松过程中事件发生的频率（单位时间内发生的次数）。 $p(T > t; k)$ 表示直到第 k 次事件发生等待的时间 T 大于 t 的概率。用符号 $p(k; T = t)$ 表示在泊松过程中 t 个时间单元内事件发生 k 次的概率。

泊松分布表示单位时间窗口内事件发生次数的概率分布，泊松分布的概率分布函数为：

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (16.2.1)$$

其表示在一个单位时间内事件发生 k 次的概率，唯一的参数 λ 表示单位时间内事件发生次数的平均值，也就是 k 的期望值， $\mathbb{E}[x] = \lambda$ 。现在要想计算 t 个单位时间内事件发生的次数，只需要用 λt 替换上式中的 λ 即可， t 个时间单元内事件发生 k 次的概率为：

$$p(k; t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (16.2.2)$$

注解：为什么用 λt 替换 λ 就可以？

泊松分布表示单位时间内，随机事件发生 k 次的概率分布。这个单位时间并没有具体的长度限制，只要保证每个时间片段的长度相同就可以。泊松分布唯一的参数就是每个时间片段内随机事件发生次数的平均值 λ ，如果要算 t 个时间片段内随机事件发生次数的概率分布，相当于原来的时间片段扩大了 t 倍，这个 t 倍的时间片段也可以看做是一个“单位时间”，这个新的“单位时间”组成一个新的泊松分布，并且其平均值参数 λ_{new} 也是原来的 t 倍，即 $\lambda_{new} = \lambda t$ 。

$$p(k) = \frac{(\lambda_{new})^k}{k!} e^{-\lambda_{new}} = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (16.2.3)$$

泊松分布表示的是一个单位时间片段内随机事件发生 **次数** 的概率分布，指数分布表示的是随机事件 **第一次**发生的需要的 **时间** 的概率分布。泊松分布描述的是 **次数**，是离散变量的分布；指数分布描述的是 **时间**，是连续值变量的分布。然而指数分布仅仅描述了事件首次发生，不能表示更多次事件发生，不具备一般化，Gamma 分布就是指数分布的扩展，能够表达事件发生任意次数所需 **时间** 的概率分布。

我们令符号 $p(T > t; k)$ 表示随机事件发生 k 所需时间大于 t 的概率。如果随机事件第 k 次发生的时间大于 t ，意味着在 t 个时间单元内，事件发生次数一定是小于等于 $k - 1$ 次。换句话说， t 个时间单元内最多只发生 $k - 1$ 次，第 k 次发生的时间一定是大于 t 。因此 $p(T > t; k)$ 可以看成是在 t 个时间单元内，事件发生 $0, 1, 2, \dots, k - 1$ 次的概率之和。

$$p(T > t; k) = \sum_{i=0}^{k-1} p(i; t) = \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!} e^{-\lambda t} \quad (16.2.4)$$

现在我们得到了第 k 次事件发生时间大于 t 的概率 $p(T > t; k)$ ，反过来，第 k 次事件发生时间小于等于 t 的概率为：

$$\begin{aligned} p(T \leq t; k) &= 1 - p(T > t; k) \\ &= 1 - \sum_{i=0}^{k-1} \frac{(\lambda t)^i}{i!} e^{-\lambda t} \end{aligned} \quad (16.2.5)$$

显然公式 (16.2.5) 是一个累积分布函数 (Cumulative Distribution Function)，是概率密度函数的积分，对其进行微分可以得到概率密度函数，这里省略微分的过程，直接给出结果

$$\begin{aligned} f(t) &= \frac{\lambda e^{-\lambda t} (\lambda t)^{k-1}}{(k-1)!} \\ &= \frac{\lambda^k e^{-\lambda t} t^{k-1}}{\Gamma(k)} \end{aligned} \quad (16.2.6)$$

公式 (16.2.6) 就是 Gamma 分布的概率密度函数，其表示随机事件发生 k 次所需时间 t 的概率分布。注意， $f(t)$ 不是具体的时间值，而是时间 t 的概率分布。参数 k 表示事件发生的次数，又被称为形状参数 (shape parameter)。参数 λ 来源于泊松分布，表示随机事件发生的频次，即单位时间发生的平均次数，又被称为速率参数 (rate parameter)。

注解：什么是形状参数 (shape parameter)?

在概率分布中，按照参数对概率分布函数曲线的影响，可以分为几种。

- 位置参数 (location parameter)，影响着图形在 x 轴上的位置；
- 尺度参数 (scale parameter)，控制着图形的拉伸和缩小，可以缩放图形。尺度参数的倒数称为速率参数 (rate parameter)，对图形的影响与尺度参数是一样的。
- 形状参数 (shape parameter)，其既不是位置参数也不是尺度参数（也不是关于这两者的函数）。形状参数直接影响图形分布的形状，而不是简单地移动分布（如位置参数）或拉伸/缩小分布（如比例参数）。

Gamma 分布的期望和方差分别为:

$$\begin{aligned}\mathbb{E}[T] &= \frac{k}{\lambda} \\ Var(T) &= \frac{k}{\lambda^2}\end{aligned}\tag{16.2.7}$$

k 是事件发生总次数, λ 是事件发生速率, 显然事件发生 k 次所需要的平均时间就是 k/λ , 这和 Gamma 分布的期望是一致的。

当 $k = 1$ 时, Gamma 分布就退化为指数分布:

$$f(t; k = 1, \lambda) = \lambda e^{-\lambda t} \quad (\lambda, t > 0)\tag{16.2.8}$$

因此有 $Gamma(1, \lambda) = Exponential(\lambda)$ 成立, 更一般的, 如果有 n 个独立的指数分布 $Exponential(\lambda)$ 随机变量, 就可以得到一个 Gamma 分布的随机变量 $Gamma(n, \lambda)$ 。

现在我们来看下形状参数和速率参数分别对概率分布函数的影响是怎样的。图 16.2.1 展示了形状参数 k 对概率分布函数的影响, 图 16.2.2 展示了速率参数 λ 对概率分布函数的影响。

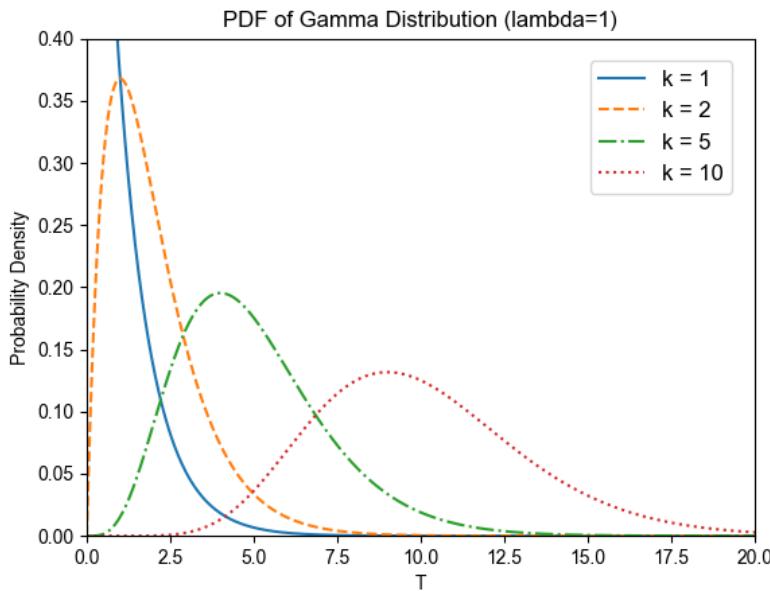


图 16.2.1: Gamma 分布不同 k 值下图形比较

通常 Gamma 分布的概率密度函数有多种参数化方式, 在计量经济学和其它一些自然科学领域, 经常使用形状参数和尺度参数进行参数化表示。

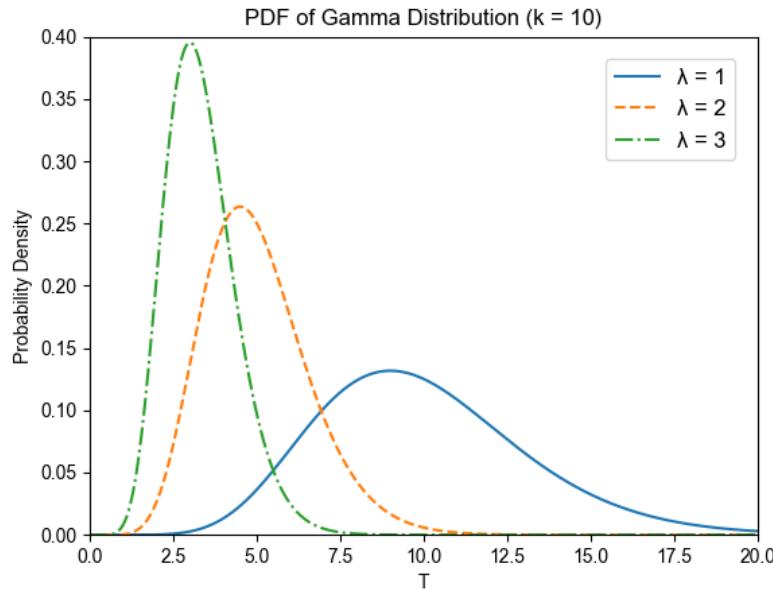
常见参数化方式 1:

令参数 $\alpha = k$ 表示形状参数, 参数 $\beta = 1/\lambda$ 表示尺度参数, 服从 Gamma 分布的随机变量 X 的概率密度函数为

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)} \quad (x, \alpha, \beta > 0)\tag{16.2.9}$$

这种形式常见于计量经济学和其它一些自然科学领域。这种形式下, Gamma 分布的期望和方差分别为

$$\begin{aligned}\mathbb{E}[X] &= \alpha\beta \\ Var(X) &= \alpha\beta^2\end{aligned}\tag{16.2.10}$$

图 16.2.2: Gamma 分布不同 λ 值下图形比较

常见参数化方式 2:

令参数 $\alpha = k$ 表示形状参数, 参数 $\beta = \lambda$ 表示尺度参数, 这种方式下, 尺度参数 β 和公式 (16.2.9) 是倒数关系, 这只是不同的参数化方法而已, 二者是等价的, 这时 Gamma 分布的概率分布函数写成

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} \quad (x, \alpha, \beta > 0) \quad (16.2.11)$$

此时的期望和方差分别为

$$\begin{aligned} \mathbb{E}[X] &= \frac{\alpha}{\beta} \\ Var(X) &= \frac{\alpha}{\beta^2} \end{aligned} \quad (16.2.12)$$

16.3 Gamma 回归模型

在 GLM 中, Gamma 模型用于响应数据只能取大于或等于 0 的连续值数据进行建模。

Gamma 分布是一个双参数分布, 包含形状参数 α 和尺度化参数 β , 形状参数 α 表示事件的发生次数, 尺度化参数 β 表示平均一次事件发生需要的时间。二者的乘积 $\mu = \alpha\beta$ 就是事件发生 α 次所需的平均时间, 即 Gamma 分布的期望值(均值)。

当把 Gamma 分布作为 GLM 家族的成员时, 需要把 α 看做一个已知的常量, 也就是人为给 α 设置一个常数, 并且对于所有观测样本都是一样的值。通常这个值一个经验值, 需要根据观测数据分布情况

GLM 中指数族分布的标准形式为

$$f(y; \theta, \phi) = \exp \left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (16.3.1)$$

现在我们把 Gamma 分布的概率密度函数转化成上述指数族分布的标准形式, 首先, 从上文已知 Gamma 分布的期望为 $\mu = \alpha\beta$ 。令 $\phi = 1/\alpha$, 则有 $\alpha = 1/\phi, \beta = \mu/\alpha = \mu\phi$, 代入到公式 (16.2.9) 可得

$$\begin{aligned}
 f(y; \mu, \phi) &= \frac{y^{\alpha-1} e^{-\frac{y}{\beta}}}{\beta^\alpha \Gamma(\alpha)} \\
 &= \exp \left[-\frac{y}{\beta} + (\alpha - 1) \ln y - \alpha \ln \beta - \ln \Gamma(\alpha) \right] \\
 &= \exp [-y/\beta - \alpha \ln \beta + (\alpha - 1) \ln y - \ln \Gamma(\alpha)] \\
 &= \exp \left[\frac{-y}{\mu\phi} - \frac{\ln(\mu\phi)}{\phi} + \left(\frac{1}{\phi} - 1 \right) \ln y - \ln[\Gamma(1/\phi)] \right] \\
 &= \exp \left[\frac{y(1/\mu)}{-\phi} - \frac{\ln \mu}{\phi} - \frac{\ln \phi}{\phi} + \left(\frac{1-\phi}{\phi} \right) \ln y - \ln[\Gamma(1/\phi)] \right] \\
 &= \exp \left[\frac{y(1/\mu)}{-\phi} + \frac{\ln \mu}{-\phi} - \frac{\ln \phi}{\phi} + \left(\frac{1-\phi}{\phi} \right) \ln y - \ln[\Gamma(1/\phi)] \right] \\
 &= \exp \left[\frac{y(1/\mu) - (-\ln \mu)}{-\phi} - \frac{\ln \phi}{\phi} + \left(\frac{1-\phi}{\phi} \right) \ln y - \ln[\Gamma(1/\phi)] \right]
 \end{aligned} \tag{16.3.2}$$

和公式 (16.2.9) 对比下, 可以直接得到各个重要组件的形式。

$$\begin{aligned}
 \theta &= 1/\mu \\
 b(\theta) &= -\ln(\mu) \\
 a(\phi) &= -\phi
 \end{aligned} \tag{16.3.3}$$

显然 Gamma 模型的标准链接函数就是倒数函数, $\eta = g(\mu) = \theta = 1/\mu$ 。现在我们看下 Gamma 分布的期望和方差函数。

$$\begin{aligned}
 b'(\theta) &= \frac{\partial b}{\partial \mu} \frac{\partial \mu}{\partial \theta} \\
 &= \left(-\frac{1}{\mu} \right) (-\mu^2) \\
 &= \mu \\
 b''(\theta) &= \frac{\partial^2 b}{\partial \mu^2} \left(\frac{\partial \mu}{\partial \theta} \right) + \frac{\partial b}{\partial \mu} \frac{\partial^2 \mu}{\partial \theta^2} \\
 &= (1)(-\mu^2) \\
 &= -\mu^2
 \end{aligned} \tag{16.3.4}$$

注意, $b''(\theta)$ 是方差函数, 体现的是方差和均值的关系, 显然 **Gamma 分布的方差是和其均值相关的**, Gamma 分布的方差为:

$$Var(y) = b''(\theta)a(\phi) = -\mu^2(-\phi) = \phi\mu^2 \tag{16.3.5}$$

最后我们整理一下 Gamma 模型的一些关键组件。

$$\begin{aligned}
 \text{标准链接函数: } \eta &= g(\mu) = \frac{1}{\mu} \\
 \text{反链接(响应)函数: } \mu &= r(\eta) = \frac{1}{\eta} \\
 \text{方差函数: } \nu(\mu) &= \mu^2 \\
 \text{分散函数: } a(\phi) &= -\phi \\
 \text{标准链接函数导数: } g' &= -\frac{1}{\mu^2}
 \end{aligned} \tag{16.3.6}$$

16.4 参数估计

16.4.1 似然函数

概率分布函数的指数族形式公式 (16.3.2) , 直接去掉底数就得到了其对数似然函数。

$$\ell(\mu, \phi; y) = \sum_{i=1}^N \left\{ \frac{y_i/\mu_i - (-\ln \mu_i)}{-\phi} + \frac{1-\phi}{\phi} \ln y_i - \frac{\ln \phi}{\phi} - \ln \Gamma(1/\phi) \right\} \quad (16.4.1)$$

根据公式 (8.1.12) , 标准链接函数的 Gamma 模型的似然函数的一阶偏导为

$$\begin{aligned} U_j &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i}{a(\phi)\nu(\mu_i)g(\mu_i)'} x_{ij} \\ &= - \sum_{i=1}^N \frac{y_i - \mu_i}{-\phi \mu_i^2 \eta_i^2} x_{ij} \end{aligned} \quad (16.4.2)$$

16.4.2 IRLS

只需要给出 W 和 Z 的计算等式就可以应用 IRLS 算法。

$$\begin{aligned} W &= \text{diag} \left\{ \frac{1}{a(\phi)\nu(\hat{\mu})(g')^2} \right\}_{(N \times N)} \\ &= \text{diag} \left\{ \frac{-\hat{\mu}^2}{\phi} \right\}_{(N \times N)} \end{aligned} \quad (16.4.3)$$

$$\begin{aligned} Z &= \{(y - \hat{\mu})g' + \eta\}_{(N \times 1)} \\ &= \left\{ \frac{-(y - \hat{\mu})}{\hat{\mu}^2} + \eta \right\}_{(N \times 1)} \end{aligned} \quad (16.4.4)$$

16.4.3 拟合优度

Gamma 模型的偏差统计量为

$$\begin{aligned} D &= 2\{\ell(y; y) - \ell(\hat{\mu}; y)\} \\ &= 2 \sum_{i=1}^N \frac{1 + \ln(y_i) - y_i/\hat{\mu}_i - \ln(\hat{\mu}_i)}{-\phi} \\ &= \frac{2}{\phi} \sum_{i=1}^N \left\{ \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \right\} \end{aligned} \quad (16.4.5)$$

Gamma 模型的皮尔逊卡方统计量为

$$\begin{aligned} \chi^2 &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\nu \hat{\mu}_i} \\ &= \sum_{i=1}^N \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2} \end{aligned} \quad (16.4.6)$$

16.5 其他连接函数

16.5.1 对数 Gamma 模型

前面我们提到, 在给定一组特定的解释变量或预测变量的情况下, 倒数链接估计模型响应的每单位速率。对数链接的 Gamma 表示响应的对数率。该模型规范与指数回归相同。当然, 这样的规范估计数据呈负指数下降。但是, 与生存分析中发现的指数模型不同, 我们不能将对数伽马模型用于审查数据。但是, 我们看到未经审查的指数模型可以符合 GLM 规范。我们将其保留到本章末尾。

对数伽玛模型, 与它的对等模型一样, 用于响应大于 0 的数据。几乎在每个学科中都可以找到示例。例如, 在健康分析中, 通常可以使用对数伽玛回归来估算住院天数 (LOS), 因为住院天数总是被约束为正数。LOS 数据通常使用泊松或负二项式回归进行估计, 因为 LOS 的元素是离散的。但是, 当存在许多 LOS 元素 (即许多不同的 LOS 值) 时, 许多研究人员发现伽马或高斯逆模型是可以接受的并且是更可取的。

在 GLM 之前, 通常使用对数转换响应的高斯回归来估计现在使用对数伽马技术估计的数据。尽管两种方法的结果通常相似, 但对数伽马技术不需要外部转换, 更易于解释, 并带有一组残差, 可用于评估模型的价值。因此, 对数伽马技术正在曾经使用过高斯技术的研究人员中得到越来越多的使用。

对数 gamma 模型是值其链接函数是对数函数, 其响应函数(反链接)是指数函数。

$$\begin{aligned}\eta &= \ln(\mu) \\ \mu &= e^\eta\end{aligned}\tag{16.5.1}$$

对数链接函数的一阶导数是 $g'(\mu) = 1/\mu$, 可以轻松的使用 IRLS 算法进行参数估计。但是由于对数链接函数表示规范链接函数, IRLS 算法的估计量和 ML 算法的估计量有不同的标准误差 (standard errors)。但是, 除了极端情况外, 标准误差的差异通常很小。在较大数据集时, 使用不同的估算方法通常不会产生任何推断差异。

16.5.2 恒等 (identity) Gamma 模型

恒等链接函数 $\eta = \mu$, 假设 μ 和 η 之间存在一一对应的关系。高斯模型的规范链接就是恒等函数, 但是 Gamma 模型中恒等函数不是规范链接。

在同一模型家族的不同链接之间进行选择有时可能很困难。McCullagh 和 Nelder (1989) 支持最小偏差的方法。他们还检查残差以观察拟合的紧密度以及标准化残差本身的正态性和独立性。在此示例中, 我们有两个我们要比较的非规范链接: 对数和恒等链接。在其他因素都相同的情况下, 最好选择偏差最小的模型。

我们也可以使用其他统计检验方法来评估链接之间的差异, 比如 BIC 和 AIC, 这些检验值较低的模型是更好一些的。关于 BIC, 如果两个模型的 BIC 之间的绝对差小于 2, 则两个模型的差异是比较小的。2 和 6 之间差值时, 两个模型之间就有了一定的区别。而 6 和 10 之间的差值, 则说明两个模型有了明显的区别。绝对差值大于 10, 那就肯定是 BIC 值较小的那个模型更好了。

过度分散

泊松回归模型是处理计数数据最基本的模型，应用最为广泛。然而泊松模型也不是万能的，它有一个局限性。泊松模型没有分散参数，并且它的方差和期望相同，这导致泊松模型无法处理方差和期望不相同的数据。本章我们讨论离散数据中一个常见的问题，即离散数据的方差大于它的期望，此时泊松模型无法很好的处理，需要借助其他的手段或者模型才行。

在 GLM 中，响应变量 Y 的方差 $V(Y)$ 等于分散函数 $a(\phi)$ 和方差函数 $\nu(\mu)$ 的乘积。

$$V(Y) = a(\phi)\nu(\mu) \quad (17.1)$$

其中方差函数 $\nu(\mu)$ 是一个关于期望 μ 的函数，其代表着方差 $V(Y)$ 和期望 μ 的关系。比如高斯模型，其方差函数为常量 $\nu(\mu) = 1$ ，说明高斯变量的方差与其期望是独立不相关的。反之，对于泊松模型，其方差函数为 $\nu(\mu) = \mu$ ，这意味着泊松模型的方差和期望存在关系。

分散函数(参数) $a(\phi)$ 相当于一个缩放 (scale) 因子，对 $\nu(\mu)$ 起到一个缩放的作用，通过这个缩放因子使得分布可以灵活适应(拟合)任何方差的数据。下表给出了 GLM 中常见分布的方差函数和分散函数，从中可以发现一些特点。

表 17.1: GLM 中的方差

分布	类型	分散函数 $a(\phi)$	方差函数 $\nu(\mu)$
Normal(Gaussian) $N(\mu, \sigma^2)$	连续分布	σ^2	1
Inverse Gaussian	连续分布	$-\sigma^2$	$-\mu^3$
Gamma	连续分布	$-\phi$	μ^2
Power	连续分布	xxx	μ^k
Bernoulli $B(\mu)$	离散分布	1	$\mu(1 - \mu)$
Binomial $B(n, \mu)$	离散分布	1	$\mu(1 - \mu/n)$
Poisson $Poisson(\mu)$	离散分布	1	μ
Categorical $Cat(K, \mu)$	离散分布	1	$\mu_k(1 - \mu_k)$
Negative binomial $NB(\mu, \alpha)$	离散分布	1	$\mu + \alpha\mu^2$

从上表中可以发现离散分布的分散函数都是常数，而其方差函数都是期望的函数，这表明离散分布的方差都是由它的期望决定的。由于没有分散参数，这就使得离散分布存在一个特有的现象，过度分散 (overdispersion)。

17.1 什么是过度分散

当训练好一个模型后, 利用模型的预测值 $\hat{\mu}$ 和分散参数 ϕ 就能计算出模型预测值的方差

$$V_e = a(\phi)\nu(\hat{\mu}_i) \quad (17.1.1)$$

模型预测值的方差称之为名义方差 (nominal variance), 用符号 V_e 表示。名义方差 V_e 就是模型 (概率分布) 的理论方差, 由方差函数计算得到的。与之相对应的, 是观测数据的真实方差, 用符号 V_o 表示, 代表观测 (observed) 方差。

$$V_o = V(Y_i) \quad (17.1.2)$$

名义方差 V_e 和观测方差 V_o 通常并不相同, 这和模型对数据的拟合程度相关, 模型对数据拟合的越好, 名义方差 V_e 和观测方差 V_o 就越接近。

模型的预测值是模型的期望值, $\hat{y}_i = \hat{\mu}_i$, 方差函数 $\nu(\hat{\mu}_i)$ 是期望值 $\hat{\mu}_i$ 的函数, 反映的是期望值, 分散函数 $a(\phi)$ 起到缩放和调节的作用, 显然, 对于名义方差 V_e 来说, 分散函数 $a(\phi)$ 的作用至关重要, 分散函数 $a(\phi)$ 使得名义方差 V_e 尽量接近和观测方差 V_o , 分散参数是用来控制模型方差的。然而, 对于离散模型, 没有分散参数 ϕ , 这就是使得名义方差 V_e 不能自由的适配任何数据的观测方差 V_o 。

当模型不存在分散参数时, 就无法拟合数据的真实方差, 就会发生 V_e 和 V_o 相差较大的现象。根据两者的关系, 有两种异常情况。

过度分散 (overdispersion) 当 V_o 大于 V_e 时, 称之为过度分散, 这种情况在离散数据中经常发生。

分散不足 (underdispersion) 当 V_o 小于 V_e 时, 称之为分散不足, 这种情况虽然理论上存在, 但实际中并不常见。因此本书我们重点讨论过度分散的问题, 分散不足的问题暂不讨论。

在实际应用中, 分散不足的情况很少见, 过度分散却是非常常见的, 尤其是对于计数数据。计数数据最基础的模型就是泊松模型, 而泊松模型的方差等于期望, 实际应用很少有数据满足这个假设。在实际的计数数据中通常都是方差大于期望的, 因此用传统的泊松模型处理计数数据必然会面临过度分散的问题, 而一旦发生过度分散, 就意味着模型无法拟合数据的实际方差。

通常在实际应用中遇到的真实数据, 并不是某个单一的、标准的概率分布产生的, 生成数据的真实分布是无从得知的。我们做的工作就是寻找一个尽可能接近数据真实分布的模型 (分布) 去拟合数据, 比如, 遇到连续值数据就用高斯分布; 遇到二值离散数据就用二项分布; 遇到计数数据就用泊松分布。**过度分散 (overdispersion) 或者分散不足 (underdispersion) 现象的本质就是: 我们选取的模型的理论方差和数据的真实方差是不相符的, 并且相差比较大。**

提示: 为什么只有离散数据会有过度分散的问题, 而连续值分布就没有呢?

这是因为连续值分布通常都有一个额外的分散参数 ϕ , 这个参数的作用就是缩放 (scale) 方差函数 $\nu(\mu)$, 使得模型的名义方差 V_e 可以更好的拟合观测样本的方差 V_o , 模型的分散参数 ϕ 就是专门用来拟合数据方差的参数, 然而离散分布模型缺少这样一个参数, 导致模型不能拟合数据的实际方差。

然而, 虽然连续值分布有分散参数 ϕ , 但这并不意味着连续值模型就不会发生过度分散现象, 如果使用不当, 同样也会导致过度分散。比如, 在应用传统的高斯模型 (线性回归) 时, 通常都是假设分散参数 ϕ 是常数 1, 这就相当于分散参数 ϕ 失去了作用, 如果响应数据的方差和这个假设并不相符, 同样会导致过度分散现象。

此外, 前面章节中我们介绍过, 可以通过皮尔逊卡方统计量 χ^2 除以其自由度的方式估计出分散参数 ϕ 。这种方式建立在分散参数与样本无关的假设之上, 即假设所有样本拥有相同的 ϕ 。如果样本方差和这个假设严重不符, 同样也会存在过度分散的可能, 只是这种严重不相符的情况很少见, 多数情况下, 这个方式足够用了。

17.2 过度分散的检测

离散分布之所以存在过度分散问题, 是因为离散分布的模型缺少一个分散参数 ϕ 去拟合数据的方差。那么我们可以先假设模型存在分散参数 ϕ , 此时观测方差 V_o 就等于 ϕ 乘上名义方差 V_e 。

$$V_o = \phi V_e \quad (17.2.1)$$

然后估计出 ϕ 的值, 如果 $\phi \approx 1$, 意味着名义方差 V_e 和观测方差 V_o 是近似相等的, 不存在过度分散, 或者说问题不严重; 反之, 如果 $\phi \gg 1$, 则意味着存在过度分散现象。在 [节 8.7](#) 和 [节 9.1.5](#) 讨论过, 可以通过皮尔逊卡方统计量估计分散参数 ϕ 。

$$\hat{\phi} = \frac{\chi^2}{N - p} \quad (17.2.2)$$

皮尔逊卡方统计量 χ^2 与其自由度 $(N - p)$ 的商称为皮尔逊分散统计量。皮尔逊分散统计量也是一个估计量, 并不是十分准确的, 因此一般情况下, 当皮尔逊分散统计量的值大于 2 时才认为发生了明显的过度分散。在早期的一些资料中, 检测过度分散也可以用偏差统计量, 然而后续的一些研究中发现, 皮尔逊卡方统计量更准确一些。

皮尔逊卡方统计量 χ^2 是模型拟合优度统计量, 本身可以用来评估模型拟合程度的。理论上模型对数据拟合的足够好时, χ^2 漸近服从自由度为 $N - p$ 的卡方分布, 它的期望值是 $N - p$ 。分散统计量的估计是建立在 χ^2 服从卡方分布并且期望为 $N - p$ 的假设基础上。如果模型对数据拟合的不够好, 那么计算出的 $\hat{\phi}$ 的也会远大于 1, 因此利用分散统计量判断是否存在过度分散并不是十分可靠的。通常我们把由于模型欠拟合导致的过度分散称为表象过度分散 (apparent overdispersion), 真正的数据过度分散称为真实过度分散 (real overdispersion)。

导致表象过度分散的直接原因是模型对数据拟合的不好, 而导致模型拟合不好的原因常见的有如下几种情况:

- 数据包含异常点。
- 缺少有效的特征。
- 缺少高阶的组合特征。
- 特征数据没有进行归一化处理。
- 链接函数不正确。

表象过度分散不同于真实过度分散, 表象过度分散仅仅是由于模型对数据拟合的不好导致, 只要针对性的解决上述问题, 并提高模型的拟合程度后就会消除过度分散现象。而真实过度分散是无法通过提高模型拟合程度解决的, 真实过度分散是数据的真实方差和模型的方差假设相违背, 显然要想解决真实过度分散就要改变模型的方差假设, 也就是要更换另一种新的分布模型才行。下一章要讨论的负二项式模型就是一种全新的计数数据分布模型, 其方差函数是 $\mu + \alpha\mu^2$, 不再像泊松分布一样限定方差和期望相等, 而是多了一个辅助参数 α 来适配数据的方差变化。

17.3 过度分散的影响

在讲 GLM 的最大似然估计时提到过, 分散参数 ϕ 并不影响协变量参数 β 的迭代, 这里从 IRLS 算法的过程再验证一下。IRLS 算法参数的迭代公式为

$$\begin{aligned} \beta &= (X^T W X)^{-1} X^T W Z \\ &= \frac{X^T W Z}{X^T W X} \end{aligned} \quad (17.3.1)$$

其中权重矩阵 W 是一个对角矩阵, 其值为

$$W = [a(\phi)\nu(\mu)g'^2]^{-1} = [V_e(\mu)g'^2]^{-1} \quad (17.3.2)$$

其中 $a(\phi) = \phi$ 是与样本无关的, 因此可以单独提取出来, 则有

$$W = [a(\phi)\nu(\mu)g'^2]^{-1} = \phi^{-1}W_o \quad (17.3.3)$$

注解: 在 $a(\phi) = \phi/w_i$ 的情形下, 同样可以单独提出分散参数 ϕ , 而把样本权重 w_i 留在 W_o 中, 结论是一样的。

代入到公式 (17.3.1) 可得

$$\begin{aligned} \beta &= \frac{X^T W Z}{X^T W X} \\ &= \frac{\phi^{-1} X^T W_o Z}{\phi^{-1} X^T W_o X} \\ &= \frac{X^T W_o Z}{X^T W_o X} \end{aligned} \quad (17.3.4)$$

可以看到在 β 的迭代过程中, 分散参数 ϕ 并没有起到作用, 有没有分散参数 ϕ 不影响参数 β 的更新。因此过度分散问题并不影响协变量参数 β 的估计。

虽然过度分散不影响 β 的迭代, 但是会影响它的标准误。参数估计量 β 的标准误的计算

$$\begin{aligned} SE(\hat{\beta}) &= \sqrt{\mathcal{J}^{-1}} \\ &= \sqrt{(X^T W X)^{-1}} \\ &= \sqrt{(X^T [a(\phi)\nu(\hat{\mu})g'^2]^{-1} X)^{-1}} \\ &= \sqrt{(X^T [V_e(\hat{\mu})g'^2]^{-1} X)^{-1}} \end{aligned} \quad (17.3.5)$$

可以看出协变量参数估计量 $\hat{\beta}$ 的标准误差的计算是依赖模型方差的, 如果模型的名义方差比观测方差小, 就会导致计算出的标准误 $SE(\hat{\beta})$ 比真实值偏小, 这会使得模型的假设检验失效。

17.4 标准误差的修正

既然清楚了过度分散的影响, 就可以针对性的去解决它, 显然最直接到的方法就是人为给模型添加一个分散参数 ϕ 。因为分散参数不影响 IRLS 算法的迭代结果, 所以可以不用修改 IRLS 算法过程, 只需要在 IRLS 算法结束后修正一下估计量的标准误差。分散参数 ϕ 的值就用皮尔逊卡分散统计量来表示, 修正后的标准误差为

$$SE(\hat{\beta}) = \sqrt{\hat{\phi}_{\chi^2}(X^T W X)^{-1}} \quad (17.4.1)$$

在对标准误差进行缩放后, 该模型不再是似然模型, 因为标准误差既不是基于预期的信息矩阵, 也不是基于观测的信息矩阵。修正标准误差的方法还有另外几种的方法, 比如, William 过程、稳健方差估计等, 这里不在详细介绍, 有兴趣的读者可以参考其它资料。

事实上, 仅仅修正标准误并没有从本质上解决过度分散问题, 过度分散的本质原因是模型假设和观测数据不匹配, 因此最根本的解决方法是更换更适合数据的模型, 下两章介绍的负二项式模型、零截断模型、零膨胀模型就是从模型根本上解决过度分散的问题。

负二项式模型

在计数数据中，应用最广泛的是泊松模型，而泊松模型假设模型的方差和期望相等， $V(Y) = \mu$ 这在多数情况下是无法满足的，大多数计数数据的方差和期望是不相等的，比较常见的是方差大于期望，称之为过度分散现象。面对过度分散的数据，传统的泊松模型就不太适合了。本周讨论另一种计数模型，负二项式模型，它的方差为 $V(Y) = \mu + \alpha\mu^2$ ，其模型方差和期望不再是相等的，而是多了一个可变的参数 α ，通过 α 可以调整方差和期望的关系，进而能适应各种方差的计数数据，负二项式分布可以更好地拟合过度分散的数据。

18.1 负二项式分布

回顾指数族的各种分布，可以发现很多分布都和伯努利分布有关，二项式分布描述的是 n 次伯努利实验成功次数的概率分布，而泊松分布又是二项式分布的极限形式，描述的是单位时间窗口内事件发生次数的概率分布。指数分布又可以从泊松分布推导而来，描述的是下一次事件发生需要等待的时间。指数分布的更一般化就得到了 Gamma 分布。本章讨论的负二项式分布，从名字就可以看出一定是和二项式分布有关。事实上，二项式分布、几何分布、负二项分布都是伯努利实验的扩展。二项式分布描述的是 n 次实验中成功的次数；几何分布描述的是第一次成功之前失败的次数；负二项分布描述的是第 r 次成功之前失败的次数。二项式分布、泊松分布、几何分布、负二项式分布都是建立在伯努利试验过程的基础上，几何分布是负二项式分布的一个特例。

负二项式分布可以有多种定义(理解)方式，有两种可以看做是二项式分布变种，另一种是可以看做是泊松-伽马混合模型。关于负二项分布的“负”有多种解释，最直观的一种就是，二项式分布描述的“成功”的次数，而负二项分布描述的是“失败”的次数，因此是“负”。

18.1.1 从二项式分布推导

二项式分布表示的时进行 n 次伯努利实验 成功的次数 的概率分布, 其概率分布函数为

$$f(y; n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad (18.1.1)$$

其中 π 是单次伯努利实验成功的概率, $1 - \pi$ 是伯努利实验失败的概率。符号 $\binom{n}{y}$ 是组合数 C_n^y , 表示 n 次实验中有任意 y 次成功的全部组合数。二项式分布的期望和方差分别是 $E[Y] = n\pi$ 和 $V(Y) = n\pi(1 - \pi)$ 。负二项式分布可以看做是二项式分布的变种, 并且有两种定义方式, 事实上这两种定义方式是等价的。

首先回顾一些组合数计算的转换公式, 其一有,

$$C_{a+b}^a = C_{a+b}^b \quad (18.1.2)$$

其二, 组合数的计算可以转换成多个 Gamma 函数的计算。Gamma 函数是阶乘在实数域的扩展, $\Gamma(n+1) = n!$, 因此有

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \quad (18.1.3)$$

第一种定义方式

随机变量 Y 表示, 伯努利实验中要得到成功 r 次的结果, 需要进行的实验总次数, r 必须是一个正整数。注意, 一共进行了 y 次实验, 最后一次实验一定是成功的, 并且是第 r 次成功。因此只需要在前面 $y-1$ 次实验中任意成功 $r-1$ 次即可, 这符合二项式分布的定义, 根据二项式分布的概率分布函数公式 (18.1.1) 可以得到 $y-1$ 次实验中成功 $r-1$ 的概率为 $Bin(y-1, r-1)$, 第 y 次是成功的, 其概率是 π , 因此实验总次数的变量 Y 的概率分布函数为

$$\begin{aligned} f(y; r, \pi) &= \underbrace{Bin(y-1, r-1)}_{\text{前 } (y-1) \text{ 次}} \times \underbrace{\pi}_{\text{第 } y \text{ 次}} \\ &= \binom{y-1}{r-1} \pi^{r-1} (1 - \pi)^{y-r} \pi \\ &= \binom{y-1}{r-1} \pi^r (1 - \pi)^{y-r} \\ &= \frac{\Gamma(y)}{\Gamma(r)\Gamma(y-r-1)} \pi^r (1 - \pi)^{y-r} \end{aligned} \quad (18.1.4)$$

第二种定义方式

变量 Y 表示事件第 r 次成功前失败的次数。注意是一共成功了 r 次, 在这之前失败了 y 次, 实验总次数是 $y+r$ 。由于第 $y+r$ 次一定是成功的, 故只要在前面的 $y+r-1$ 次中找出成功的 $r-1$ 次的组合次数即可, 这符合二项式分布 $Bin(y+r-1, r-1)$, 最终变量 Y 的概率分布为

$$\begin{aligned} f(y; r, \pi) &= \underbrace{Bin(y+r-1, r-1)}_{\text{前 } (y+r-1) \text{ 次}} \times \underbrace{\pi}_{\text{第 } (y+r) \text{ 次}} \\ &= \binom{y+r-1}{r-1} \pi^{r-1} (1 - \pi)^y \pi \\ &= \binom{y+r-1}{r-1} \pi^r (1 - \pi)^y \\ &= \binom{y+r-1}{y} \pi^r (1 - \pi)^y \\ &= \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \pi^r (1 - \pi)^y \end{aligned} \quad (18.1.5)$$

无论哪种方式, r 的含义都是一样的, 公式 (18.1.5) 是公式 (18.1.4) 的平移, 相当于把 y 平移到 $y+r$, 二者的概率分布曲线是完全一致的, 只是对变量 Y 的定义有些差异, 但这不影响使用, 二者可以看做是等价的。

18.1.2 泊松-伽马混合分布

假设变量 Y 是泊松变量, 其概率质量函数为

$$P(Y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad (18.1.6)$$

其中唯一的参数 λ 是泊松分布的期望参数, 在本书之前的所有内容中, 都是假设概率分布的参数是数值参数, 不是随机量。现在我们假设泊松分布的期望参数 λ 也是一个随机量, 并且它的概率分布是 Gamma 分布, 我们采用 Gamma 分布公式 (16.2.11) 的参数化方法, 两个参数分别是形状 (shape) 参数 α 和尺度 (scale) 参数 β 。

$$f(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (\lambda, \alpha, \beta > 0) \quad (18.1.7)$$

此时响应变量 Y 和参数变量 λ 的联合概率分布为 $p(Y, \lambda) = p(\lambda)p(Y|\lambda)$, 需要通过边缘化 (积分消掉参数变量) 的方法得到边缘概率 $p(Y)$ 。

$$P(Y) = \int_0^\infty P(\lambda)P(Y|\lambda)d\lambda \quad (18.1.8)$$

把公式 (18.1.6) 和公式 (18.1.7) 代入到公式 (18.1.8), 并计算积分得到变量 Y 边缘概率分布函数。

$$\begin{aligned} f(y) &= \int_0^\infty \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \frac{\lambda^y e^{-\lambda}}{y!} d\lambda \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)y!} \lambda^{\alpha-1} e^{-\beta\lambda} \lambda^y e^{-\lambda} d\lambda \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)y!} \lambda^{\alpha+y-1} e^{-(\beta+1)\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \int_0^\infty \lambda^{\alpha+y-1} e^{-(\beta+1)\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \int_0^\infty \frac{\Gamma(\alpha+y)}{(\beta+1)^{\alpha+y}} \frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)} \lambda^{\alpha+y-1} e^{-(\beta+1)\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \frac{\Gamma(\alpha+y)}{(\beta+1)^{\alpha+y}} \int_0^\infty \underbrace{\frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)} \lambda^{\alpha+y-1} e^{-(\beta+1)\lambda}}_{\text{积分为 1}} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)y!} \frac{\Gamma(\alpha+y)}{(\beta+1)^{\alpha+y}} \\ &= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta+1)^\alpha (\beta+1)^y} \\ &= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y \end{aligned} \quad (18.1.9)$$

现在重新参数化公式 (18.1.9), 令 $r = \alpha, \pi = \beta/(\beta+1)$, 则公式 (18.1.9) 可以转化成

$$f(y; r, \pi) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \pi^r (1-\pi)^y \quad (18.1.10)$$

这和公式 (18.1.5) 完全一样的, 可见泊松-伽马混合模型本质上就是负二项式分布。

18.1.3 辅助参数 α 的影响

负二项式分布有两个参数, 一个是期望参数 μ , 一个是辅助参数 α , 根据前面的推导过程可知辅助参数 α 一定是大于 0 的, 首先给 α 赋值一个接近 0 的值, 观察下当 $\alpha \approx 0$ 时负二项式分布的特点。图 18.1.1 是 $\alpha = 0.001$ 时负二项式分布概率质量函数的曲线图。

注解: 注意, 负二项式分布是离散变量的分布, 离散分布的概率分布函数称为概率质量函数, 概率质量函数应该是离散的点图, 为了方便观测概率的变化规律, 我们把点图用线连接起来, 观察曲线的变化。

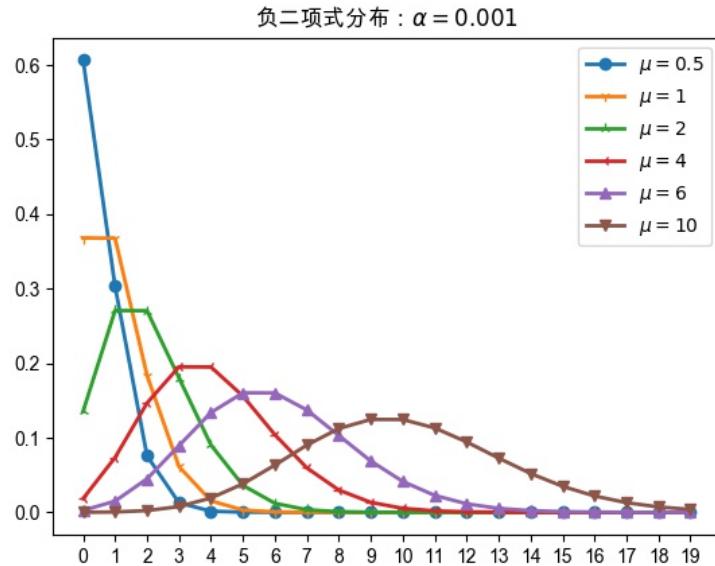


图 18.1.1: 当 $\alpha = 0.001$ 时, 不同 μ 值下负二项式分布的概率质量函数

对比下图 18.1.1 和泊松分布的概率质量函数的分布图(图 14.1.1), 可以发现二者基本是一致的。因此负二项式分布的辅助参数 $\alpha = 0$ 时, 就等价于泊松分布, 泊松分布可以看做是负二项式分布的一个特例。

现在我们逐渐增大 α 的值, 图 18.1.2 是 $\alpha = 0.33$ 的分布图, 可以看出曲线上凸起的部分向左移动了一些。我们继续增大 α 的值, 如图 18.1.3, 把 α 设置为 0.67, 可以发现原来凸起的部分逐渐消失。

把 α 的值进一步增大到 1.0, 变成了图 18.1.3 所示的图形, 图形基本都变成了下凹的形状。 $\alpha = 1.0$ 的负二项式分布又叫做几何 (geometric) 分布, 几何分布是负二项式分布的一个特例。对比下图 18.1.4 和指数分布的图形(图 15.1.1), 可以发现二者的图形几乎是一致的。事实上几何分布是指数分布的离散版本, 指数分布是一个连续值概率分布, 而几何分布与指数分布是离散相关的。

注解: 指数分布是连续值概率分布, 连续值随机变量的概率分布函数叫做概率密度函数 (probability density function, pdf); 离散随机变量的概率分布函数叫做概率质量函数 (probability mass function, pmf)。概率密度函数的曲线图纵坐标不是概率值, 需要积分才能得到概率值; 而概率质量函数的曲线图纵坐标直接就是对应的概率值。

继续增大 α 的值, 如图 18.1.5 和图 18.1.6 所示, 图形的左下角会越来越凹陷, 并且随着 α 的增加, 各个不同的期望 μ 值的曲线会逐渐重合在一起, 这意味着当 α 足够大时, 期望参数 μ 的影响力逐渐变小。

最后我们固定 μ 的值, 直接对比不同的 α 下图形的差异。图 18.1.7 是 μ 固定为 4.0, α 取不同值时负二项式分布的图形。可以看出, α 等于 0 时, 负二项式分布在期望值附近的概率时最大的, 而随着 α 增大, 负二项

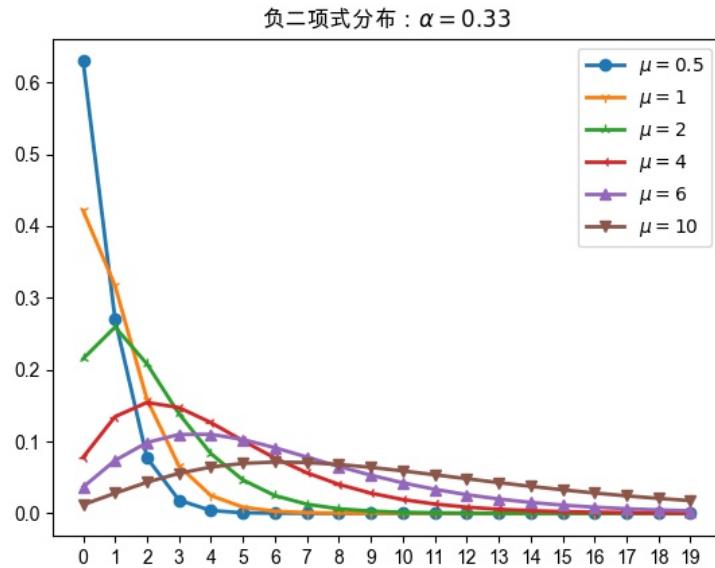


图 18.1.2: 当 $\alpha = 0.33$ 时, 不同 μ 值下负二项式分布的概率质量函数。相比于 $\alpha = 0.001$ 时, 曲线凸起的部分向左移动了。

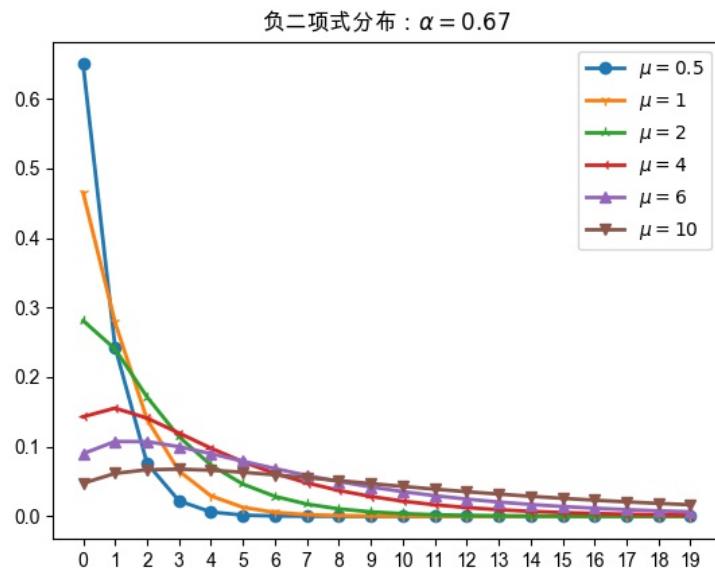


图 18.1.3: 当 $\alpha = 0.67$ 时, 图形左侧凸起逐渐消失了。

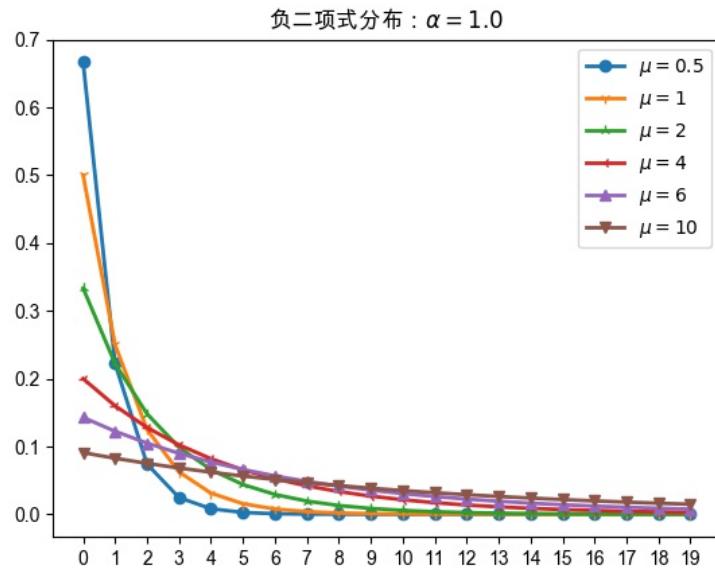


图 18.1.4: 当 $\alpha = 1.0$ 时, 负二项式分布又称为几何分布, 并且和指数分布是离散相关的。

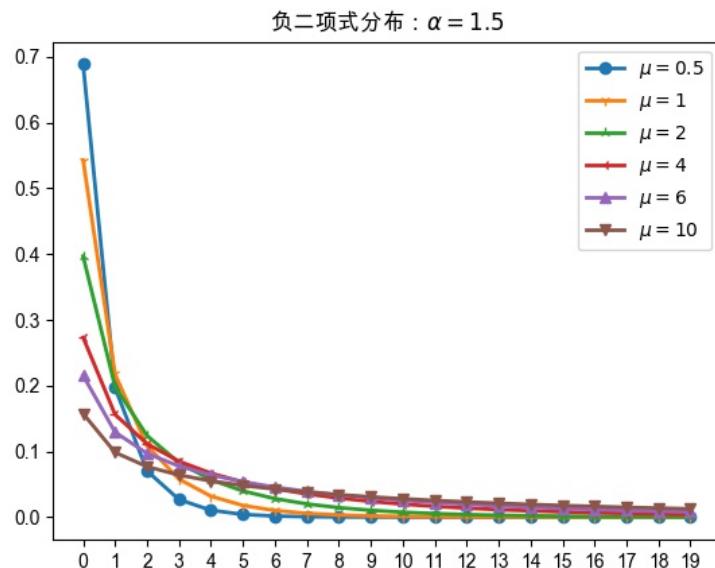


图 18.1.5: 当 $\alpha = 1.0$ 时, 不同 μ 值下负二项式分布的概率质量函数

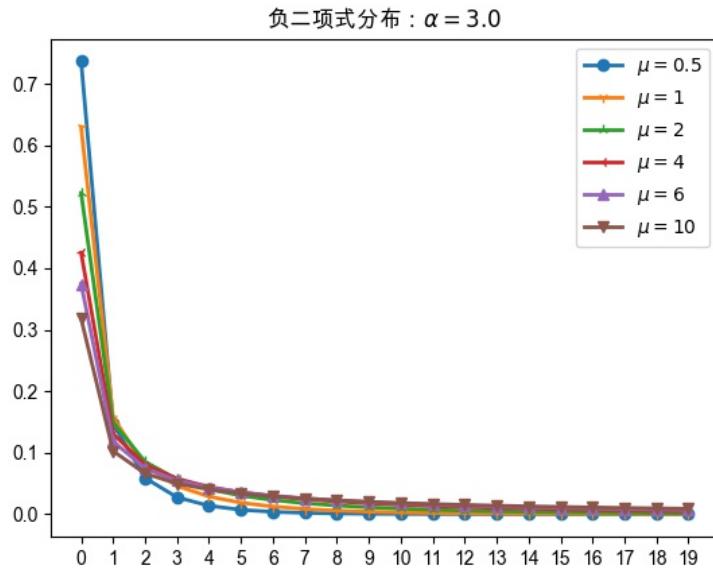


图 18.1.6: 当 $\alpha = 1.0$ 时, 不同 μ 值下负二项式分布的概率质量函数

式分布中 0 的概率逐步增大。

反过来, 固定 α 的值, μ 越大, 0 的概率就越小, 图形就越接近高斯分布。最后我们总结下负二项式分布中 μ 和 α 的关系。

- 固定 μ , α 越大, 0 的概率越大。
- 固定 α , μ 越大, 0 的概率越小。

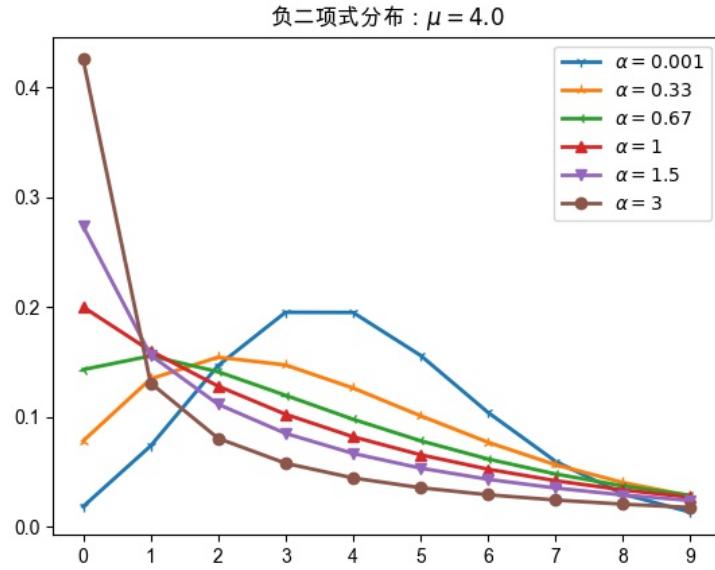
18.2 负二项回归模型

负二项式分布同样属于指数族分布, 因此负二项式回归模型可以纳入到 GLM 框架中, 作为 GLM 的一员。先把公式 (18.1.10) 作为负二项式分布的概率分布函数, 并把它转化成指数族的形式。

$$\begin{aligned} f(y; r, \pi) &= \exp \left\{ \ln \left[\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \pi^r (1-\pi)^y \right] \right\} \\ &= \exp \left\{ y \ln(1-\pi) + r \ln \pi + \ln \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \right\} \end{aligned} \quad (18.2.1)$$

因此, 有

$$\begin{aligned} \text{自然参数} \quad \theta &= \ln(1-\pi) \\ \pi &= 1 - e^\theta \\ \text{累积分布函数} \quad b(\theta) &= -r \ln \pi = -r \ln(1 - e^\theta) \\ \text{分散函数} \quad a(\phi) &= 1 \end{aligned} \quad (18.2.2)$$

图 18.1.7: 固定 $\mu = 4.0$, 不同 α 的值对图形的影响。

现在来看下负二项式分布的期望与方差, 指数族分布的期望与方差函数可以分别由累积分布函数 $b(\theta)$ 的一阶导和二阶导得到。

$$\begin{aligned}
 b'(\theta) &= \frac{\partial b}{\partial \pi} \frac{\partial \pi}{\partial \theta} \\
 &= \left(-\frac{r}{\pi} \right) \{ -(1 - \pi) \} \\
 &= \frac{r(1 - \pi)}{\pi} \\
 &= \mu
 \end{aligned} \tag{18.2.3}$$

$$\begin{aligned}
 b''(\theta) &= \frac{\partial^2 b}{\partial \pi^2} \left(\frac{\partial \pi}{\partial \theta} \right)^2 + \frac{\partial b}{\partial \pi} \frac{\partial^2 \pi}{\partial \theta^2} \\
 &= \left(\frac{r}{\pi^2} \right) (1 - \pi)^2 + \frac{r}{\pi} (1 - \pi) \\
 &= \frac{r(1 - \pi)^2 + r\pi(1 - \pi)}{\pi^2} \\
 &= \frac{r(1 - \pi)}{\pi^2} \\
 &= \mu + \frac{\mu^2}{r} \\
 &= \nu(\mu)
 \end{aligned} \tag{18.2.4}$$

有了方差函数后, 可得到方差为

$$V(Y) = a(\phi)\nu(\mu) = \mu + \frac{\mu^2}{r} \tag{18.2.5}$$

这个形式下的方差函数, 参数 r 在分母的位置, 不是很“美观”, 通常情况下会重新参数化一下, 令 $\alpha = 1/r$, 使用参数 α 重新参数化负二项式模型后, 负二项式分布的期望和方差分别为

$$\begin{aligned} \text{期望} \quad \mu &= \frac{1 - \pi}{\alpha\pi} \\ \pi &= \frac{1}{1 + \alpha\mu} \\ \text{方差} \quad V(Y) &= \mu + \alpha\mu^2 \end{aligned} \tag{18.2.6}$$

负二项式分布的概率分布函数公式 (18.2.1) 用期望参数 μ 和参数 α 重新参数化后为

$$\begin{aligned} f(y; \alpha, \mu) &= \exp \left\{ y \ln(1 - \pi) + r \ln \pi + \ln \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)} \right\} \\ &= \exp \left\{ y \ln \left(\frac{\alpha\mu}{1 + \alpha\mu} \right) - \frac{1}{\alpha} \ln(1 + \alpha\mu) + \ln \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \right\} \end{aligned} \tag{18.2.7}$$

公式 (18.2.7) 是负二项式模型的标准指数族形式, 其标准连接函数为

$$\eta = \theta = \ln \left(\frac{\alpha\mu}{1 + \alpha\mu} \right) \tag{18.2.8}$$

标准连接函数的响应函数为

$$\mu = \frac{\exp(\eta)}{\alpha[1 - \exp(\eta)]} \tag{18.2.9}$$

采用标准连接函数的负二项式模型通常简称为 NB-C 模型,

因此有

$$\begin{aligned} \text{自然参数} \quad \theta &= \ln \left(\frac{\alpha\mu}{1 + \alpha\mu} \right) \\ \text{累积分布函数} \quad b(\theta) &= \frac{1}{\alpha} \ln(1 + \alpha\mu) \\ \text{期望} \quad b'(\theta) &= \mu \\ \text{方差函数} \quad \nu(\mu) &= b''(\theta) = \mu + \alpha\mu^2 \\ \text{分散函数} \quad a(\phi) &= 1 \\ \text{方差} \quad V(Y) &= a(\phi)\nu(\mu) = \mu + \alpha\mu^2 \\ \text{标准连接函数} \quad g(\mu) &= \ln \left(\frac{\alpha\mu}{1 + \alpha\mu} \right) \\ \text{标准连接函数一阶导} \quad g'(\mu) &= \frac{1}{\mu + \alpha\mu^2} \\ \text{响应函数} \quad \mu &= \frac{\exp(\eta)}{\alpha[1 - \exp(\eta)]} \end{aligned} \tag{18.2.10}$$

18.3 参数估计

参数 α 通常被称为辅助 (ancillary) 参数或者尺度 (scale) 参数, 只有 α 是一个常数的时, 负二项式回归模型才能纳入到 GLM 的框架下, 这是因为 GLM 框架的 IRLS 算法只能估计协变量参数 β , 无法同时估计出额外的参数, 需要在应用 IRLS 算法前通过其它的方法确定 α 的值, 然后把 α 的值代入到 GLM 中, 当做一个常量值。我们先给出 NB-C 模型的 IRLS 算法过程, 然后再讨论如何用最大似然估计同时估计参数 α 和协变量参数 β 。

18.3.1 IRLS

NB-C 模型的概率质量函数为公式 (18.2.7) , 其对数似然函数为

$$\ell(\mu; y, \alpha) = \sum_{i=1}^N \left\{ y_i \ln \left(\frac{\alpha \mu_i}{1 + \alpha \mu_i} \right) - \frac{1}{\alpha} \ln(1 + \alpha \mu_i) + \ln \Gamma(y_i + 1/\alpha) - \ln \Gamma(y_i + 1) - \ln \Gamma(1/\alpha) \right\} \quad (18.3.1)$$

NB-C 模型采用的标准链接函数, 标准链接函数为

$$g(\mu) = \ln \left(\frac{\alpha \mu}{1 + \alpha \mu} \right) = -\ln[1 + 1/(\alpha \mu)] \quad (18.3.2)$$

标准链接函数的导数为

$$g'(\mu) = \frac{1}{\mu + \alpha \mu^2} \quad (18.3.3)$$

依此可以给出 IRLS 算法过程中的权重矩阵 W 和工作响应矩阵 Z , 分别为

$$\begin{aligned} W_{ii} &= \frac{1}{a(\phi)\nu(\hat{\mu}_i)(g')^2} \\ &= \hat{\mu}_i + \alpha \hat{\mu}_i^2 \end{aligned} \quad (18.3.4)$$

$$\begin{aligned} Z_i &= (y_i - \hat{\mu}_i)g' + \eta_i \\ &= \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i + \alpha \hat{\mu}_i^2} + \eta_i \end{aligned} \quad (18.3.5)$$

偏差统计量为

$$D = 2 \sum_{i=1}^N \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - \left(y_i + \frac{1}{\alpha} \right) \ln \left(\frac{1 + \alpha y_i}{1 + \alpha \hat{\mu}_i} \right) \right\} \quad (18.3.6)$$

NB-C 模型 IRLS 算法过程:

```

 $\mu = \{y + \text{mean}(y)\}/2$ 
 $\eta = -\ln\{1/(\alpha \mu) + 1\}$ 
 $\text{ WHILE ( abs( } \Delta \text{Dev} \text{ ) > tolerance) } \{$ 
     $W = \mu + \alpha \mu^2$ 
     $Z = \{\eta + (y - \mu)/W\}$ 
     $\beta = (X^T W X)^{-1} X^T W Z$ 
     $\eta = X \beta$ 
     $\mu = 1/\{\alpha(e^{-\eta} - 1)\}$ 
     $\text{OldDev}=\text{Dev}$ 
     $\text{Dev} = 2 \sum \{y \ln(y/\mu) - (y + 1/\alpha) \ln[(1 + \alpha y)/(1 + \alpha \mu)]\}$ 
     $\Delta \text{Dev} = \text{Dev} - \text{OldDev}$ 
}
 $\chi^2 = \sum [(y - \mu)^2 / (\mu + \alpha \mu^2)]$ 

```

```

 $\mu = (y - \text{mean}(y)) / 2$            /// initialization
 $\eta = -\ln(1/\alpha\mu + 1)$            /// canonical link
WHILE(ABS( $\Delta\text{Dev}$ ) > tolerance {
 $w = \mu + \alpha\mu^2$            /// variance
 $z = \eta + (y - \mu)/w - \text{offset}$ 
 $\beta = (X'wX)^{-1}X'wz$ 
 $\eta = X'\alpha + \text{offset}$            /// linear predictor
 $\mu = 1/(\alpha(\exp(-\eta) - 1))$            /// inverse link
oldDev = Dev
Dev =  $2\sum\{y\ln(y/\mu) - (y + 1/\alpha)\ln((1 + \alpha y)/(1 + \alpha\mu))\}$ 
 $\Delta\text{Dev} = \text{Dev} - \text{oldDev}$ 
}

```

图 18.3.1: NB-C 模型的 IRLS 算法过程

18.3.2 参数 α 的估计

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^N \frac{y_i - \mu_i}{\mu_i(1 + \alpha\mu_i)} \quad (18.3.7)$$

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N \frac{x_{ij}(y_i - \mu_i)}{1 + \alpha\mu_i} \quad (18.3.8)$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^N \left\{ \frac{1}{\alpha^2} \left[\ln(1 + \alpha\mu_i) + \frac{\alpha(y_i - \mu_i)}{1 + \alpha\mu_i} \right] + \psi\left(y_i + \frac{1}{\alpha}\right) - \psi\left(\frac{1}{\alpha}\right) \right\} \quad (18.3.9)$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^N \left[-x_{ij}x_{ik} \frac{\mu_i(1 + \alpha y_i)}{(1 + \alpha\mu_i)^2} \right] \quad (18.3.10)$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \alpha} = \mathbb{E} \left[- \sum_{i=1}^N \frac{\mu_i(y_i - \mu_i)x_{ij}}{(1 + \alpha\mu_i)^2} \right] \quad (18.3.11)$$

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{i=1}^N & \left\{ -\frac{1}{\alpha^3} \left[\frac{\alpha(1 + 2\alpha\mu)(y_i - \mu_i) - \alpha\mu_i(1 + \alpha\mu_i)}{(1 + \alpha\mu_i)^2} + 2\ln(1 + \alpha\mu_i) \right] \right. \\ & \left. + \psi'\left(y_i + \frac{1}{\alpha}\right) - \psi'\left(\frac{1}{\alpha}\right) \right\} \end{aligned} \quad (18.3.12)$$

18.4 负二项式模型扩展

18.4.1 对数连接函数

采用对数连接的负二项式模型简称为 NB-2 模型, $\alpha = 0$ 的 NB-2 模型就等价于泊松模型。传统的泊松回归模型经常会面临过度分散的问题, NB-2 模型通常会看做是泊松模型的“改进版”, 是处理泊松过度分散数据最常用的方法, 而标准连接的 NB-C 模型很少使用。

把 NB-C 模型转换成 NB2 模型, 只需要更换连接函数和响应函数(反连接函数)即可。

- 连接函数: $\eta = \ln(\mu)$
- 响应函数: $\mu = \exp(\eta)$

NB2 模型的概率质量函数的指数族形式为

$$f(y; \alpha, \mu) = \exp \left\{ y \ln(\mu) - \frac{1}{\alpha} \ln(1 + \alpha\mu) + \ln \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \right\} \quad (18.4.1)$$

NB2 模型的偏差统计量为

$$D = 2 \sum_{i=1}^N \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - \frac{1}{\alpha} \ln \left(\frac{1 + \alpha y_i}{1 + \alpha \hat{\mu}_i} \right) \right\} \quad (18.4.2)$$

NB2 模型的 IRLS 算法过程和 NB-C 模型几乎没有差别, 只是把对应的连接函数及其导数部分替换一下即可。NB2 模型关键组件为

$$\begin{aligned} \text{连接函数} \quad & \eta = g(\mu) = \ln(\mu) \\ \text{连接函数一阶导} \quad & g'(\mu) = 1/\mu \\ \text{连接函数二阶导} \quad & g''(\mu) = -1/\mu^2 \\ \text{方差} \quad & V(\mu) = a(\phi)\nu(\mu) = \mu + \alpha\mu^2 \\ \text{方差一阶导} \quad & V'(\mu) = 1 + 2\alpha\mu \\ \text{方差的平方} \quad & V^2(\mu) = (\mu + \alpha\mu^2)^2 \end{aligned} \quad (18.4.3)$$

IRLS 算法过程的权重 W 和工作响应 Z 分别为

$$\begin{aligned} W &= \text{diag} \left\{ \frac{1}{V(\hat{\mu})(g')^2} \right\}_{(N \times N)} \\ &= \text{diag} \left\{ \frac{\mu}{1 + \alpha\mu} \right\}_{(N \times N)} \end{aligned} \quad (18.4.4)$$

$$\begin{aligned} Z &= \{(y - \hat{\mu})g' + \eta\}_{(N \times 1)} \\ &= \left\{ \frac{(y - \hat{\mu})}{\hat{\mu}} + \eta \right\}_{(N \times 1)} \end{aligned} \quad (18.4.5)$$

在传统的 GLM 算法中, 参数估计算法 IRLS 使用的是期望信息矩阵 EIM, 对于采用标准连接函数的 NB-C 模型来说, 期望信息矩阵 EIM 和观测信息矩阵 OIM 是等价的。然而 NB-2 模型是非标准连接函数, 此时期望信息矩阵 EIM 和观测信息矩阵 OIM 不再相等。在 NB-2 的参数估计过程中, 要想利用观测信息矩阵 OIM 计算参数的标准误, 就需要对 IRLS 算法做一些改动。

注解: 回顾一下, GLM 的参数估计算法, 传统的最大似然估计是利用观测信息矩阵 OIM 计算参数估计量的标准误, IRLS 算法利用期望信息矩阵 EIM 替代了观测信息矩阵 OIM, 而利用 EIM 计算出的参数估计量标

准误是有一定误差的。对于采用标准连接函数的 GLM 模型, OIM 和 EIM 是等价的, 使用 EIM 也没有关系。然而非标准连接的 GLM 模型, OIM 和 EIM 是不一样的, 此时要想得到准确的参数估计量标准误, 就需要对 IRLS 做一些改动。

NB-2 模型和泊松模型连接函数都是对数函数, 两个模型不同的是方差, 泊松模型的方差等于期望 μ , 而 NB-2 模型的方差是 $\mu + \alpha\mu^2$, NB-2 模型的方差多出来一项 $\alpha\mu^2$, 就是多出来的一项使得 NB-2 模型的理论方差不再是和期望相同, 并且可以通过参数 α 调节方差和期望的大小关系, 可以拟合泊松过度分散的数据, 因此 NB-2 模型是最常用的用于替代泊松模型处理泊松过度分散数据的方案。

18.4.2 参数 α 的估计

辅助参数 α 和分散参数 ϕ 是不同的。分散参数 ϕ 不影响协变量参数 β 的估计过程, 因此可以在 IRLS 算法之后利用皮尔逊分散统计量估计。然而, 辅助参数 α 是会影响协变量参数 β 的估计过程的, 所以 α 的估计是不同于 ϕ 的。

负二项式模型中 α 参数的确定一般有两种方法, 一种是在 IRLS 以外, 用某种方法确定 α 的值, 然后把它当成一个常量代入 IRLS 过程。此时负二项式模型就是一个标准的单参数 GLM 模型, 可以使用 IRLS 进行协变量参数估计。

如果需要模型自行估计 α , 就需要修改 IRLS 算法的过程, 在 IRLS 迭代的过程中插入 α 的估计的过程。在 IRLS 迭代的每一步中插入一个 α 的估计过程, α 和 β 交替估计。

在一个迭代步骤中, 先假设 α 是已知的, 执行 β 的估计过程, 然后再利用 β 的估计值, 估计 α 的值。

在 β 已知的情况下, 可以利用皮尔逊分散统计量来估计 α 。理论上, 负二项式模型的分散统计量值为 1, 可以利用这个

$$\hat{\phi} = \frac{\chi^2}{N - p} = \sum \frac{(y_i - \mu_i)^2}{(N - p)(\mu_i - \alpha\mu_i^2)} = 1 \quad (18.4.6)$$

18.4.3 几何模型

当 $\alpha = 1$ 时, 负二项式分布就变成了几何分布, 几何分布是负二项式分布的一个特例。分别把 NB-C 模型和 NB-2 模型的 α 设置为 1 就得到了标准连接和对数连接的几何分布。

标准连接的 NB-C 模型的概率质量函数公式 (18.2.7) 转换成几何分布的概率质量函数为

$$f(y; \mu) = \exp \left\{ y \ln \left(\frac{\mu}{1 + \mu} \right) - \ln(1 + \mu) \right\} \quad (18.4.7)$$

对数连接的 NB-2 模型的概率质量函数公式 (18.4.7) 转换成几何分布的概率质量函数为

$$f(y; \mu) = \exp \{ y \ln(\mu) - \ln(1 + \mu) \} \quad (18.4.8)$$

相比于负二项式分布, 几何分布的概率质量函数没有了 Gamma 函数的部分, 变得更加简洁一些。几何分布和指数分布的图形几乎是已知的, 不同的是, 指数分布是连续值分布, 而几何分布是离散分布, 一般称几何分布和指数分布是离散相关的。

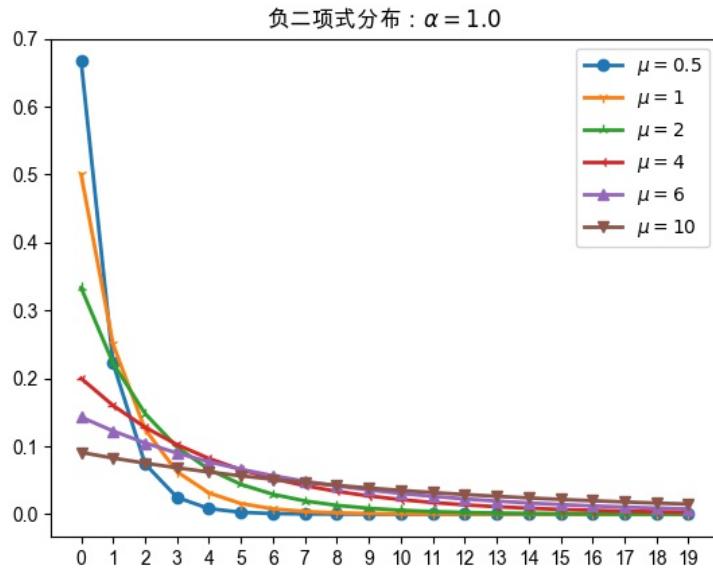


图 18.4.1: 几何分布和指数分布近乎是一致的,

几何分布的期望是 μ ，方差是 $\mu + \mu^2$ ，各个关键组件为

$$\begin{aligned}
 \text{自然参数} \quad \theta &= \ln \left(\frac{\mu}{1 + \mu} \right) \\
 \text{累积分布函数} \quad b(\theta) &= \ln(1 + \mu) \\
 \text{期望} \quad b'(\theta) &= \mu \\
 \text{方差函数} \quad \nu(\mu) &= b''(\theta) = \mu + \mu^2 \\
 \text{分散函数} \quad a(\phi) &= 1 \\
 \text{方差} \quad V(Y) &= a(\phi)\nu(\mu) = \mu + \mu^2 = \mu(1 + \mu) \\
 \text{标准连接函数} \quad g(\mu) &= \ln \left(\frac{\mu}{1 + \mu} \right) \\
 \text{标准连接函数一阶导} \quad g'(\mu) &= \frac{1}{\mu + \mu^2}
 \end{aligned} \tag{18.4.9}$$

标准连接的对数似然函数为

$$\begin{aligned}
 \ell(\mu; y) &= \sum_{i=1}^N \left\{ y_i \ln \left(\frac{\mu_i}{1 + \mu_i} \right) - \ln(1 + \mu_i) \right\} \\
 &= \sum_{i=1}^N \{ y_i \ln(\mu_i) - (1 + y_i) \ln(1 + \mu_i) \}
 \end{aligned} \tag{18.4.10}$$

标准连接的偏差统计量为

$$\mathcal{D} = 2 \sum_{i=1}^N \left\{ y_i \ln \left(\frac{y_i}{\mu_i} \right) - (1 + y_i) \ln \left(\frac{1 + y_i}{1 + \mu_i} \right) \right\} \tag{18.4.11}$$

几何模型是一个单参数模型，因此可以直接应用 IRLS 算法进行参数估计。

18.4.4 广义负二项式模型

负二项式模型的方差函数为 $\mu + \alpha\mu^2$ ，其中有一个 μ 的二次项，如果把二次项改成一次就变成了一个线性参数化方程，我们把方差函数为 $\mu + \alpha\mu$ 的模型称为线性负二项式模型，简称为 NB-1 模型。泊松模型的方差等于期望 μ ，而这些扩展模型都是在泊松方差的基础上乘上一个因子，不同的乘数因子形成了不同的泊松扩展模型。

- 泊松: $V = \mu$
- NB1: $V = \mu(1 + \alpha)$
- NB2: $V = \mu(1 + \alpha\mu)$
- 几何: $V = \mu(1 + \mu)$

NB-1 模型，或者说线性负二项式模型，也可以看做是泊松-伽马混合模型，只不过在混合模型的定义和推导上和 NB-2 模型有些差别。

NB-1 模型方差函数是 $\mu + \alpha\mu$ ，NB-2 模型方差函数是 $\mu + \alpha\mu^2$ ，二者的差别在于 μ 的幂次不一样，按照这个规律是不是可以有更高幂次的模型？更进一步，是不是可以把幂次也参数化？比如用参数 Q 参数化方差函数的最高幂次，则方差函数变为

$$V = \mu + \alpha\mu^Q \quad (18.4.12)$$

其中 Q 是一个待估计的未知参数。把方差函数的幂次参数化，相当于对负二项式模型的方差函数进行了一般化扩展，因此我们称之为广义负二项式模型 (generalized negative binomial)，一般简称为 NB-P 模型。

NB-P 模型是一个三参数模型，三个参数分别是期望参数 μ ，辅助参数 α ，以及参数 Q ，显然已经不再属于 GLM 框架下的模型，无法用 GLM 原版的 IRLS 算法进行参数估计。

零计数问题

在计数数据中，有一个非常特殊的数字：0。0表示事件发生的次数为0，常用的计数模型，泊松分布、负二项式分布都是支持0次数的，但是这些分布中0的期望（或者说发生概率）是有一定的限度的，然而，实际应用中观测数据中的0经常与模型分布的假设相违背，比如数据中没有0或者拥有过多的0。无论泊松分布还是负二项式分布，0的概率是不能为0的，也就是正常的分布采样数据中必须是存在0数据的。同理，分布中0的概率也不是任意大的，观测数据中过多的0也是和分布不匹配的。

数据中没有0或者0太多，都与先前计数模型的假设相违背。通常，把没有0的数据称为零截断（zero-truncate）数据，把0过多的数据称为零膨胀（zero-inflate）数据。虽然仍然可以使用泊松模型和负二项式模型处理这两类数据，但在效果上总是差强人意的。本章，我们介绍针对这两种情况的改进计数模型，处理零截断数据的模型称为零截断模型，处理零膨胀数据的模型称为零膨胀模型。

19.1 零截断模型

在计数类场景中，有时会存在没有0的情况，比如，住院病人住院的天数，不会出现住院天数是0的情况，这是一个典型的非零数据。当然这个例子中研究的对象是住院病人，那些不需要住院的病人不在研究范围内。泊松分布和负二项式分布都是包含0的，虽然仍然可以使用这两个模型处理非零数据，但显然它们并不是那么契合此类数据。要想使得泊松分布或负二项式能更好的适配非零数据，最直接的方法就是调整下概率分布函数，把0从概率分布函数中去掉。

假设随机变量 Y 的概率分布函数是 $f(y)$ ，根据概率和为1约束，可知 $\sum_y f(y) = 1$ ，如果要从 $f(y)$ 中去掉一个值，就不满足上述约束了，此时就需要对 $f(y)$ 重新归一化以满足概率约束。假设 Y 是一个计数型离散随机变量， $Y = 0$ 的概率是 $f(0)$ ，那么 $Y > 0$ 的概率分布函数为

$$f(y|y > 0) = \frac{f(y)}{1 - f(0)} \quad (19.1.1)$$

计数数据模型中最常见的泊松模型和负二项式模型，都可以使用这种方法转换成零截断数据的模型。

19.1.1 零截断泊松模型

先以泊松分布为例, 泊松分布的概率分布函数为

$$f(y; \mu) = \exp\{y \ln(\mu) - \mu - \ln \Gamma(y + 1)\} \quad (19.1.2)$$

$Y = 0$ 的概率为

$$f(y = 0; \mu) = \exp\{-\mu\} \quad (19.1.3)$$

不包含 0 的泊松分布的概率分布函数为

$$f(y; \mu | y > 0) = \frac{\exp\{y \ln(\mu) - \mu - \ln \Gamma(y + 1)\}}{1 - \exp\{-\mu\}} \quad (19.1.4)$$

对数似然函数为

$$\ell(\mu; y | y > 0) = \sum_{i=1}^N \left\{ \underbrace{y_i \ln(\mu_i) - \mu_i - \ln \Gamma(y_i + 1)}_{\text{原来泊松分布部分}} - \underbrace{\ln[1 - \exp\{-\mu_i\}]}_{\text{新的归一化项}} \right\} \quad (19.1.5)$$

仔细观察下, 零截断泊松模型的对数似然函数相比原始泊松模型对数似然函数, 只是多了一个归一化项。由于对数似然函数结构发生了变化, IRLS 算法不适用于零截断模型, 严格来说, 零截断模型已经不属于 GLM 框架。对于零截断模型的参数估计, 可以单独运用完整最大似然估计求解, 也就是梯度法或者牛顿法。

19.1.2 零截断负二项式模型

负二项式模型的概率分布函数为

$$f(y; \alpha, \mu) = \exp \left\{ y \ln \left(\frac{\alpha \mu}{1 + \alpha \mu} \right) - \frac{1}{\alpha} \ln(1 + \alpha \mu) + \ln \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1) \Gamma(1/\alpha)} \right\} \quad (19.1.6)$$

$Y = 0$ 的概率为

$$f(y = 0; \alpha, \mu) = \exp \left\{ -\frac{1}{\alpha} \ln(1 + \alpha \mu) \right\} = (1 + \alpha \mu)^{-1/\alpha} \quad (19.1.7)$$

用符号 $f(y; \alpha, \mu)_{NB}$ 代指公式 (19.1.6), 符号 $f(0; \alpha, \mu)_{NB}$ 代指公式 (19.1.7), 零截断负二项式模型的概率分布函数为

$$f(y; \alpha, \mu | y > 0) = \frac{f(y; \alpha, \mu)_{NB}}{1 - f(0; \alpha, \mu)_{NB}} \quad (19.1.8)$$

用符号 ℓ_{NB} 表示原始负二项式模型对数似然函数, 则零截断负二项式模型的对数似然函数为

$$\ell(\mu; y | y > 0) = \ell_{NB} - \sum_{i=1}^N \left\{ \ln[1 - (1 + \alpha \mu_i)^{-1/\alpha}] \right\} \quad (19.1.9)$$

19.2 零膨胀模型

零膨胀 (zero-inflate) 表示数据中的 0 太多了, 超过了泊松分布和负二项式分布中 0 的极限。对于零膨胀的计数数据, 常见的处理思想就是把数据进行分割, 模型分成两段式。所谓两段式是指, 假设观测数据由两个过程生成, 先进行一个二分类过程, 再进行一个计数过程, 实际应用中, 有两种常见的实现方法,

0 和非 0

把观测数据分成 0 和非 0 两部分, 先用一个二分类模型判断是 0 还是非 0, 然后用一个零截断模型处理非 0 的数据。采用这种方式的模型称为栅栏 (Hurdle) 模型。

“多余”的 0 和“正常”的 0

把数据中的 0 分成两部分, 一部分认为是计数分布中产生的 0, 这些 0 和非零数据组成计数模型的部分。剩余的 0 认为是“多余”的, 用二分类模型处理。这种方法的模型习惯上称为零膨胀模型 (zero-inflated model, ZIM)。

栅栏模型和零膨胀模型都是两段式模型, 第一阶段是二分类模型, 第二阶段是计数模型。二者在数据的划分上有些许差别, 这个差别使得两种模型有不一样的特性, 零膨胀模型只能处理有过多 0 的数据, 而栅栏模型是可以处理 0 过少的数据的。为了和零膨胀模型区分, Hurdle 模型一般被称为 zero-altered 模型。

19.2.1 Hurdle 模型

Hurdle 在英文语境里指栅栏、障碍, 顾名思义, Hurdle 寓意在数据生成过程中有一个 0 的栅栏阻碍, 没有跨过去就为 0, 跨过去就是非零的正整数, Hurdle 模型相当于是一个二分类模型和零截断模型组成的一个混合模型, 其概率分布函数可以用一个分段函数表示。

$$f(y; \pi, \lambda) = \begin{cases} \text{Binary}(0; \pi) & \text{if } y = 0 \\ [1 - \text{Binary}(0; \pi)] \times \text{Zero-Truncate}(y; \lambda) & \text{if } y > 0 \end{cases} \quad (19.2.1)$$

其中二分类模型和零截断模型都有可以多种不同的选择。

- 二分类模型: logit 回归, probit 回归, complementary log-log 回归等。
- 零截断模型: 零截断泊松模型, 零截断负二项式模型。

因为 Hurdle 模型又被称为 zero-altered 模型, 因此采用零截断泊松模型的 Hurdle 模型简称为 ZAP 模型, 采用零截断负二项式模型的 Hurdle 模型简称为 ZANB 模型。

Hurdle 模型中二分类部分和零截断计数部分一般拥有不同的线性预测器, 假设二分类模型的线性预测器为 $\eta^\gamma = z^T \gamma$, 响应函数为 $F(\eta^\gamma)$ 。零截断计数模型的线性预测器为 $\eta^\beta = x^T \beta$, 标准连接的响应函数为 $R(\eta^\beta)$ 。下面我们分别看下 ZAP 模型和 ZANB 模型的定义。

ZAP

当 Hurdle 模型中的零截断模型是零截断泊松模型时, 简称为 ZAP 模型, 零截断泊松模型的概率分布函数是公式 (19.1.4), 代入到公式 (19.2.1) 的模板可以给出 ZAP 的概率分布函数

$$\begin{aligned} P(y_i = 0) &= F(\eta_i^\gamma) \\ P(y_i > 0) &= [1 - F(\eta_i^\gamma)] \frac{\exp\{y_i \eta_i^\beta - R(\eta_i^\beta) - \ln \Gamma(y_i + 1)\}}{1 - \exp\{-R(\eta_i^\beta)\}} \end{aligned} \quad (19.2.2)$$

Hurdle 模型的对数似然函数可以按照样本是否为 0 分成两部分, 假设观测样本中为 0 的子集是 \mathcal{D}_0 , 则 Hurdle 模型的对数似然函数为

$$\ell = \sum_{i \in \mathcal{D}_0} \ln P(y_i = 0) + \sum_{i \notin \mathcal{D}_0} \ln P(y_i > 0) \quad (19.2.3)$$

按照这个模板, ZAP 模型的对数似然函数为

$$\begin{aligned} \ell &= \sum_{i \in \mathcal{D}_0} \ln F(\eta_i^\gamma) \\ &+ \sum_{i \notin \mathcal{D}_0} \left\{ \ln [1 - F(\eta_i^\gamma)] + \left[y_i \eta_i^\beta - R(\eta_i^\beta) - \ln \Gamma(y_i + 1) \right] - \ln \left\{ 1 - \exp[-R(\eta_i^\beta)] \right\} \right\} \end{aligned} \quad (19.2.4)$$

ZANB

当 Hurdle 模型中的零截断模型是零截断负二项式模型时, 简称为 ZANB 模型, 零截断泊松模型的概率分布函数是公式 (19.1.8), 代入到公式 (19.2.1) 的模板可以给出 ZANB 的概率分布函数

$$\begin{aligned} P(y_i = 0) &= F(\eta_i^\gamma) \\ P(y_i > 0) &= [1 - F(\eta_i^\gamma)] \frac{f(y)_{NB}}{1 - f(0)_{NB}} \end{aligned} \quad (19.2.5)$$

19.2.2 Zero-inflate 模型

零膨胀 (Zero-inflate) 计数模型是由 Lambert (1992) 首次提出的, 它提供了另一种解决过多零计数的方法。与 Hurdle 模型一样, 它也是两部分模型, 由二分类和计数模型组成。与 Hurdle 模型不同的地方在于, 零膨胀模型提供了同时使用二分类和计数过程对零计数进行建模的功能。Hurdle 模型将零的建模与计数的建模分开, 这意味着只有一个过程会生成零, 从对数似然函数中也能看到这一点。与之不同的时, 零膨胀模型将零计数合并到二分类和计数过程中。

和 Hurdle 模型是类似的是, 零膨胀模型也是一个两段式的模型, 二分类模型和计数模型组合在一起, 不同的地方在于, 划分的方法不太一样。零膨胀模型把数据中的 0 看做两部分, 一部分是二分类模型产生, 另一部分由计数模型产生。这里的计数模型既负责一部分 0 的数据, 又负责非 0 的数据, 所以这里的计数模型是一个完整的计数模型, 而不是零截断计数模型, 这一点和 Hurdle 模型不同。

$$f(y) = \begin{cases} \text{Binary}(0; \pi) + [1 - \text{Binary}(0; \pi)]\text{Count}(0; \lambda) & \text{if } y = 0 \\ [1 - \text{Binary}(0; \pi)]\text{Count}(y; \lambda) & \text{if } y > 0 \end{cases} \quad (19.2.6)$$

二分类模型部分, 通常使用的是伯努利模型, 连接函数常用的是 logit, probit 和 complementary log-log。计数模型部分常用的是泊松模型和负二项式模型, 采用泊松模型的零膨胀模型通常简称为 ZIP 模型, 采用负二项式模型的零膨胀模型通常简称为 ZINB 模型。

二分类部分和计数部分一般拥有不同的线性预测器, 假设二分类模型的线性预测器为 $\eta^\gamma = z^T \gamma$, 响应函数为 $F(\eta^\gamma)$ 。计数模型的线性预测器为 $\eta^\beta = x^T \beta$, 响应函数为 $R(\eta^\beta)$ 。下面我们分别看下 ZIP 模型和 ZINB 模型的定义。

ZIP 模型

计数模型是泊松模型的零膨胀模型称为 ZIP 模型, 把泊松模型的概率分布函数公式 (19.1.2) 代入到公式 (19.2.6) 可得到 ZIP 模型的概率分布函数。

$$\begin{aligned} P(y_i = 0) &= F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)] \exp[-R(\eta_i^\beta)] \\ P(y_i > 0) &= [1 - F(\eta_i^\gamma)] \exp \left\{ y_i \eta_i^\beta - R(\eta_i^\beta) - \ln \Gamma(y_i + 1) \right\} \end{aligned} \quad (19.2.7)$$

用 \mathcal{D}_0 表示观测样本集中 0 的子集, ZIP 模型的对数似然函数可以写成两部分的和。

$$\ell = \sum_{i \in \mathcal{D}_0} \ln \left\{ F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)] \exp[-R(\eta_i^\beta)] \right\} + \sum_{i \notin \mathcal{D}_0} \left\{ \ln[1 - F(\eta_i^\gamma)] + y_i \eta_i^\beta - R(\eta_i^\beta) - \ln(y_i!) \right\} \quad (19.2.8)$$

对数似然函数的一阶偏导数为

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= - \sum_{i \in \mathcal{D}_0} x_{ji} \frac{[1 - F(\eta_i^\gamma)] R(\eta_i^\beta) \exp[-R(\eta_i^\beta)]}{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)] \exp[-R(\eta_i^\beta)]} + \sum_{i \notin \mathcal{D}_0} x_{ji} (y_i - R(\eta_i^\beta)) \\ \frac{\partial \ell}{\partial \gamma_j} &= \sum_{i \in \mathcal{D}_0} z_{ji} \frac{F'(\eta_i^\gamma) \left\{ 1 - \exp[-R(\eta_i^\beta)] \right\}}{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)] \exp[-R(\eta_i^\beta)]} - \sum_{i \notin \mathcal{D}_0} z_{ji} \frac{F'(\eta_i^\gamma)}{1 - F(\eta_i^\gamma)} \end{aligned} \quad (19.2.9)$$

其中 $F'(\eta_i^\gamma)$ 表示 $F(\eta_i^\gamma)$ 对 γ 的导数。

ZINB 模型

计数模型是负二项式模型的零膨胀模型称为 ZINB 模型, 把负二项式模型的概率分布函数公式 (19.1.6) 代入到公式 (19.2.6) 可得到 ZINB 模型的概率分布函数。

$$\begin{aligned} m &= 1/\alpha \\ p_i &= 1/(1 + \alpha\mu_i) \\ \mu_i &= R(\eta_i^\beta) = \exp(x_i\beta) \\ P(y_i = 0) &= F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)]p_i^m \\ P(y_i > 0) &= [1 - F(\eta_i^\gamma)] \frac{\Gamma(m + y_i)}{\Gamma(y_i + 1)\Gamma(m)} p_i^m (1 - p_i)^y \end{aligned} \quad (19.2.10)$$

ZINB 模型的对数似然函数为

$$\begin{aligned} \ell &= \sum_{i \in \mathcal{D}_0} \ln \{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)]p_i^m\} \\ &+ \sum_{i \notin \mathcal{D}_0} \{\ln[1 - F(\eta_i^\gamma)] + \ln \Gamma(m + y_i) - \ln \Gamma(y_i + 1) - \ln \Gamma(m) + m \ln p_i + y_i \ln(1 - p_i)\} \end{aligned} \quad (19.2.11)$$

其一阶偏导导数为

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i \in \mathcal{D}_0} x_{ji} \frac{-[1 - F(\eta_i^\gamma)]\mu_i p_i^{m+1}}{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)]p_i^m} + \sum_{i \notin \mathcal{D}_0} x_{ji} p_i (y_i - \mu_i) \quad (19.2.12)$$

$$\frac{\partial \ell}{\partial \gamma_j} = \sum_{i \in \mathcal{D}_0} z_{ji} \frac{F'(\eta_i^\gamma)(1 - p_i^m)}{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)]p_i^m} + \sum_{i \notin \mathcal{D}_0} z_{ji} \frac{-F'(\eta_i^\gamma)}{1 - F(\eta_i^\gamma)} \quad (19.2.13)$$

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= - \sum_{i \in \mathcal{D}_0} \frac{m^2 p_i^m \ln p_i - m \mu_i p_i^{m-1}}{F(\eta_i^\gamma) + [1 - F(\eta_i^\gamma)]p_i^m} \\ &- \sum_{i \notin \mathcal{D}_0} \alpha^{-2} \left\{ \frac{\alpha(\mu_i - y_i)}{1 + \alpha\mu_i} - \ln(1 + \alpha\mu_i) + \psi(y_i + 1/\alpha) - \psi(1/\alpha) \right\} \end{aligned} \quad (19.2.14)$$

其中 ψ 表示双伽马函数 (digamma function)

多项式模型

<https://www.freesion.com/article/4717201261/>

<https://www.freesion.com/article/4717201261/>

<https://www.jianshu.com/p/99c4238f067e>

20.1 类别分布

类别分布可以看做是伯努利分布的扩展，在讲类别分布之前，先回顾一下伯努利分布。伯努利变量只有两个离散状态，一般用 0 和 1 表示，伯努利分布的概率质量函数为

$$f(y; \pi) = \pi^y (1 - \pi)^{1-y} \quad (20.1.1)$$

为方便扩展到类别分布，我们先定义一个指示函数， $\mathbb{I}(\text{true}) = 1$ ，指示函数只有两个值 0 或者 1，当满足条件时函数值为 1，否则为 0。

$$\mathbb{I}(x = A) = \begin{cases} 1 & \text{if } x = A \\ 0 & \text{otherwise} \end{cases} \quad (20.1.2)$$

用指示函数可以把公式 (20.1.1) 改写成

$$f(y; \pi) = \prod_k \pi_k^{\mathbb{I}(y=k)}, \quad k \in \{0, 1\} \quad (20.1.3)$$

其中 k 表示伯努利分布的两个状态， $k = 0$ 表示状态 0， $k = 1$ 表示状态 1。 π_0 表示变量 $Y = 0$ 的概率， π_1 表示变量 $Y = 1$ 的概率。

$$\begin{aligned} P(Y = 0) &= f(y = 0; \pi) = \pi_0^1 \pi_1^0 = \pi_0 \\ P(Y = 1) &= f(y = 1; \pi) = \pi_0^0 \pi_1^1 = \pi_1 \end{aligned} \quad (20.1.4)$$

参数 $\pi = [\pi_0, \pi_1]$ 表示的是概率值，满足约束 $\sum_k \pi_k = \pi_0 + \pi_1 = 1$ ， π_0 可以由 π_1 计算得到 $\pi_0 = 1 - \pi_1$ 。因此对于伯努利分布来说，实际上不需要两个未知参数，只需要一个参数就可以，伯努利分布的概率质量函数一般会写成公式 (20.1.1) 的单参数形式。

伯努利变量只有两个状态, 如果增加到多个状态就是类别分布。类别分布是伯努利分布的扩展, 由两个离散状态扩展到更多的离散状态。假设随机变量 Y 是 K 个离散状态的类别变量, 公式 (20.1.1) 可以扩展成类别分布的概率质量函数。

$$\begin{aligned} f(y; \pi_1, \pi_2, \dots, \pi_K) &= \pi_1^{\mathbb{I}(y=1)} \pi_2^{\mathbb{I}(y=2)} \cdots \pi_K^{\mathbb{I}(y=K)} \\ &= \prod_{k=1}^K \pi_k^{\mathbb{I}(y=k)} \quad (1 \leq k \leq K) \end{aligned} \quad (20.1.5)$$

π_k 表示类别 k 的概率值, $P(Y = y_k) = \pi_k$, 所有的 π_k 满足概率约束

$$\sum_{k=1}^K \pi_k = 1 \quad (20.1.6)$$

既然所有的 π_k 相加为 1, 就没有必要使用 K 个参数来表示 K 个类别的概率, 最后一个类别 K 的概率可以写成

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k \quad (20.1.7)$$

公式 (20.1.5) 形式有 K 个参数, 可以转换成只有 $K - 1$ 个参数的形式。

$$\begin{aligned} f(y; \pi_1, \pi_2, \dots, \pi_{K-1}) &= \pi_1^{\mathbb{I}(y=1)} \pi_2^{\mathbb{I}(y=2)} \cdots \pi_{(K-1)}^{\mathbb{I}(y=K-1)} \left(1 - \sum_{k=1}^{K-1} \pi_k\right)^{\mathbb{I}(y=K)} \\ &= \left(1 - \sum_{k=1}^{K-1} \pi_k\right)^{\mathbb{I}(y=K)} \prod_{k=1}^{K-1} \pi_k^{\mathbb{I}(y=k)} \end{aligned} \quad (20.1.8)$$

由于类别变量 Y 有多个离散状态值, 其期望是一个向量, 方差是一个协方差矩阵。类别分布的期望向量为

$$\mathbb{E}[Y] = [\pi_1, \pi_2, \dots, \pi_K]^T \quad (20.1.9)$$

方差 $V(Y_k)$ 和协方差 $Cov(Y_p, Y_q)$ 为

$$\begin{aligned} V(Y_k) &= \pi_k(1 - \pi_k) \\ Cov(Y_p, Y_q) &= -\pi_p \pi_q, \quad (p \neq q) \end{aligned} \quad (20.1.10)$$

重要: 注意, 类别变量 $Y = [y_1, y_2, \dots, y_K]$ 的各个离散状态值是没有任何大小和顺序的关系, 各个离散状态值是不可比较的。

20.2 softmax 回归模型

20.2.1 模型定义

现在我们推导下 GLM 中类别分布的回归模型, 在推导过程中, 需要用如下一个等式。

$$\mathbb{I}(y = K) = 1 - \sum_{k=1}^{K-1} \mathbb{I}(y = k) \quad (20.2.1)$$

类别分布的概率质量函数公式 (20.1.5) 转换成指数族的形式过程为

$$\begin{aligned}
 f(y; \pi_1, \pi_2, \dots, \pi_{K-1}) &= \pi_K^{\mathbb{I}(y=K)} \prod_{k=1}^{K-1} \pi_k^{\mathbb{I}(y=k)} \\
 &= \exp \left[\ln \left(\pi_K^{\mathbb{I}(y=K)} \right) + \ln \left(\prod_{k=1}^{K-1} \pi_k^{\mathbb{I}(y=k)} \right) \right] \\
 &= \exp \left[\mathbb{I}(y=K) \ln \pi_K + \sum_{k=1}^{K-1} \ln \pi_k^{\mathbb{I}(y=k)} \right] \\
 &= \exp \left\{ \left[1 - \sum_{k=1}^{K-1} \mathbb{I}(y=k) \right] \ln \pi_K + \sum_{k=1}^{K-1} \mathbb{I}(y=k) \ln \pi_k \right\} \\
 &= \exp \left\{ \ln \pi_K - \sum_{k=1}^{K-1} \mathbb{I}(y=k) \ln \pi_K + \sum_{k=1}^{K-1} \mathbb{I}(y=k) \ln \pi_k \right\} \\
 &= \exp \left\{ \sum_{k=1}^{K-1} \mathbb{I}(y=k) \ln \frac{\pi_k}{\pi_K} + \ln \pi_K \right\}
 \end{aligned} \tag{20.2.2}$$

上式中存在指示函数不方便处理, 为方便处理需要转换一下。响应变量 Y 是一个有 K 个可能取值的离散变量, 我们把 Y 转换成一个长度为 K 的向量。令 $T(Y)$ 表示一个长度为 K 的向量, 向量 $T(Y)$ 只能有一个元素为 1, 其余元素为 0, $T(Y)_k = 1$ 表示向量 $T(Y)$ 第 k 个元素为 1, 等价于响应变量 $Y = k$ 。

$$T(y)_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, T(y)_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, T(y)_{K-1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, T(y)_K = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{20.2.3}$$

指示函数和向量 $T(Y)$ 的关系可以写为

$$T(Y)_k = \mathbb{I}(y=k) \tag{20.2.4}$$

此外, 我们再定义一个长度为 K 的参数向量 θ ,

$$\theta = \begin{bmatrix} \ln \pi_1 / \pi_K \\ \ln(\pi_2 / \pi_K) \\ \vdots \\ \ln(\pi_k / \pi_K) \\ \vdots \\ \ln(\pi_{K-1} / \pi_K) \\ \ln(\pi_K / \pi_K) = 0 \end{bmatrix} \tag{20.2.5}$$

依据向量积的定理则有

$$\begin{aligned}
 \theta^T T(y) &= \sum_{k=1}^K T(y)_k \ln \frac{\pi_k}{\pi_K} \\
 &= T(y)_K \ln \frac{\pi_K}{\pi_K} + \sum_{k=1}^{K-1} T(y)_k \ln \frac{\pi_k}{\pi_K} \\
 &= 0 + \sum_{k=1}^{K-1} T(y)_k \ln \frac{\pi_k}{\pi_K} \\
 &= \sum_{k=1}^{K-1} T(y)_k \ln \frac{\pi_k}{\pi_K} \\
 &= \sum_{k=1}^{K-1} \mathbb{I}(y = k) \ln \frac{\pi_k}{\pi_K}
 \end{aligned} \tag{20.2.6}$$

此外根据概率约束有

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k \tag{20.2.7}$$

通过把累加转化成向量积的方式, 把公式 (20.2.2) 写成指数族的形式

$$\begin{aligned}
 f(y; \theta) &= \exp \left\{ \sum_{k=1}^{K-1} \mathbb{I}(y = k) \ln \frac{\pi_k}{\pi_K} + \ln \pi_K \right\} \\
 &= \exp \{ \theta^T T(y) + A(\theta) \}
 \end{aligned} \tag{20.2.8}$$

上式就是类别分布概率质量函数的指数族形式, 其中 θ 就是自然参数, 并且 θ 和 $T(y)$ 分别是一个向量。可以看到类别分布的指数族形式是不存在分散函数 $a(\phi)$ 的, 或者说 $a(\phi) = 1$ 。

标准连接函数

自然参数 θ 和期望参数 $\mu = \pi$ 的关系为

$$\theta_k = \ln \frac{\pi_k}{\pi_K} = \ln \frac{\pi_k}{(1 - \sum_{k=1}^{K-1} \pi_k)} \tag{20.2.9}$$

根据标准连接函数的定义, 当线性预测 η 等于自然参数 θ 时得到的就是标准连接函数因此其标准连接函数为

$$\begin{aligned}
 \eta_k &= \theta_k = \ln \frac{\pi_k}{\pi_K} \\
 &= \ln \frac{\pi_k}{(1 - \sum_{k=1}^{K-1} \pi_k)} \\
 &= \ln \frac{\mu_k}{(1 - \sum_{k=1}^{K-1} \mu_k)}
 \end{aligned} \tag{20.2.10}$$

响应函数

响应函数是连接函数的反函数, 现在我们推导下公式 (20.2.10) 的反函数。首先等号两边取值指数操作, 可得

$$\pi_k = \pi_K e^{\eta_k} \tag{20.2.11}$$

上式中仍然存在 π_K ，需要推导下 π_K 如何用 η 表示。

$$\begin{aligned}
 1 &= \sum_{k=1}^K \pi_k \\
 &= \pi_K + \sum_{k=1}^{K-1} \pi_k \\
 &= \pi_K + \sum_{k=1}^{K-1} \pi_K e^{\eta_k} \\
 &= \pi_K \left(1 + \sum_{k=1}^{K-1} e^{\eta_k} \right)
 \end{aligned} \tag{20.2.12}$$

因此，有

$$\pi_K = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\eta_k}} \tag{20.2.13}$$

代入到公式 (20.2.10) 可得

$$\mu_k = \pi_k = \pi_K e^{\eta_k} = \frac{e^{\eta_k}}{1 + \sum_{k=1}^{K-1} e^{\eta_k}}, \quad \text{for } k = 1, 2, \dots, K-1 \tag{20.2.14}$$

公式 (20.2.14) 就是类别分布的响应函数，此函数通常叫作 softmax 函数，它是 logistic 函数的扩展，当 $K = 2$ 时，softmax 就是退化成 logistic 函数。

对于有 K 个类别的 softmax 回归模型来说，其线性预测器只有 $K - 1$ 个，这意味着有 $K - 1$ 个协变量参数向量， $\beta = [\beta_1, \beta_2, \dots, \beta_{K-1}]$ 。

$$\eta_k = \ln \frac{\mu_k}{(1 - \sum_{k=1}^{K-1} \mu_k)} = \beta_k x^T, \quad \text{for } k = 1, 2, \dots, K-1 \tag{20.2.15}$$

这 $K - 1$ 个类别的响应函数就是 softmax 函数。

$$\mu_k = \frac{e^{\eta_k}}{1 + \sum_{k=1}^{K-1} e^{\eta_k}}, \quad \text{for } k = 1, 2, \dots, K-1 \tag{20.2.16}$$

第 K 个类别不需要独立的线性预测器，依据概率的约束，它的期望可以通过其余 $K - 1$ 个计算得到。

$$\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\eta_k}} \tag{20.2.17}$$

最后我们整理下 softmax 回归的各个关键部分。

$$\begin{aligned}
 \text{自然参数} \quad \theta_k &= \ln \frac{\pi_k}{\pi_K} = \ln \frac{\pi_k}{(1 - \sum_{k=1}^{K-1} \pi_k)} \\
 \text{累积分布函数} \quad b(\theta) &= \ln \pi_K = (1 - \sum_{k=1}^{K-1} \pi_k) \\
 \text{期望} \quad \mu_k &= b'(\theta_k) = \pi_k \\
 \text{方差} \quad V(Y_k) &= \pi_k(1 - \pi_k) \\
 \text{协方差} \quad Cov(Y_p, Y_q) &= -\pi_p \pi_q, \quad (p \neq q) \\
 \text{分散函数} \quad a(\phi) &= 1 \\
 \text{标准连接函数} \quad g(\mu_k) &= \ln \frac{\pi_k}{\pi_K} = \ln \frac{\pi_k}{(1 - \sum_{k=1}^{K-1} \pi_k)} \\
 \text{标准连接函数一阶导} \quad g'(\mu_k) &= \frac{\pi_k + \pi_K}{\pi_k \pi_K} = \frac{\pi_k + 1 - \sum_{k=1}^{K-1} \pi_k}{\pi_k(1 - \sum_{k=1}^{K-1} \pi_k)} \\
 \text{响应函数} \quad r(\eta_k) &= \frac{e^{\eta_k}}{1 + \sum_{k=1}^{K-1} e^{\eta_k}}
 \end{aligned} \tag{20.2.18}$$

类别分布是伯努利分布在多类别上的扩展，因此二者在链接函数、响应函数等方面都有一定的关联性。二值离散变量的概率分布是伯努利分布，其标准连接对应的响应函数是 logistic 函数，其回归模型是称为 logistic 回归，可用于处理二分类响应数据。多值离散变量的概率分布是类别分布，其标准连接对应的响应函数是 softmax 函数，其回归模型称为 softmax 回归，可用于处理多分类的响应数据。由于 softmax 是 logistic 的扩展，有时也会把 softmax 回归称为多元逻辑回归。

20.2.2 参数估计

softmax 回归模型也是 GLM 中的一员，并且它是 logistic 回归模型的扩展，它在 IRLS 估计算法的各个组件和 logistic 模型是类似的。

softmax 回归模型的对数似然函数为

$$\begin{aligned}
 \ell(\hat{\pi}; y) &= \sum_{i=1}^N \left\{ \sum_{k=1}^{K-1} \left[\mathbb{I}(y_i = k) \ln \frac{\hat{\pi}_{ik}}{\pi_{iK}} \right] + \ln \hat{\pi}_{iK} \right\} \\
 &= \sum_{i=1}^N \left\{ \sum_{k=1}^{K-1} \left[T(y_i)_k \ln \frac{\pi_{ik}}{\hat{\pi}_{iK}} \right] + \ln \hat{\pi}_{iK} \right\} \\
 &= \sum_{i=1}^N \left\{ \sum_{k=1}^{K-1} [T(y_i)_k \ln \hat{\pi}_{ik}] - \sum_{k=1}^{K-1} [T(y_i)_k \ln \hat{\pi}_{iK}] + \ln \hat{\pi}_{iK} \right\} \\
 &= \sum_{i=1}^N \left\{ \sum_{k=1}^{K-1} [T(y_i)_k \ln \hat{\pi}_{ik}] - \sum_{k=1}^{K-1} [T(y_i)_k \ln \hat{\pi}_{iK}] + \ln \hat{\pi}_{iK} \right\}
 \end{aligned} \tag{20.2.19}$$

不同于二元逻辑回归模型，softmax 回归模型有 $K - 1$ 个协变量参数向量，参数向量 β_k 的权重矩阵 W_k 和工作响应矩阵 Z_k 的计算公式分别为：

$$\begin{aligned}
 W_{iik} &= \frac{1}{a(\phi)\nu(\hat{\mu}_{ik})(g'_{ik})^2} \\
 &= \frac{\hat{\pi}_{ik}\hat{\pi}_{iK}^2}{(\hat{\pi}_{ik} + \hat{\pi}_{iK})^2(1 - \hat{\pi}_{ik})}
 \end{aligned} \tag{20.2.20}$$

$$\begin{aligned}
Z_{ik} &= [T(y_i)_k - \hat{\mu}_{ik}]g'_{ik} + \eta_{ik} \\
&= \frac{[T(y_i)_k - \hat{\mu}_{ik}](\hat{\pi}_{ik} + \hat{\pi}_{iK})}{\hat{\pi}_{ik}\hat{\pi}_{iK}} + \eta_{ik}
\end{aligned} \tag{20.2.21}$$

20.3 多项式分布

伯努利分布表示单次伯努利实验成功或者失败的概率，多次伯努利实验成功次数的概率分布是二项式分布。按照这种方式，多次类别分布实验各个状态发生次数的概率分布就叫做多项式分布 (multinomial distribution)。假设实验总次数为 n ，每个类别发生的次数为 y_k ，多项式分布的概率质量函数为

$$f(y; \pi) = \frac{n!}{y_1!y_2!\dots y_K!} \prod_{k=1}^K \pi_k^{y_k} \tag{20.3.1}$$

多项式分布变量的期望为

$$\mathbb{E}[y_k] = n\pi_k \tag{20.3.2}$$

方差 $V(Y_k)$ 和协方差 $Cov(Y_p, Y_q)$ 为

$$\begin{aligned}
V(Y_k) &= n\pi_k(1 - \pi_k) \\
Cov(Y_p, Y_q) &= -n\pi_p\pi_q, \quad (p \neq q)
\end{aligned} \tag{20.3.3}$$

可以看到多项式分布和类别分布的区别就是多了一个实验的总次数 n ，当 $n = 1$ 时，多项式分布就等价于类别分布，类别分布是多项式分布的一个特例。当 $K = 2$ 时，多项式分布就等价于二项式分布，因此二项式分布也是多项式分布的一个特例。多项式分布于类别分布的关系，就好比二项式分布于伯努利分布的关系，读者可以回顾一下二项式模型的章节加深理解。此外多项式分布也是二项式分布的扩展，二项式分布变量只有两个状态，而多项式分布扩展的更多的状态。

需要注意的是，虽然通常会使用连续的整数表示多项式变量的各个离散状态，但这些状态间是没有大小顺序关系的。存在顺序关系的离散状态随机变量，我们下一章再探讨。

20.4 多项式回归模型

类别分布可以用来处理多分类的响应数据，而多项式分布分科

$$\begin{aligned}
f(y; \pi) &= \frac{n!}{y_1!y_2!\dots y_K!} \prod_{k=1}^K \pi_k^{y_k} \\
&= \exp \left\{ \ln \left[\frac{n!}{y_1!y_2!\dots y_K!} \right] + \ln \prod_{k=1}^K \pi_k^{y_k} \right\} \\
&= \exp \left\{ \sum_{k=1}^K y_k \ln \pi_k + \ln \left[\frac{n!}{y_1!y_2!\dots y_K!} \right] \right\}
\end{aligned} \tag{20.4.1}$$

有序离散模型

有序多分类模型，可以看做是二分类逻辑回归模型的扩展，因此，有序多分类模型也被称为有序逻辑回归模型（Ordered Logistic Regression）。

21.1 有序逻辑回归

`logistic` 函数是一个累计概率分布函数。

$$\begin{aligned} P(Y = 1) &= \frac{e^{x^T \beta}}{1 + e^{x^T \beta}} \\ P(Y = 0) &= 1 - \frac{e^{x^T \beta}}{1 + e^{x^T \beta}} = \frac{e^{-x^T \beta}}{1 + e^{-x^T \beta}} \end{aligned} \quad (21.1.1)$$

线性预测器 $\eta = x^T \beta$ 的取值范围是 $(-\infty, \infty)$ ，

假设有一个分割点 c ，当 η 的值落在区间 $(-\infty, c]$ 时，响应变量 $Y = 0$ ；反之，当 η 的值落在区间 (c, ∞) 时，响应变量 $Y = 1$ 。

$$y = \begin{cases} 0 & \text{if } -\infty < \eta \leq c \\ 1 & \text{if } c < \eta < \infty \end{cases} \quad (21.1.2)$$

由于变量 Y 是一个随机变量，不能直接用上面的分段函数输出 Y 的值，而是要给出 Y 的一个概率分布。假设 $Y = 0$ 的概率和 η 与分割点 c 的距离（有方向）有关， η 与分割点 c 的距离为 $c - \eta$ ， $c - \eta$ 越大， $P(Y = 0)$ 也就越大， $c - \eta$ 越小， $P(Y = 0)$ 也就越小。此时可以选择一个累计概率分布函数作为 $P(Y = 0)$ 与 $c - \eta$ 之间的关系，累计概率分布函数有多种选择，比如 `logistic`, `probit`, `clog-log` 等等，这里以 `logistic` 为例。 $P(Y = 0)$ 可以写成

$$P(Y = 0) = \frac{e^{c-\eta}}{1 + e^{c-\eta}} \quad (21.1.3)$$

- 当 η 趋近于 $-\infty$ 时， $c - \eta$ 趋近于 $+\infty$ ， $P(Y = 0)$ 趋近于 1。
- 当 η 趋近于 c 时， $c - \eta$ 趋近于 0， $P(Y = 0)$ 趋近于 0.5。

- 当 η 趋近于 ∞ 时, $c - \eta$ 趋近于 $-\infty$, $P(Y = 0)$ 趋近于 0。

$P(Y = 1)$ 为

$$\begin{aligned}
 P(Y = 1) &= 1 - P(Y = 0) \\
 &= 1 - \frac{e^{c-\eta}}{1 + e^{c-\eta}} \\
 &= \frac{1}{1 + e^{c-\eta}} \\
 &= \frac{e^{-(c-\eta)}}{1 + e^{-(c-\eta)}} \\
 &= \frac{e^{\eta-c}}{1 + e^{\eta-c}}
 \end{aligned} \tag{21.1.4}$$

线性预测器 η 展开为

$$\eta = x^T \beta = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_p \times x_p \tag{21.1.5}$$

$\eta - c$ 为

$$\begin{aligned}
 \eta - c &= \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_p \times x_p - c \\
 &= (\beta_0 - c) + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_p \times x_p
 \end{aligned} \tag{21.1.6}$$

虽然 β_0 和 c 都是未知参数, 但是可以把 $(\beta_0 - c)$ 看做一个整体并作为截距参数, $\beta_0 - c \Rightarrow \beta_0$, 对于模型来说, $\eta - c$ 和 η 是等价的。从这里可以看出, 在逻辑回归模型中, 线性部分的截距 β_0 是和分割点有关的。最后, 利用分割点分方式推导出的逻辑回归和前面章节逻辑回归的定义是等价的。

核心思想就是, 在线性预测器 η 的空间中有一个分割点, 当 η 的值落在分割点的左侧时, 响应变量 Y 为 0 的概率大, 当 η 的值落在分割点的右侧时, Y 为 1 的概率大, Y 的概率分布与间距 $c - \eta$ 正相关, 两者可以通过一个累计概率分布函数相连接。

假设 η 的整个实数域空间中, 有 $K + 1$ 个分割点, $c_0, c_1, c_2, \dots, c_K$, 并且令 $c_0 = -\infty, c_K = +\infty$, 则整个空间被分成 K 个段, 分别对应着响应变量 Y 的 K 个值。参考公式 (21.1.2), 则有

$$y = \begin{cases} 1 & \text{if } -\infty = c_0 < \eta \leq c_1 \\ 2 & \text{if } c_1 < \eta \leq c_2 \\ 3 & \text{if } c_2 < \eta \leq c_3 \\ \vdots & \\ K & \text{if } c_{K-1} < \eta < c_K = +\infty \end{cases} \tag{21.1.7}$$

假设累计概率分布函数为 F , 其一阶导数也就是对应的概率分布函数用符号 f 表示。任意选择一个分割点 c_k , $F(c_k - \eta)$ 表示响应变量 Y 小于等于类别 k 的概率。注意, $F(c_k - \eta)$ 不是 $Y = k$ 的概率, 而是 $Y \leq k$ 的概率。

$$\begin{aligned}
 P(Y \leq k) &= F(c_k - \eta) \\
 P(Y > k) &= 1 - P(Y \leq k) = 1 - F(c_k - \eta)
 \end{aligned} \tag{21.1.8}$$

$Y = k$ 的概率可以表示成

$$\begin{aligned}
 P(Y = k) &= P(Y \leq k) - P(Y \leq k - 1) \\
 &= F(c_k - \eta) - F(c_{k-1} - \eta)
 \end{aligned} \tag{21.1.9}$$

特殊的 $P(Y = 1)$ 和 $P(Y = K)$ 分别为

$$\begin{aligned}
 P(Y = 1) &= F(c_1 - \eta) - F(c_0 - \eta) \\
 &= F(c_1 - \eta) - F(-\infty - \eta) \\
 &= F(c_1 - \eta)
 \end{aligned} \tag{21.1.10}$$

$$\begin{aligned}
P(Y = K) &= F(c_K - \eta) - F(c_{K-1} - \eta) \\
&= F(+\infty - \eta) - F(c_{K-1} - \eta) \\
&= 1 - F(c_{K-1} - \eta) \\
&= 1 - P(Y \leq K - 1)
\end{aligned} \tag{21.1.11}$$

完整的概率质量函数为

$$f(y) = \prod_{k=1}^K [F(c_k - \eta) - F(c_{k-1} - \eta)]^{\mathbb{I}(y=k)}, \quad c_0 = -\infty, c_K = +\infty \tag{21.1.12}$$

累计分函数 F 可以有多种选择, 如果是 `logistic` 函数, 就是有序逻辑回归模型, 如果是高斯累积分布函数就是 `probit` 有序回归模型, 其它的还有 `log-log`, `clog-log` 等等。

模型的参数除了 β 外, 增加了未知参数 $c^T = [c_1, c_2, \dots, c_{K-1}]$, 由于参数 c 替代线性部分的截距参数 β_0 , 因此有序多分类模型中的线性部分不再需要解决参数 β_0 。此外注意, 所有的类别是共用 β 参数的。

21.2 参数估计

有序多分类模型同样不能使用 `IRLS` 算法进行参数估计, 需要使用牛顿法等完全最大似然估计。其对数似然函数为

$$\ell(y; c, \beta) = \sum_{i=1}^N \sum_{k=1}^K \mathbb{I}(y_i = k) \ln [F(c_k - \eta_i) - F(c_{k-1} - \eta_i)] \tag{21.2.1}$$

对数似然函数的一阶导数为

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^N x_{ij} \sum_{k=1}^K \left[\frac{-f(c_k - \eta_i) + f(c_{k-1} - \eta_i)}{F(c_k - \eta_i) - F(c_{k-1} - \eta_i)} \right] \mathbb{I}(y_i = k) \tag{21.2.2}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial c_k} &= \sum_{i=1}^N \left[\frac{f(c_k - \eta_i)}{F(c_k - \eta_i) - F(c_{k-1} - \eta_i)} \mathbb{I}(y_i = k) \right. \\
&\quad \left. - \frac{f(c_k - \eta_i)}{F(c_{k+1} - \eta_i) - F(c_k - \eta_i)} \mathbb{I}(y_i = k+1) \right], \quad 1 \leq k \leq K-1
\end{aligned} \tag{21.2.3}$$

二阶导数为

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_t} &= \sum_{i=1}^N x_{ij} x_{it} \sum_{k=1}^K \left\{ \frac{f'(c_k - \eta_i) - f'(c_{k-1} - \eta_i)}{F(c_k - \eta_i) - F(c_{k-1} - \eta_i)} \right. \\
&\quad \left. - \frac{[-f(c_k - \eta_i) + f(c_{k-1} - \eta_i)]^2}{[F(c_k - \eta_i) - F(c_{k-1} - \eta_i)]^2} \right\} \mathbb{I}(y_i = k)
\end{aligned} \tag{21.2.4}$$

$$\frac{\partial^2 \ell}{\partial c_k \partial c_t} = - \sum_{i=1}^N \frac{f(c_k - \eta_i) f(c_t - \eta_i)}{[F(c_k - \eta_i) - F(c_t - \eta_i)]^2} \mathbb{I}(y_i = \max(k, t)) \mathbb{I}(|k - t| = 1) \tag{21.2.5}$$

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial c_k \partial c_k} &= \sum_{i=1}^N \left\{ \right. \\
&\quad \left. \left[\frac{f'(c_k - \eta_i)}{F(c_k - \eta_i) - F(c_{k-1} - \eta_i)} - \frac{f(c_k - \eta_i) [f(c_k - \eta_i) - f(c_{k-1} - \eta_i)]}{[F(c_k - \eta_i) - F(c_{k-1} - \eta_i)]^2} \right] \mathbb{I}(y_i = k) \right. \\
&\quad \left. - \left[\frac{f'(c_k - \eta_i)}{F(c_{k+1} - \eta_i) - F(c_k - \eta_i)} - \frac{f(c_k - \eta_i) [f(c_{k+1} - \eta_i) - f(c_k - \eta_i)]}{[F(c_{k+1} - \eta_i) - F(c_k - \eta_i)]^2} \right] \mathbb{I}(y_i = k+1) \right\}
\end{aligned} \tag{21.2.6}$$

$$\frac{\partial^2 \ell}{\partial c_k \partial \beta_t} = - \sum_{i=1}^N [x_{it} \mathcal{A}_i \mathbb{I}(y_i = t) - \mathcal{B}_i \mathbb{I}(y_i = t+1)] \quad (21.2.7)$$

其中

$$\begin{aligned} \mathcal{A}_i &= \frac{f(c_k - \eta_i)[F(c_k - \eta_i) - F(c_{k-1} - \eta_i) + f(c_k - \eta_i) - f(c_{k-1} - \eta_i)]}{[F(c_k - \eta_i) - F(c_{k-1} - \eta_i)]^2} \\ \mathcal{B}_i &= \frac{f(c_k - \eta_i)[F(c_{k+1} - \eta_i) - F(c_k - \eta_i) + f(c_{k+1} - \eta_i) - f(c_k - \eta_i)]}{[F(c_{k+1} - \eta_i) - F(c_k - \eta_i)]^2} \end{aligned} \quad (21.2.8)$$

21.3 连接函数

现在, 我们讨论下有序多分类模型常用的连接函数。上一节中, 用符号 F 表示累积分布函数, 符号 f 表示 F 的导数, 也就是累积分布函数对应的概率密度函数。 F 的作用就类似于逻辑回归模型中的响应函数, F 的反函数就是连接函数。有序多分类模型可以看做逻辑回归模型的一种扩展, 因此逻辑回归模型的连接函数在有序多分类模型中都可以使用。

21.3.1 logit

`logistic` 函数是 `logit` 函数的反函数, 它同时也是标准逻辑分布 (standard logistic distribution) 的累积分布函数。当累积分布函数 F 采用 `logistic` 函数时, 就相当于采用 `logit` 连接函数。

$$F(a) = \frac{\exp(a)}{1 + \exp(a)} \quad (21.3.1)$$

$$\begin{aligned} f(a) &= F'(a) \\ &= \frac{\exp(a)}{[1 + \exp(a)]^2} \\ &= F(a)[1 - F(a)] \end{aligned} \quad (21.3.2)$$

$$\begin{aligned} f'(a) &= \frac{\exp(a) - \exp(2a)}{[1 + \exp(a)]^3} \\ &= F(a)[1 - F(a)]\{F(a) - [1 - F(a)]\} \end{aligned} \quad (21.3.3)$$

21.3.2 probit

当 F 采用累积正态分布时, 就是 `probit` 模型, 用符号 Φ 表示累积正态分布函数, 符号 ϕ 表示正态分布的概率密度函数。

$$\begin{aligned} F(a) &= \Phi(a) \\ f(a) &= \phi(a) \\ f'(a) &= -a\phi(a) \end{aligned} \quad (21.3.4)$$

21.3.3 clog-log

clog-log 连接函数的反函数是广义耿贝尔分布 (generalized Gumbel distribution) 的累积分布函数。 F 也可以是 clog-log 函数反函数。

$$\begin{aligned} F(a) &= 1 - \exp\{-\exp(a)\} \\ f(a) &= [F(a) - 1] \ln[1 - F(a)] \\ f'(a) &= f(a)\{1 + \ln[1 - F(a)]\} \end{aligned} \tag{21.3.5}$$

21.3.4 log-log

当 F 是 log-log 反函数时, 有

$$\begin{aligned} F(a) &= \exp\{-\exp(-a)\} \\ f(a) &= -F(a) \ln[F(a)] \\ f'(a) &= f(a)\{1 + \ln[F(a)]\} \end{aligned} \tag{21.3.6}$$

21.3.5 cauchit

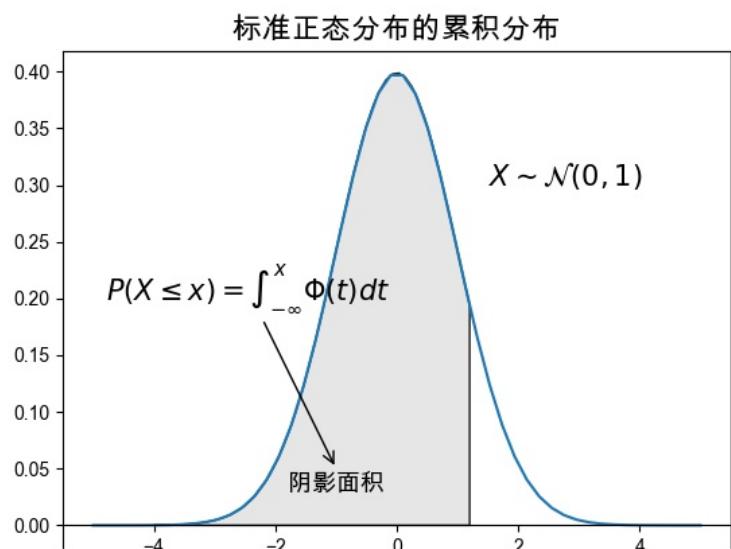
柯西分布 (Cauchy distribution) 的定义域也是整个实数域, 因此柯西分布的累积分布函数也可以拿来用。当响应函数 F 是标准柯西分布的累积分布函数时, 连接函数称为 cauchit。

$$\begin{aligned} F(a) &= 0.5 + \pi^{-1} \operatorname{atan}(-a) \\ f(a) &= -\frac{1}{\pi(1 + a^2)} \\ f'(a) &= f(a)2\pi a \end{aligned} \tag{21.3.7}$$

21.4 总结

附录

22.1 标准正态累积分布表

图 22.1.1: 标准正态分布表表示的是 $X \leq x$ 概率, 即阴影部分面积。

X	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141

下页继续

表 22.1.1 - 续上页

X	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999

22.2 卡方分布临界值表

自由度/显著水平	0.5	0.25	0.1	0.05	0.01	0.005	0.001
1	0.455	1.323	2.706	3.841	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	9.21	10.597	13.816
3	2.366	4.108	6.251	7.815	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	13.277	14.86	18.467
5	4.351	6.626	9.236	11.07	15.086	16.75	20.515
6	5.348	7.841	10.645	12.592	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	20.09	21.955	26.124
9	8.343	11.389	14.684	16.919	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	24.725	26.757	31.264
12	11.34	14.845	18.549	21.026	26.217	28.3	32.909
13	12.34	15.984	19.812	22.362	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	30.578	32.801	37.697

下页继续

表 22.2.1 - 续上页

自由度/显著水平	0.5	0.25	0.1	0.05	0.01	0.005	0.001
16	15.338	19.369	23.542	26.296	32	34.267	39.252
17	16.338	20.489	24.769	27.587	33.409	35.718	40.79
18	17.338	21.605	25.989	28.869	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	36.191	38.582	43.82
20	19.337	23.828	28.412	31.41	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	42.98	45.559	51.179
25	24.337	29.339	34.382	37.652	44.314	46.928	52.62
26	25.336	30.435	35.563	38.885	45.642	48.29	54.052
27	26.336	31.528	36.741	40.113	46.963	49.645	55.476
28	27.336	32.62	37.916	41.337	48.278	50.993	56.892
29	28.336	33.711	39.087	42.557	49.588	52.336	58.301
30	29.336	34.8	40.256	43.773	50.892	53.672	59.703
31	30.336	35.887	41.422	44.985	52.191	55.003	61.098
32	31.336	36.973	42.585	46.194	53.486	56.328	62.487
33	32.336	38.058	43.745	47.4	54.776	57.648	63.87
34	33.336	39.141	44.903	48.602	56.061	58.964	65.247
35	34.336	40.223	46.059	49.802	57.342	60.275	66.619
36	35.336	41.304	47.212	50.998	58.619	61.581	67.985
37	36.336	42.383	48.363	52.192	59.893	62.883	69.346
38	37.335	43.462	49.513	53.384	61.162	64.181	70.703
39	38.335	44.539	50.66	54.572	62.428	65.476	72.055
40	39.335	45.616	51.805	55.758	63.691	66.766	73.402
41	40.335	46.692	52.949	56.942	64.95	68.053	74.745
42	41.335	47.766	54.09	58.124	66.206	69.336	76.084
43	42.335	48.84	55.23	59.304	67.459	70.616	77.419
44	43.335	49.913	56.369	60.481	68.71	71.893	78.75

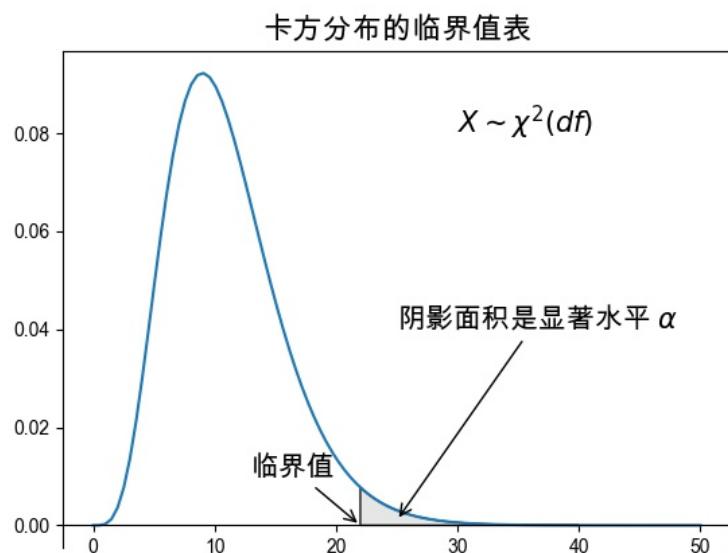


图 22.2.1: 单侧卡方检验临界值

CHAPTER 23

参考文献

1. Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory*. Wiley, Chichester, 2014. ISBN 9781118857502.
2. Stefan Boes and Rainer Winkelmann. Ordered response models. *Allgemeines Statistisches Archiv*, 90(1):167–181, March 2006. URL: <http://link.springer.com/10.1007/s10182-006-0228-y>, doi:10.1007/s10182-006-0228-y.
3. Annette J. Dobson and Adrian G. Barnett. *An introduction to generalized linear models*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton, 3rd ed edition, 2008. ISBN 9781584889502. OCLC: ocn213602344.
4. Ludwig Fahrmeir and Heinz Kaufmann. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics*, 13(1):342 –368, 1985. URL: <https://doi.org/10.1214/aos/1176346597>, doi:10.1214/aos/1176346597.
5. Jeff Gill. *Generalized linear models: a unified approach*. Number v. 134 in Quantitative applications in the social sciences. Sage Publications, Inc, Thousand Oaks, Calif, 2001. ISBN 9780761920557.
6. Susanne Gschlößl and Claudia Czado. Modelling count data with overdispersion and spatial effects. *Statistical Papers*, 49(3):531–552, July 2008. URL: <http://link.springer.com/10.1007/s00362-006-0031-6>, doi:10.1007/s00362-006-0031-6.
7. James W. Hardin and Joseph M. Hilbe. *Generalized linear models and extensions*. Stata Press, College Station, Texas, fourth edition edition, 2018. ISBN 9781597182256.
8. Joseph M. Hilbe. *Negative binomial regression*. Cambridge University Press, Cambridge, UK ; New York, 2nd ed edition, 2011. ISBN 9780521198158. OCLC: ocn694679188.
9. John Hinde and Clarice Demetrio. *Overdispersion: models and estimation*. CRC, Boca Raton, Fla.; London, 2006. ISBN 9781584882893. OCLC: 249290191.
10. John Hinde and Clarice G.B. Demétrio. Overdispersion: Models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170, April 1998. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167947398000073>, doi:10.1016/S0167-9473(98)00007-3.
11. Silvie Kafková and Lenka Křivánková. Generalized Linear Models in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2):383–388, 2014. URL: <https://acta.mendelu.cz/62/2/0383/>, doi:10.11118/actaun201462020383.

12. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Springer US, Boston, MA, 1989. ISBN 9780412317606 9781489932426. URL: <http://link.springer.com/10.1007/978-1-4899-3242-6>, doi:10.1007/978-1-4899-3242-6.
 13. Jorge G. Morel and Nagaraj K. Neerchal. *Overdispersion models in SAS*. SAS Institute, Cary, N.C, 2012. ISBN 9781607648819. OCLC: ocn777626941.
 14. Hui Zhang, Stanley B. Pounds, and Li Tang. Statistical Methods for Overdispersion in mRNA-Seq Count Data. *The Open Bioinformatics Journal*, 7(1):34–40, December 2013. URL: <https://openbioinformaticsjournal.com/VOLUME/7/PAGE/34/>, doi:10.2174/1875036201307010034.
 15. Yunlong Zhang, Zhirui Ye, and Dominique Lord. Estimating Dispersion Parameter of Negative Binomial Distribution for Analysis of Crash Data: Bootstrapped Maximum Likelihood Method. *Transportation Research Record: Journal of the Transportation Research Board*, 2019(1):15–21, January 2007. URL: <http://journals.sagepub.com/doi/10.3141/2019-03>, doi:10.3141/2019-03.
- matplotlib 支持的 latex 符号 <https://matplotlib.org/3.3.3/tutorials/text/mathtext.html>

非字母

统计量, **41**

K

KL 散度 (Kullback–Leibler divergence), **80**

V

嵌套模型 (nested model), **118**

W

最 大 似 然 估 计 (maximum likelihood estimation, MLE), **39**

概率公理, **4**