logistic 回归

授课教师：吴翔
邮箱：wuhsiang@hust.edu.cn

March 27, 2021

1. 二项 logistic 回归

2. 多项 logistic 回归

3. 次序 logistic 回归（略）

Section 1

二项 logistic 回归

课程存储地址

- 课程存储地址：https://github.com/wuhsiang/Courses
- 资源：课件、案例数据及代码

图 1：课程存储地址

参考教材

- 丹尼尔·鲍威斯，谢宇．分类数据分析的统计方法（第二版）．北京：社会科学文献出版社．2018.

数据的测量类型

- 定量测量：数值有实质含义。包括连续变量（或定距变量）、离散变量（通常是计数变量）。
- 定性测量：数值**无实质含义**。包括次序变量和名义变量。
- 实践中的处理：李克特量表

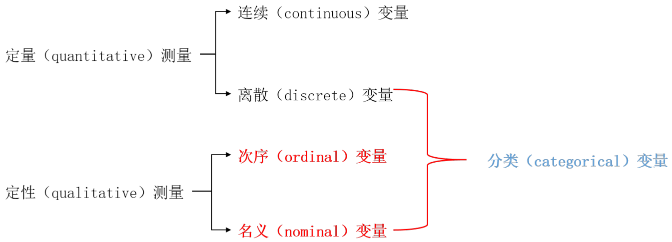


图 2: A typology of measurement

线性回归回顾

线性回归中，一组预测变量向量 X 只对应一个预测值 $\hat{y}$，总体回归线穿过 $(X^k, E(y|X^k))$。

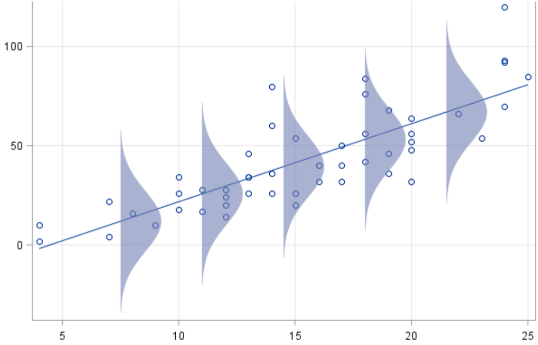


图 3: Linear regression line

分类因变量与线性回归模型

线性回归模型

$$y = \beta X + \epsilon$$

最关键的推导和设定包括两步：

$$E(y|X) = \beta X + E(\epsilon|X), \text{ and } E(\epsilon|X) = 0.$$

从而剥离出误差项 ϵ，并通过普通最小二乘法（OLS）得到最佳线性无偏估计量（best linear unbiased estimator, BLUE）。

$E(y|X)$ 对分类因变量不适用，因此分类因变量需要新的统计模型！

分类因变量与 logistic 回归

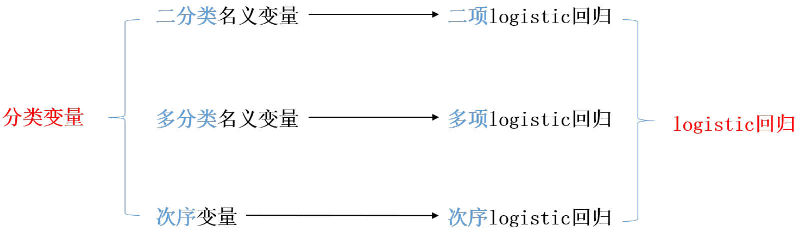


图 4: Categorical dependent variables and logit models

二分类因变量

因变量只能在两个可能的数值中取值，要么“是”或者“发生”，要么“否”或者“未发生”。例如，患病、犯罪、抑郁、自杀等健康管理研究议题。

二分类因变量 (binary dependent variable)：取值为二分类，两种可能结果被描述为“发生”或者“不发生”。研究关注的结果视作“发生”，且编码为 1；另一结果则被视为“不发生”，且编码为 0。即因变量 $y \in 0, 1$。但 0 和 1 不具有数值上的实质意义。

研究者目的在于，估计或预测事件发生的概率如何受到自变量的影响。相应地，每个独立样本可以视作一次**伯努利试验**（Bernoulli trial），试验结果要么是 1（发生），要么是 0（不发生）。

线性概率模型

研究者目的在于，估计或预测事件发生的概率p 如何受到自变量的影响。

线性概率模型（linear probability model, LPM）**直接**用自变量 X 来解释事件发生概率 p：

$$p_i = \beta X + \epsilon_i.$$

但 LPM 存在异方差问题，同时预测值 $\hat{p}_i$ 很可能落在 $[0, 1]$ 区间以外。因而，LPM 随即被 logit 和 probit 模型取代。

发生比（odds）

事件的发生比（odds，也称发生比率、比数），定义为事件发生的概率 p 与不发生的概率 $(1-p)$ 的比率：

$$\text{odds} = \frac{p}{1-p}.$$

此时 odds $\in [0, \infty]$。

进一步，对数发生比（log-odds），也称为发生概率 p 的logit：

$$\text{logit}(p) = \log(\frac{p}{1-p}).$$

显然，$\text{logit}(p) \in (-\infty, \infty)$。$p \to 0$ 时，$\text{logit}(p) \to -\infty$；$p \to 1$ 时，$\text{logit}(p) \to \infty$。

二项 logistic 回归

二项 logistic 回归认为，$\text{logit}(p_i)$ 是自变量 X_i 的线性函数。

$$\log(\frac{p_i}{1-p_i}) = \text{logit}(p_i) = \beta X_i + \epsilon_i.$$

从而，事件发生概率

$$p_i = \text{logistic}(\beta X_i) = \frac{\exp(\beta X_i)}{1+\exp(\beta X_i)}.$$

注：$\text{logit}(\cdot)$ 与 $\text{logistic}(\cdot)$ 互为逆函数（inverse function）。

通过最大似然估计（maximum likelihood estimation, MLE）方法，得到参数 β 的估计值 $\hat{\beta}$。

二项 logistic 回归案例

考虑如下问题：**哪些民众更倾向使用互联网作为健康信息来源？** 当观测样本 i 使用互联网作为健康信息来源时，记作 $y_i = 1$；否则，记作 $y_i = 0$。将所有其它变量纳入模型作为自变量，用以解释民众使用互联网作为健康信息来源的概率 p。

$$\text{logit}(p_i) = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Race_i + \beta_4 Educ_i + \beta_5 Inc_i + \epsilon_i.$$

进一步将分类自变量其虚拟变量化，得到最终二项 logistic 回归模型中，最后一项 $\beta_5 Inc_i$ 则变成两个虚拟变量项：

$$\beta_5 IncM_i + \beta_6 IncH_i$$

案例更多细节，详见二项 logistic 回归案例：健康信息搜寻行为

二项 logistic 回归结果

```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.43365    0.17298   -2.51   0.0122 *
## age                          -0.05043    0.00431  -11.69  < 2e-16 ***
## genderMale                   -0.03720    0.11918   -0.31   0.7550
## raceWhite                     0.64694    0.14190    4.56  5.1e-06 ***
## educationCollege and above    0.37010    0.12278    3.01   0.0026 **
## income$20,000 to $74,999      0.87564    0.16555    5.29  1.2e-07 ***
## income$75,000 or more         1.26223    0.18502    6.82  9.0e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2124.4  on 1813  degrees of freedom
## Residual deviance: 1813.9  on 1807  degrees of freedom
## AIC: 1828
```

图 5: Estimated parameters

参数解释

二项 logistic 回归得到如下参数估计结果：

$$\text{logit}(p_i) = -0.43 - 0.05Age_i + 0.65RaceW_i + 0.37EducH_i \\ + 0.88IncM_i + 1.26IncH_i + \epsilon_i.$$

然而，该如何解释参数 β 的含义？我们分以下四个情形逐一讨论：

- 截距项
- 二分类自变量：种族、教育水平
- 多分类自变量：收入
- 连续自变量：年龄

参数解释：截距项

截距项 $\hat{\beta}_0 = -0.43$，其含义为：当其它变量取值均为 0 的时候，对数发生比 $\log(\text{odds}) = -0.43$。

其它变量取值均为 0 的含义：年龄 55 岁（样本均值）、非白人种族、教育程度在大学以下、年收入在 20,000 美元以下。

$\log(\text{odds}) = -0.43$ 的含义是，使用互联网获取健康信息的概率与不使用互联网获取健康信息的概率之比（odds）为：

$$\text{odds} = \exp(-0.43) = 0.65.$$

注：年龄进行对中（centering）处理之前，截距项估计值 $\hat{\beta}_0 = 2.35$，可计算相应的对数发生比 $\text{odds} = \exp(2.35) = 10.49$，但此数值无法解释。

参数解释：二分类自变量

二分类自变量种族（$\hat{\beta}_3 = 0.65$）和教育水平（$\hat{\beta}_4 = 0.37$）的系数的含义是什么?

其它条件不变（ceteris paribus）时，种族变量对民众使用互联网获取健康信息的对数发生比的净效应是 0.65。换言之，其它条件不变时，非白人族群使用互联网获取健康信息的对数发生比记为 α_0，那么相应的白人族群对应的对数发生比为 $\alpha_0 + 0.65$。由此，白人族群使用互联网获取健康信息的发生比（odds）与非白人族群使用互联网获取健康信息的发生比的比率为：

$$\frac{\text{odds}_{RaceW}}{\text{odds}_{RaceNW}} = \frac{\text{exp}(\alpha_0 + 0.65)}{\text{exp}(\alpha_0)} = \exp(0.65) = 1.91.$$

因此，$\hat{\beta}_3 = 0.65$ 的含义是：白人族群使用互联网获取健康信息的发生比是非白人族群的 1.91 倍。

参数解释：多分类自变量

多分类自变量收入（$\hat{\beta}_5 = 0.88$，$\hat{\beta}_6 = 1.26$）的系数的含义是什么?

其它条件不变时，低收入群体使用互联网获取健康信息的对数发生比记为 α_0。那么，中等收入群体和高收入群体的对数发生比相应为 $\alpha_0 + 0.88$ 和 $\alpha_0 + 1.26$。类似地，

$$\frac{\text{odds}_{IncM}}{\text{odds}_{IncL}} = \exp(0.88) = 2.41, \text{ and } \frac{\text{odds}_{IncH}}{\text{odds}_{IncL}} = \exp(1.26) = 3.53.$$

课堂讨论：中等收入群体和高收入群体之间是否可比?

参数解释：连续自变量

连续自变量年龄（$\hat{\beta}_1 = -0.05$）的系数的含义是什么?

其它条件不变时，由于没有参考水平，我们取年龄为 x_0，而除年龄以外的其它项对应的对数发生比记为 α_0。那么，年龄 x_0 的群体使用互联网获取健康信息的对数发生比为 $\alpha_0 + \hat{\beta}_1 x_0$；而年龄增加 1 岁，相应的对数发生比为 $\alpha_0 + \hat{\beta}_1(x_0 + 1)$。由此，

$$\frac{\text{odds}_{x_0+1}}{\text{odds}_{x_0}} = \frac{\text{exp}[\alpha_0 + \hat{\beta}_1(x_0 + 1)]}{\text{exp}(\alpha_0 + \hat{\beta}_1 x_0)} = \exp(\hat{\beta}_1) = \exp(-0.05) = 0.95.$$

因此，年龄每增加 1 岁，使用互联网获取健康信息的发生比降低 5%。

参数解释：事件发生概率的预测

问题：一位年龄 50 岁的白人男性民众，受教育程度在大学以下，年收入为 20,000 至 74,999 美元区间（中等收入水平)。请问他使用互联网获取健康信息的概率是多少?

分析：首先预测对数发生比

$$\text{logit}(\hat{p}_i) = \log(\widehat{\text{odds}}_i) = -0.43 - 0.05 \times (50 - 55) + 0.65 + 0.88 = 1.35.$$

从而发生比的预测值为 $\widehat{\text{odds}}_i = \exp(1.35) = 3.86$，该民众使用互联网获取健康信息的概率预测值为 $\hat{p}_i = 3.86/4.86 = 0.79$。

发生比的比率（odds ratio）

以上各种类型的自变量（二分类自变量、多分类自变量、连续变量）的系数解释时，都使用了以下概念：发生比的比率（odds ratio, OR）。

假定有 A 组和 B 组，我们通常会考虑两组的发生比的比率：

$$\mathrm{OR} = \frac{\mathrm{odds}_A}{\mathrm{odds}_B}.$$

使用 OR 解释系数含义更加直观，因此我们通常报告 OR 及相应的 CI。

课堂讨论：解释以上各个系数时，相应的 A 组和 B 组是什么？

报告 OR 及相应的 CI

$\hat{\beta} > 0$，则有 $OR > 1$；若 β 系数显著不等于 0，则 OR 的置信区间（confidence interval，CI）不包含 1。

	Estimate	OR	2.5 %	97.5 %	Pr(>|z|)
intercept	-0.43	0.65	0.46	0.91	0.012
age	-0.05	0.95	0.94	0.96	1.4e-31
male	-0.037	0.96	0.76	1.2	0.75
white	0.65	1.9	1.4	2.5	5.1e-06
college and above	0.37	1.4	1.1	1.8	0.0026
$20,000 to 74,999	0.88	2.4	1.7	3.3	1.2e-07
$75,000 or more	1.3	3.5	2.5	5.1	9e-12

图 6: Coefficients, OR, and corresponding CI

似然函数

问题：箱子里有 10 个球，或是白球，或是黑球。从中有放回地取出 5 个球，得到结果：$\{$白球、白球、白球、黑球、白球$\}$。请估计，箱子中有几个白球、几个黑球?

建模：令 $p \in [0,1]$：箱子中白球的比例，事件 A：取出的球是白球。那么，单次伯努利试验中事件 A 发生的概率为 p。样本观测值为：{1,1,1,0,1}。

分析：给定参数 p，得到以上观测数据 D 的概率是，
$\text{Prob}(D|p) = p \times p \times p \times (1-p) \times p = p^4(1-p)$。

以上概率是未知参数 p 的函数，称为**似然函数**（likelihood function），表述为 $L(p) = \text{Prob}(D|p) = p^4(1-p)$。

最大似然估计

更一般化，给定**参数**θ 和**观测数据**D，似然函数 $L(\theta) = \text{Prob}(D|\theta)$ 是未知参数 θ 的函数，刻画了给定参数 θ 时观测到数据 D 的概率。

最大似然估计（maximum likelihood estimation，MLE）的逻辑：找到 $\theta = \hat{\theta}$，使似然函数 $L(\theta)$ 取最大值。换言之，使得数据 D 以最大可能性被观测到的参数值 $\hat{\theta}$ 即为最大似然估计值。通常 $L(\theta) \in (0, 1)$ 极小，因而参数估计时使用其对数 $LL(\theta)$。

以上例子中，$LL(p) = 4\text{log}(p) + \text{log}(1-p)$。当 $p = 0.8$ 时，$LL(p)$ 取得最大值。因此，我们估计箱子中白球的比例是 $\hat{p} = 0.8$，亦即箱子中有 8 个白球、2 个黑球。

二项 logistic 回归的参数估计

二项 logistic 回归的似然函数

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)}$$

进一步，对数似然函数

$$LL(\beta) = \sum_{i=1}^{n}[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)].$$

可以使用最大似然估计得到参数估计值 $\hat{\beta}$。

模型拟合优度

logistic 回归模型采用最大似然估计方法估计参数，因而模型拟合优度也应基于似然函数。

假定模型中待估计的参数个数为 k，样本量为 n，似然函数的最大值为 $\hat{L}$，则可计算 AIC（Akaike information criterion）和 BIC（bayesian information criterion）统计量：

$$\text{AIC} = 2k - 2LL \text{ and } \text{BIC} = \log(n) \cdot k - 2LL.$$

此外，McFadden 提出的伪 R^2 统计量为：

$$\text{pseudo } R^2 = \frac{LL_0 - LL_c}{LL_0}.$$

其中 LL_0 为空模型对应的对数似然值。

评估模型拟合优度

在以上例子中，样本量 $n = 1814$，参数个数 $k = 7$，对数似然函数最大值 $LL = -907$。因而，

$$\text{AIC} = 2 \times 7 - 2 \times (-907) = 1828.$$

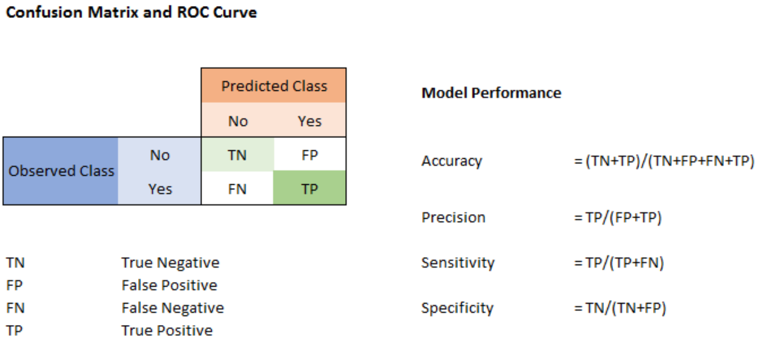
以及

$$\text{BIC} = \log(1814) \times 7 - 2 \times (-907) = 1866.$$

最后估计空模型，得到相应对数似然值 $LL_0 = -1257$，由此得到

$$\text{pseudo } R^2 = \frac{LL_0 - LL_c}{LL_0} = 0.28.$$

预测准确率?



Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

图 7: confusion matrix

如何报告二项 logistic 回归结果?

Section 2

多项 logistic 回归

多项 logistic 回归

当因变量取值是多分类变量时，需要使用多项 logistic 回归模型。其基本逻辑是：一次比较两个结果。

假定因变量有 J 个类别，我们将第 $j(1 \leq j \leq J)$ 个分类与第一个分类（参考水平，reference level）进行比较，从而得到第 j 个分类的基线 logistic 回归模型：

$$\mathrm{BL}_j = \log[\frac{\mathrm{Prob}(y=j)}{\mathrm{Prob}(y=1)}] = \beta_j X + \epsilon_j.$$

估计方法和其余细节都与二项 logistic 回归模型类似。

多项 logistic 回归案例

考虑如下问题：**哪些因素影响了民众选择健康信息来源?**

$$\text{logit}(p_j/p_1) = \beta_{0j} + \beta_{1j}Age_i + \beta_{2j}Gender_i + \beta_{3j}Race_i + \beta_{4j}Educ_i + \beta_{5j}Inc_i +$$

案例更多细节，详见多项 logistic 回归案例: 健康信息搜寻行为

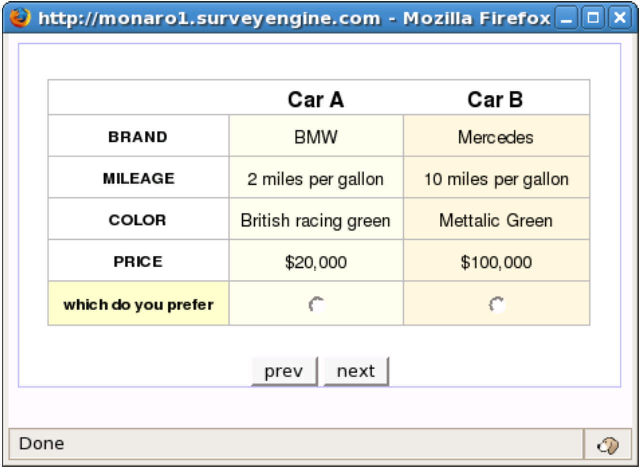
离散选择模型

消费决策过程中，通常面临几个候选项。这与多项 logistic 回归模型的设定是一致的。

同时，每个候选项可以由具体的属性刻画，例如产品属性。

给定候选项及其属性的时候，消费者如何决策?

离散选择试验



http://monaro1.surveyengine.com - Mozilla Firefox

	Car A	Car B
BRAND	BMW	Mercedes
MILEAGE	2 miles per gallon	10 miles per gallon
COLOR	British racing green	Mettalic Green
PRICE	$20,000	$100,000
which do you prefer		

prev next

Done

图 8: An example of discrete choice experiment

Section 3

次序 logistic 回归（略）

次序 logistic 回归（略）

logistic 回归总结

1. 二项 logistic 回归
2. 多项 logistic 回归
3. 次序 logistic 回归（略）