北京航空航天大学
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 7 讲：Logistic 回归

康雁飞

数量经济与商务统计系

Outline

Outline

Logistic 回归的目的

- 因变量为分类变量的回归模型
- 它是广义线性模型（Generalized Linear Models，简称 GLMs）的一种
- 它可以输出样本点属于某一类的概率
- 例如：降雨概率，违约概率，股票涨跌概率
- 是一种实践中尤其常用的分类模型

回忆线性回归模型

- OLS：对回归模型中的自变量、回归系数以及残差项的取值都没有任何限制，作为自变量函数的因变量就必须能够在 $(-\infty, +\infty)$ 范围内自由取值。
- 如果因变量只取分类值，或者只取两类值，就会严重违反因变量为连续型变量的假设。

Outline

Logistic 回归

假设因变量 Y_i 取值 0 或者 1，并且 $P(Y_i = 1) = p_i$（伯努利分布），

我们可以把 p_i 与自变量联系起来：

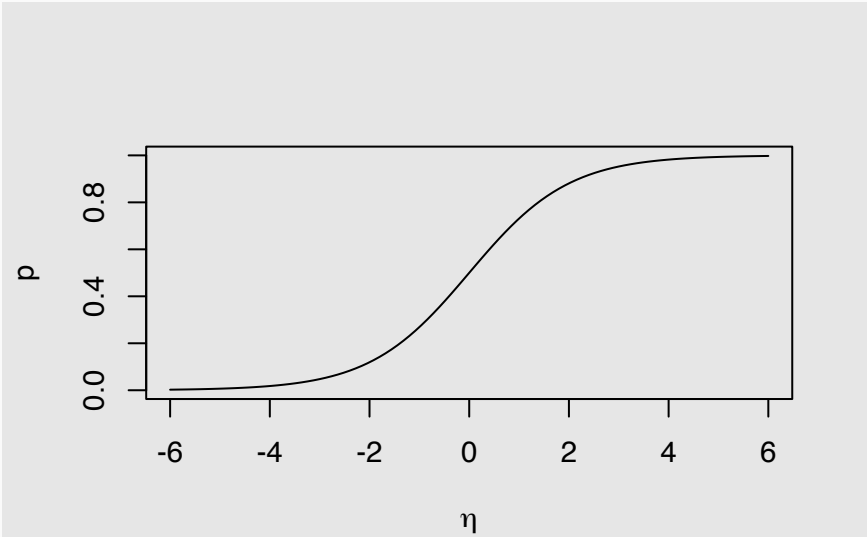
$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$
$$\eta_i = g(p_i)$$

- 链接函数 g 是单调的。
- 对任意的 η，$0 \leq g^{-1}(\eta) \leq 1$。

链接函数: *logit* 函数

- $\eta = g(p) = \log(\frac{p}{1-p})$ 叫作 *logit* 链接函数，完成映射 $(0,1) \to \mathbb{R}$。
- 通过链接函数克服了概率因变量的取值受到限制的困难。
- 逆 logit 就是 $g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$，完成映射 $\mathbb{R} \to (0,1)$。
- $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = P(Y_i = 1)$。

逆 logit 函数



Outline

对数似然函数

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_q x_{i,q}$$

$$\begin{aligned}
\log L(\beta) &= \sum_{i=1}^{n} \log(p_i) 1_{Y_i=1} + \log(1-p_i) 1_{Y_i=0} \\
&= \sum_{i=1}^{n} y_i[\eta_i - \log(1+e^{\eta_i})] + (1-y_i)\log\left(1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}\right) \\
&= \sum_{i=1}^{n} y_i[\eta_i - \log(1+e^{\eta_i})] + (1-y_i)\log\left(\frac{1}{1+e^{\eta_i}}\right) \\
&= \sum_{i=1}^{n} y_i[\eta_i - \log(1+e^{\eta_i})] - (1-y_i)\log(1+e^{\eta_i}) \\
&= \sum_{i=1}^{n} [y_i\eta_i - \log(1+e^{\eta_i})]
\end{aligned}$$

R 中的 Logistic 回归模型估计

```
service.logis <- glm(OPINION ~ AGE + SEX,
                     family=binomial,
                     data=service)
```

- 伯努利分布等价于只有两个水平（0 或者 1）的二项分布。
- glm 默认使用 logit 链接函数和极大似然估计（当 family=binomial 时）。

Logistic 回归模型中如何解释?

$$\text{Odds} = p/(1-p).$$

- 优势比（Odds）是无界的。
- 对数优势比：log(odds) = $\log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$。
- x_2 不变的情况下，x_1 每增加一个单位，对数优势比增加 β_1 个单位。
- x_2 不变的情况下，x_1 每增加一个单位，优势比增加 e^{β_1} 个单位。

Outline

模型评价

$$\text{AIC} = -2\log L + 2q$$

- AIC 取值越小越好。
- Logistic 回归方程求解采用最大似然估计方法，其似然函数值表达的是一种概率，即在假定模型为真实情况时，能够观察到这一特定样本数据的概率。因此，这个数据处于 $[0, 1]$ 之间，数值往往极小，取对数是一个负数。所以通常采用 $-2\log L$。
- 为什么 $+2q$?

模型的显著性检验

- 似然比统计量: $2\log\frac{L_l}{L_s}$, 其中 L_l 和 L_s 分别是一个具有 l 和 s 个变量的模型 ($s < l$)。
- 似然比检验: $2\log L_l - 2\log L_s \sim \chi^2_{l-s}$.

```
service.logis <- glm(OPINION ~ AGE + SEX, family=binomial,
            data=service)
anova(service.logis, test="Chisq")
drop1(service.logis, test="Chisq")
service.logis2 <- glm(OPINION ~ AGE, family='binomial',
         data=service)
anova(service.logis, service.logis2, test="Chisq")
```

回归系数的置信区间和假设检验

- MLE 渐近正态。
- β_i 的 $100(1-\alpha)\%$ 置信区间为：

$$\hat{\beta}_i \pm z^{\alpha/2} se(\hat{\beta}_i).$$

- R：`confint()`.

二分类的 Logistic 回归

- 我们得到的是 p_i，如何根据 p_i 进行分类?
- 模型准确率的判断方法

分类表（Classification Table）或混淆矩阵（Confusion Matrix）：把案例分成预测事件发生或不发生。建立一个 2×2 的交互表，来比较预测情况和实际发生的情况。

多项 Logistic 回归

得到的是判为每一类的概率。