概率统计 B
第七章 回归分析方法

根据李东风老师课件修改

2017 春季学期

本节目录

回归分析方法

- 回归分析方法是数理统计的重要工具，是处理多个变量之间**相关关系**的一种数学方法。
- **函数关系**: 确定性关系。如自由落体运动

$$s = \frac{1}{2}gt^2 \quad (0 \le t \le T)$$

- **相关关系**是给定了 x 的值后并不能确定 y 的值，但 y 的值与 x 的值有关。
- 即使是确定性关系的变量，其测量值因为含有误差所以也有不确定性。
- 回归分析可以建立变量间的关系的数学表达式（经验公式），并可判断这样的公式的有效性，以及如何利用所得到的经验公式去达到预测、控制等目的。

一元线性回归

- 在一元线性回归分析中，考察随机变量 Y 与一个普通变量 x(非随机) 之间的联系。
- 数据成对观测：

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

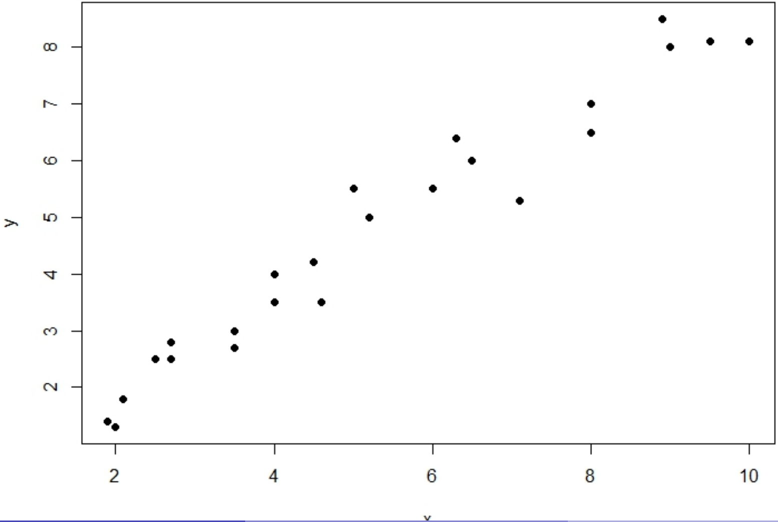
例 1.1

- **例 1.1** 某种合成纤维的强度与其拉伸倍数有关。
- 有 24 个纤维样品的强度与相应的拉伸倍数的实测记录。
- 设拉伸倍数为 x，强度为 Y，希望根据观测数据找出 x 和 Y 的关系式。
- 数据（部分）如

$$(1.9, 1.4), (2.0, 1.3), (2.1, 1.8), \ldots, (9.5, 8.1), (10.0, 8.1)$$

- 一种直观的考察方式是**散点图**（演示）。

纤维强度对拉伸倍数的散点图



- 散点图中每个点以 x 为横坐标，以 y 为纵坐标。
- 从散点图看散点围绕在一条直线周围，有

$$\hat{y} = a + bx \tag{1.1}$$

其中 $\hat{y}$ 表示建立 Y 与 x 的关系后用 x 对 Y 做的预测。

- 于是借助于散点图确定了经验公式的形式。只需要确定 (1.1) 中的 a 和 b。
- b 叫做**回归系数**，关系式 $\hat{y} = a + bx$ 叫做**回归方程**。
- **线性**: x 每增加 1，y 的变化量是恒定的。
- **非线性**: x 每增加 1，y 的变化量不是恒定的。

求解回归直线

- 要找一条直线与散点图中所有点尽可能最接近。
- 直接作图过于粗略，且无法推广到多个自变量的情形。
- 定义距离：用

$$[y_i - (a + bx_i)]^2$$

衡量点 (x_i, y_i) 到直线 $\hat{y} = a + bx$ 的距离。

- 这是两个纵坐标的距离平方，不是点到直线的垂直距离。
- **理由**: 需要衡量的是对 Y 的预测精度。

最小二乘原则

- 平方和

$$Q(a,b)=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2 \tag{1.2}$$

衡量直线 $\hat{y}=a+bx$ 与所有散点的距离远近。

- 求回归直线问题化为：找两个数 $\hat{a},\hat{b}$，使得二元函数 $Q(a,b)$ 在 $a=\hat{a},b=\hat{b}$ 处达到最小。
- 这种方法叫做**最小二乘法**。

最小二乘求解的微分法

- 为了求 $Q(a,b)$ 的最小值点，求解其一阶偏导数都等于零的方程：

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}[y_i-(a+bx_i)] = 0 \tag{1.3}$$

$$\frac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}[y_i-(a+bx_i)]\cdot x_i = 0 \tag{1.4}$$

- 由 (1.3) 解得

$$na = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i$$

$$a = \bar{y} - b\bar{x} \tag{1.5}$$

- 由 (1.4) 得

$$\sum_{i=1}^{n} x_i[y_i - a - bx_i] = 0$$

- 把 (1.5) 代入上式可得

$$\sum_{i=1}^{n} x_i[y_i - \bar{y} - b(x_i - \bar{x})] = 0$$

$$b\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

$$b\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.6}$$

- 代入 (1.5) 可得 $\hat{a} = \bar{y} - \hat{b}\bar{x}$。

- 当 $x_1, x_2, \ldots, x_n$ 不全相等时有解。
- 可以证明这样用微分法求得的 $(\hat{a}, \hat{b})$ 是 $Q(a, b)$ 的最小值点。
- 事实上，二阶导数矩阵，即海色阵为

$$H = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2\right) \end{pmatrix}$$

易见其主子式都大于零，H 正定，$Q(a, b)$ 的唯一的一阶偏导数等于零的点一定是全局最小值点。

最小二乘解的配方法

- 拆分平方与交叉项：

$$
\begin{aligned}
Q(a,b) &= \sum_{i=1}^{n}\{(y_i-\bar{y})+[\bar{y}-(a+b\bar{x})]-b(x_i-\bar{x})\}^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2+n[\bar{y}-(a+b\bar{x})]^2 \\
&\quad + b^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})
\end{aligned}
$$

- 其中 $[\bar{y}-(a+b\bar{x})]$ 是常数，所以它与 $x_i-\bar{x}$ 和 $y_i-\bar{x}$ 的交叉项为零。

- 记

$$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad l_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$l_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- 则

$$\begin{aligned} Q(a,b) =& l_{yy} + n[\bar{y} - (a + b\bar{x})]^2 + l_{xx}b^2 - 2l_{xy}b \\ =& l_{yy} + n[\bar{y} - (a + b\bar{x})]^2 + l_{xx}\left(b - \frac{l_{xy}}{l_{xx}}\right)^2 - \frac{l_{xy}^2}{l_{xx}} \\ \geq& l_{yy} - \frac{l_{xy}^2}{l_{xx}} \end{aligned}$$

- 等于号成立当且仅当

$$b = \frac{l_{xy}}{l_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, a = \bar{y} - b\bar{x}$$

- $Q(a,b)$ 的最小值为

$$
\begin{aligned}
Q(\hat{a},\hat{b}) =& l_{yy} - \frac{l_{xy}^2}{l_{xx}} \\
=& l_{yy} - \hat{b} l_{xy} \\
=& \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{b} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\end{aligned}
$$

- 得到了 $\hat{a},\hat{b}$ 就确定了回归的经验公式，确定了回归直线。
- 易见点 $(\bar{x},\bar{y})$ 落在回归直线上：

$$
\bar{y} = \hat{a} + \hat{b}\bar{x}
$$

- 回归一般使用统计软件计算。比如在 R 中

```
lm1 <- lm(y ~ x)
summary(lm1)
plot(lm1)
```

可以计算并显示回归结果、画回归诊断图形。

- 对于例 1.1 的 24 个点，计算得 $\hat{b} = 0.859, \hat{a} = 0.15$，纤维强度 (Y) 与拉伸倍数 (x) 的经验公式为

$$\hat{y} = 0.15 + 0.859x$$

- 经验公式也叫**回归方程**, 相应的直线叫做**回归直线**。
- 回归系数 b 的含义：拉伸倍数 (x) 每增加一个单位，强度 (Y) 平均增加 0.859 个单位。

非线性关系线性化

- 某些非线性关系可以通过变换转化为线性关系。
- **例 1.2** 彩色显影中，染料光学密度 Y 与析出银的光学密度 x 有如下类型的关系

$$Y \approx Ae^{-B/x}, \quad B > 0$$

- 这不是线性关系。两边取对数得

$$\ln Y \approx \ln A - B\frac{1}{x}$$

- 令

$$Y^* = \ln Y \qquad\qquad x^* = \frac{1}{x}$$

- 则 $Y^* \approx \ln A - Bx^*$ 为线性关系。

- 从 n 组数据 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 得到变换的数据 $(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)$。
- 对变换后的数据建立线性回归方程

$$\hat{y}^* = \hat{a} + \hat{b}x^*$$

- 反变换得

$$\hat{A} = e^{\hat{a}} \qquad \hat{B} = -b$$

- 则有

$$\hat{Y} = \hat{A}e^{-\hat{B}/x}$$

- **例 1.3** 炼钢钢包随使用次数增加而容积增大。
- 测量了 13 组这样的数据（部分）：

$$(2, 106.42), (3, 108.20), (4, 109.58), \ldots, (19, 111.20)$$

- 画出了散点图（演示）。用双曲线

$$\frac{1}{y} \approx a + b\frac{1}{x}$$

- 令 $x^* = 1/x, y^* = 1/y$，化为线性模型

$$y^* \approx a + bx^*$$

- 解得 $\doteq 0.008967$, $\hat{b} = 0.0008292$, 经验公式为

$$\frac{1}{\hat{y}} = 0.008967 + 0.0008292\frac{1}{x}$$

线性相关性

- 只要数据中 $x_1, x_2, \ldots, x_n$ 不全相等，最小二乘法存在唯一解，总可以得到经验公式

$$\hat{y} = \hat{a} + \hat{b}x$$

- 所以，经验公式并不都能反映实际情况。
- 需要判别 x 与 Y 之间是否真的具有线性相关关系：Y 是否随着 x 增大而线性地增大（或者线性地减小）。

平方和分解公式

- 对于任意 n 组数据 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，只要 $x_1, x_2, \ldots, x_n$ 不全相等，就有

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{1.7}$$

- 其中 $\bar{y}$ 是 $y_1, y_2, \ldots, y_n$ 的平均值，

$$\hat{y}_i = \hat{a} + \hat{b}x_i \quad (i = 1, 2, \ldots, n)$$

- 证

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$
$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

- 注意

$$\begin{aligned}\hat{y}_i=&\hat{a}+\hat{b}x_i=\bar{y}-\hat{b}\bar{x}+\hat{b}x_i\\=&\bar{y}+\hat{b}(x_i-\bar{x})\end{aligned}$$

- 所以交叉项

$$
\begin{aligned}
&\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \\
=&\sum_{i=1}^{n}[y_i-\bar{y}-\hat{b}(x_i-\bar{x})]\hat{b}(x_i-\bar{x}) \\
=&\hat{b}\left\{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-\hat{b}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \\
=&0
\end{aligned}
$$

- 于是

$$
\begin{aligned}
&\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
=&\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2
\end{aligned}
$$

- 即平方和分解公式 (1.7) 成立。

平方和分解公式的解释

- (1.7) 式坐标的 $\sum_{i=1}^{n}(y_i - \bar{y})^2$ 是因变量的离差（偏差）平方和（Corrected Sum of Squares），描述了因变量的分散程度，是我们要用模型解释的目标。记为 l_{yy}。
- 考虑分解的第一项 $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$。这是用模型得到的因变量**拟合值** $\hat{y}_i$ 与实际因变量值 y_i 之间的差距的一个度量，是最小二乘法最后得到的最小化的目标函数值。这个平方和越小，说明模型与实际数据越相符。
- 记

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad (i = 1, 2, \ldots, n)$$

称为**残差** (residual)。记

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2,$$

- 分解的第二项：

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- 易见

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i =& \frac{1}{n}\sum_{i=1}^{n}(\hat{a}+\hat{b}x_i) \\ =& \hat{a}+\hat{b}\bar{x} = \bar{y}\end{aligned}$$

- 所以 U 是拟合值 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ 的离差平方和。
- U 受什么因素影响呢？

- $$\begin{aligned} U &= \sum_{i=1}^{n} (\hat{a} + \hat{b}x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2 \\ &= \hat{b}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \hat{b}^2 l_{xx} \end{aligned}$$

- l_{xx} 是自变量的离差平方和。所以 U 代表了自变量对因变量的变化的解释，在分解中 U 越大，残差平方和 Q 越小，模型对数据拟合越好。
- U 叫做回归平方和。

- 所以，平方和分解公式把因变量的变差（离差平方和）分解为两部分：

$$l_{yy} = Q + U$$

- U 来源于自变量 x 的分散程度，通过线性关系影响了因变量 Y 造成 Y 的变差，是模型可以解释的部分；
- Q 是模型拟合的误差的度量，是自变量和模型不能解释的部分。
- 分解中，U 越大，Q 越小，模型越准确描述自变量和因变量之间的线性相关关系。
- 反之，如果 Q 很大，则自变量和因变量之间没有线性相关关系。
- 取统计量

$$F = \frac{U}{Q/(n-2)} \tag{1.9}$$

则当 F 相当大时，表明 x 对 Y 的线性影响越强，两者有线性相关性。否则没有线性相关性。

- F 值多大才认为线性相关性成立？这是一个假设检验问题。
- 需要对模型进行进一步细化。
- 设数据满足如下结构

$$
\begin{aligned}
Y_1 &= a + bx_1 + \varepsilon_1 \\
Y_2 &= a + bx_2 + \varepsilon_2 \\
&\cdots\cdots\cdots \\
Y_n &= a + bx_n + \varepsilon_n
\end{aligned}
\tag{1.10}
$$

- 其中 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 是独立同分布随机变量列，共同分布为 $\mathsf{N}(0, \sigma^2)$(σ^2 未知)。
- 这样的模型是单总体模型、两总体模型的进一步推广。它等价于

$$Y_i \sim \mathsf{N}(a + bx_i, \sigma^2), \quad i = 1, 2, \ldots, n$$

且相互独立。这给出了 n 维随机向量 $(Y_1, Y_2, \ldots, Y_n)$ 的联合分布。

- 在承认了 (1.10) 模型基础上，x 与 Y 之间有无线性相关关系的问题变成假设

$$H_0 : b = 0$$

- 当 H_0 成立时，模型 (1.10) 退化为

$$Y_i = a + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

模型中不含自变量 x，所以这时 x 与 Y 没有线性相关关系。

- 当 H_0 不成立时，Y 与 x 有线性相关关系。

相关性检验

- 当 H_0 成立时，统计量 F 服从 $\mathsf{F}(1, n-2)$ 分布。
- 对检验水平 α，取 $\mathsf{F}(1, n-2)$ 分布的右侧 α 临界值 λ，H_0 的否定域为

$$W = \{F > \lambda\}$$

- 从样本中计算统计量 F 的值，当 $F > \lambda$ 时拒绝 H_0，认为 x 与 Y 之间存在线性相关性（有**显著的线性相关性**）；
- 当 $F \leq \lambda$ 时 H_0 相容，认为 x 与 Y 之间没有显著的线性相关性。

随机误差方差估计

- 可以证明

$$\frac{1}{\sigma^2}Q \sim \chi^2(n-2)$$

- 从而

$$\begin{aligned} E\left(\frac{1}{\sigma^2}Q\right) &= n-2 \\ E\left(\frac{1}{n-2}Q\right) &= \sigma^2 \end{aligned}$$

- 所以

$$\hat{s}^2 \triangleq \frac{1}{n-2}Q = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

是 σ^2 的无偏估计。

（样本）相关系数

- 设 U, V 是两个随机变量，其相关系数定义为

$$\rho = \frac{\mathrm{Cov}(U, V)}{\sqrt{D(U)D(V)}}$$

- 若 (U, V) 有样本 $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)$，则可计算样本相关系数

$$R = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^n (u_i - \bar{u})^2 \sum_{i=1}^n (v_i - \bar{v})^2}}$$

- 在线性回归中虽然 x 是非随机的变量，但也可以定义样本相关系数为

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- 当 $|R|$ 相当大时拒绝 $H_0 : b = 0$。

复相关系数平方

- 易见

$$
\begin{aligned}
R^2 = & \frac{l_{xy}^2}{l_{xx}l_{yy}} = \frac{l_{xy}^2}{l_{xx}^2} \cdot \frac{l_{xx}}{l_{yy}} \\
= & \frac{\hat{b}^2 l_{xx}}{l_{yy}} = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}}
\end{aligned}
$$

- 一般地，定义

$$
R^2 = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}},
$$

称 R^2 为复相关系数平方，这个定义在多元回归时照样适用。

- R^2 取值于 $[0,1]$，代表了回归平方和在总平方和中的比例，R^2 越接近于 1，回归模型对数据拟合得越好。
- 当 $R^2 = 1$ 时，$Q = 0$，所有的 n 个数据点

$$(x_i, y_i), \quad i = 1, 2, \ldots, n$$

都落在直线

$$\hat{y} = \hat{a} + \hat{b}x$$

上。
- $R^2 = 1$ 的情况一般只出现在确定性关系中。

F 统计量与 R^2

- 检验 $H_0: b=0$ 用的 F 统计量与 R^2 一一对应：

$$\begin{aligned} F =& \frac{U}{Q/(n-2)} = (n-2)\frac{U}{l_{yy}-U} \\ =&(n-2)\frac{R^2}{1-R^2} = \frac{n-2}{\frac{1}{R^2}-1} \end{aligned}$$

- 两者为一一对应的严格单调递增关系。

两个平方和的计算公式

- 按定义，

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 但是，在有了 l_{yy}, l_{xx}， l_{xy}, $\hat{b}$ 后可简单计算为

$$U = \hat{b}^2 l_{xx} = \hat{b} l_{xy} \tag{1.11}$$

$$Q = l_{yy} - U \tag{1.12}$$

例 1.4

- **例 1.4** 炼钢基本是氧化脱碳的过程，原来碳含量越高，需要的冶炼时间越长。
- 有某平炉 34 炉的熔毕碳 (x) 与精炼时间 (y) 的记录如下（部分）：

$$(180, 200), (104, 100), \ldots, (143, 160)$$

- 散点图见演示。

- 计算过程：见 R 程序演示。
- 主要结果

$$\hat{a} = -23.20, \qquad \hat{b} = 1.270$$
$$F = 145.0 \qquad R^2 = 0.8192$$

- F(1,32) 右侧 0.01 分位数为 $\lambda = 4.15$，$F > \lambda$，可以认为 x, Y 之间存在线性相关关系，或：直线回归是显著的。

回归模型的作用

- 揭示变量之间的数量关系；
- 预报；
- 控制。

预报

- 设

$$Y = a + bx + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2)$$

- 由数据 $(x_i, y_i), i = 1, 2, \ldots, n$ 得到参数最小二乘估计 $\hat{a}, \hat{b}$ 和误差方差估计 s^2。
- 对新的自变量值 x_0，设

$$Y_0 = a + bx_0 + \varepsilon_0$$

- 用

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

预报 Y_0 的值。

- 还需要衡量预报精度。
- 若 $\varepsilon_0, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{iid} \sim \mathrm{N}(0, \sigma^2)$，则

$$t \triangleq \frac{Y_0 - \hat{y}_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim \mathrm{t}(n-2)$$

- 为了求 Y_0 的**预报区间**, 设 λ 为 $\mathrm{t}(n-2)$ 分布的双侧 α 临界值，由

$$P\left(\left|\frac{Y_0 - \hat{y}_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}}\right| \le \lambda\right) = 1 - \alpha \tag{1.13}$$

- 得 Y_0 的置信度 $1-\alpha$ 的预报区间为

$$\hat{y}_0 \pm \lambda s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} \tag{1.14}$$

- 演示：熔毕碳与精炼时间的预报区间，连线为曲线，但只有单点意义。

- x_0 离 $\bar{x}$ 越远，预报区间长度越长。
- 注意：回归模型的应用范围不能超出原数据的范围。
- 作为 (1.14) 的近似，当 n 较大且 x_0 离 $\bar{x}$ 不远的时候，

$$\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\approx 1$$

所以预测区间近似为

$$[\hat{y}_0-\lambda s,\hat{y}_0+\lambda s]$$

- 当 n 较大时 λ 可以用标准正态分布的双侧 α 临界值，如 $\alpha=0.05$ 时用 $\lambda=1.96$。
- 误差标准差估计 s 越小，预报区间越短，预报越精确。

控制问题

- 控制问题是：要求控制 Y 在区间 $[A, B]$ 内，如何选取 x 的值？
- 办法是要求 (1.14) 得到的上下限都在 $[A, B]$ 内，反解符合要求 x_0 的区间。

回归诊断和残差分析

- 即使线性相关性检验否定了 $H_0: b = 0$，也并不说明模型就是合适的。
- 常见问题包括：
- 缺少重要自变量；
- 有非线性相关；
- 误差项方差非恒定；
- 误差项存在序列相关；
- 自变量严重共线（多元回归中）；
- 数据有异常值或强影响点。
- 可以用残差散点图等进行回归诊断。

残差分析

- 残差

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- 令

$$\begin{aligned} h_i =& \frac{1}{n} + \frac{(x_i - \bar{x})^2}{l_{xx}} \\ s =& \sqrt{\frac{Q}{n-2}} \\ r_i =& \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} \end{aligned}$$

- 则在 (1.10) 模型成立时 $r_1, r_2, \ldots, r_n$ 近似相互独立，且近似服从标准正态分布。

- 有

$$P(|r_i| > 2) \approx 0.05$$

- 当 n 比较大时，$r_i, i = 1, 2, \ldots, n$ 应该只有约 $[0.05n]$ 个绝对值大于 2。
- 这可以用来检验模型关于误差项的假设是否成立，以及发现异常值点。

本节目录

多元线性回归

- 考虑多个自变量与因变量的关系。
- 要解决的问题与一元回归相同。
- 解决方法类似。

模型

- 设因变量 Y 与自变量 $x_1, x_2, \ldots, x_k$ 有关系式

$$Y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

- 其中自变量 $x_1, x_2, \ldots, x_k$ 是非随机的变量，ε 是随机项。
- 有 n 组数据

$$\begin{aligned}
&(y_1; x_{11}, x_{12}, \ldots, x_{1k}) \\
&(y_2; x_{21}, x_{22}, \ldots, x_{2k}) \\
&\ldots\ldots\ldots \\
&(y_n; x_{n1}, x_{n2}, \ldots, x_{nk})
\end{aligned} \tag{2.1}$$

- 假定数据满足

$$\begin{cases} Y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_k x_{1k} + \varepsilon_1 \\ Y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_k x_{2k} + \varepsilon_2 \\ \quad \cdots\cdots\cdots \\ Y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_k x_{nk} + \varepsilon_n \end{cases} \tag{2.2}$$

  这里 Y_t 写成大写是为了强调在模型中它是随机变量。
- 其中 $b_0, b_1, \ldots, b_k$ 是待估参数，$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 相互独立且服从相同的 $\mathrm{N}(0, \sigma^2)$ 分布。

说明

- “多元”是指自变量有多个，但因变量还是只有一个。另外，自变量是非随机的普通变量，因变量是随机变量。
- (2.1) 中的各个 y_t 是数据值，(2.2) 中大写的 Y_t 看作随机变量。把 y_t 看作 Y_t 的观测值。
- (2.2) 表示 Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系。对于某些非线性关系，可以通过变换转化为线性。比如一元多项式回归

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k$$

只要记 $x_1 = x, x_2 = x^2, \ldots, x_k = x^k$ 就变成自变量为 $x_1, x_2, \ldots, x_k$ 的多元线性回归

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- 为估计未知参数，最小化误差平方和

$$
\begin{aligned}
&Q(b_0, b_1, \ldots, b_k) \\
=&\sum_{t=1}^{n}[y_t - (b_0 + b_1 x_{t1} + b_2 x_{t2} + \cdots + b_k x_{tk})]^2
\end{aligned}
$$

- 使 $Q(b_0, b_1, \ldots, b_k)$ 达到最小值的点 $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ 称为参数 $b_0, b_1, \ldots, b_k$ 的**最小二乘估计**。

- 记

$$
\begin{aligned}
\bar{y} =& \frac{1}{n}\sum_{t=1}^{n} y_t, \\
\bar{x}_i =& \frac{1}{n}\sum_{t=1}^{n} x_{ti},\ \ i = 1, 2, \ldots, k \\
l_{ij} =& l_{ji} = \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j),\ \ i, j = 1, 2, \ldots, k \\
l_{iy} =& \sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(y_t - \bar{y}),\ \ i = 1, 2, \ldots, k
\end{aligned}
$$

- 则 $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$ 为如下 n 阶线性方程组的解：

$$\begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1k} \\ l_{21} & l_{22} & \cdots & l_{2k} \\ \cdots & \cdots & \cdots & \\ l_{k1} & l_{k2} & \cdots & l_{kk} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{ky} \end{pmatrix}$$

- 而

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_k\bar{x}_k$$

- 这 $k+1$ 个方程组成的方程组叫做**正规方程**。
- 可以证明，最小二乘估计一定存在，而且 $b_0, b_1, \ldots, b_k$ 是最小二乘估计的充分必要条件为满足正规方程。

- 当如下矩阵

$$\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

为满秩矩阵（要求 $n > k+1$，满秩指列满秩）时正规方程的解唯一，所以最小二乘估计唯一。

平方和分解公式

- 平方和分解公式：

$$
\begin{aligned}
l_{yy} =& Q + U \\
l_{yy} =& \sum_{t=1}^{n} (y_t - \bar{y})^2 \\
Q =& \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \quad (\text{残差平方和}) \\
U =& \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 \quad (\text{回归平方和}) \\
=& \hat{b}_1 l_{1y} + \hat{b}_2 l_{2y} + \cdots + \hat{b}_k l_{ky} \\
\hat{y}_t =& \hat{b}_0 + \hat{b}_1 x_{t1} + \hat{b}_2 x_{t2} + \cdots + \hat{b}_k x_{tk}, \ t = 1, 2, \ldots, n
\end{aligned}
$$

σ^2 的无偏估计

- $Q/\sigma^2 \sim \chi^2(n-k-1)$，所以

$$
\begin{aligned}
E(Q/\sigma^2) =& n-k-1 \\
E\left(\frac{1}{n-k-1}Q\right) =& \sigma^2
\end{aligned}
$$

- 记

$$
\hat{\sigma}^2 = \frac{1}{n-k-1}Q
$$

  $\hat{\sigma}^2$ 为 σ^2 的无偏估计。

- 有时记为 s^2。

相关性检验

- 最小二乘估计总存在，所以不管 Y 和 $x_1, x_2, \ldots, x_k$ 之间有没有线性相关关系总能建立回归方程。
- 必须检验线性相关关系是否成立。
- 转化为：

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

  的检验。
- 当 H_0 成立时，模型中不出现自变量 $x_1, x_2, \ldots, x_k$，所以没有线性相关关系。
- 当 H_0 不成立时，Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系。

- 检验统计量为

$$F = \frac{U/k}{Q/(n-k-1)}$$

- 在 H_0 下 $F \sim \mathsf{F}(k, n-k-1)$。
- 给定检验水平 α 后查 $\mathsf{F}(k, n-k-1)$ 的临界值表得 λ。
- 计算 F 的值后，当且仅当 $F > \lambda$ 时拒绝 H_0，认为 Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系，也称**回归方程显著**。
- 若 F 的值为 v，可以计算检验的 p 值

$$p = P(F > v)$$

其中 F 为服从 $\mathsf{F}(k, n-k-1)$ 分布的随机变量，当且仅当 p 值小于 α 时拒绝 H_0。

因素主次的判别

- 多元回归时，即使能否定 $H_0: b_1 = b_2 = \cdots = b_k = 0$，仍然有可能一部分自变量与 Y 没有线性相关关系。
- 或者，虽然某自变量 x_i 与 Y 有线性相关关系，但是其它自变量能够代表它，所以 x_i 也不需要出现在模型中。
- 另外，即使部分自变量都是在模型中有意义的，也会有因素主次之分。

偏回归平方和

- 在平方和分解中，回归平方和 U 代表了所有 k 个自变量的作用。
- 为了研究某个自变量的贡献，不妨考虑 x_k 的作用。
- 从原来的数据中建立 Y 对 $x_1, x_2, \dots, x_{k-1}$ 的回归，得到一个回归平方和 $U_{(k)}$，一定有 $U_{(k)} \le U$。
- 称

$$u_k = U - U_{(k)}$$

为 $x_1, x_2, \dots, x_k$ 中 x_k 的**偏回归平方和**。
- 类似可以定义每个自变量的偏回归平方和 $u_i, i = 1, 2, \dots, k$。
- 注意偏回归平方和都是在一个变量集合的前提下讨论的。

偏回归平方和的计算

- u_i 的计算不需要真的重新拟合回归模型，而是有公式

$$u_i = \frac{\hat{b}_i^2}{c_{ii}}$$

其中 c_{ii} 为

$$L = (l_{ij})_{k\times k}$$

的逆矩阵的第 i 个主对角线元素。

- 为了检验 $H_0 : b_i = 0$，可以用

$$F_i = \frac{u_i}{s^2}$$

在 H_0 下 $F_i \sim \mathsf{F}(1, n-k-1)$。

单个自变量的显著性

- 设 F_i 的值为 v，则

$$p = P(F > v)$$

(其中 F 为 F$(1, n-k-1)$ 分布随机变量) 是检验的 p 值。
- 当 p 值小于 0.05 时称变量 x_i 是显著的。
- 当 p 值小于 0.01 时称变量 x_i 是高度显著的。
- 当 F_i 的值很小时，应该从回归方程中剔除自变量 x_i。
- 注意：当 x_i 不显著时，可能有两种原因：
  - x_i 对 Y 没有线性的影响；
  - x_i 对 Y 有线性的影响，但存在其它的自变量能够代替 x_i 对 Y 施加相同的影响。
- 即使回归方程显著，所有自变量显著，也不能断言模型就是符合实际的，还可能有各种模型设定错误或缺陷（类似一元回归时所述）。

多元回归的计算

- 各统计软件都可以很容易地计算多元回归。
- 比如，在 R 软件中输入了自变量 x1，x2和因变量 y后，只要用

```
lm1 <- lm(y ~ x1 + x2)
summary(lm1)
plot(lm1)
```

就可以得到回归结果并绘制回归诊断图形。

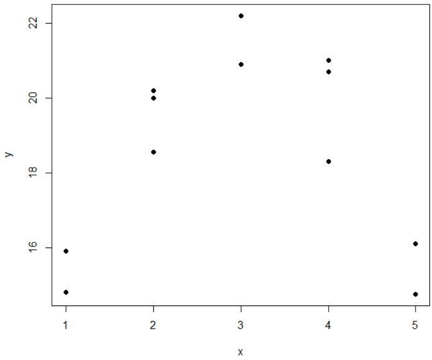
例 2.1（广告策略）

- **例 2.1（广告策略）** 研究广告费用 x 与纯利润 y 之间的关系，以确定最佳的广告策略。
- 数据：

x	1	1	2	2	2	3
y	14.80	15.90	20.20	20.00	18.55	22.20
x	3	4	4	4	5	5
y	20.90	21.00	18.30	20.70	16.10	14.75

- 试找出 y 与 x 的相关关系是并确定最优的广告策略。

- 画出散点图：



- 可以看出 y 与 x 不是线性关系。

- 最简单的非线性关系是一元二次多项式，设

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

其中 ε 是随机项。

- 若令 $x_1 = x, x_2 = x^2$，则方程化为

$$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

- 但是，在多项式回归时为了避免共线性问题，令

$$x_1 = x \qquad\qquad x_2 = (x-3)^2$$

- 用统计软件计算得

$$\begin{aligned}\hat{y} =& 21.26 + 0.07045x - 1.504(x-3)^2 \\ =& 7.627 - 9.094x - 1.504x^2 \\ s =& 0.9788 \\ F =& 35.08, \quad \text{p 值} < 0.0001\end{aligned}$$

- 为了求 $\hat{y}$ 的最大值（纯利润最大值），求导得

$$x = \frac{-9.093}{2 \times (-1.504)} = 3.02$$

时达到最大值。

例 2.2（生理节律模型）

- **例 2.2（生理节律模型）** 为了测定一个人在 24 小时内的生理节律（例如血压如何随时间变化），一些学者提出了如下模型：

$$f(t) = M + A\cos(\omega t + \phi)$$

- 其中 M 是基准值，A 是振幅，ϕ 是想为，ω 是角频率（周期 $T = 2\pi/\omega$）。
- 问：设有观测值

$$y_j = f(t_j) + \varepsilon_j,\ \ j = 1, 2, \ldots, n$$

  这里 t_j 是第 j 个观测时刻，$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 是相互独立的随机项，$\varepsilon_j \sim \mathrm{N}(0, \sigma^2)(\sigma^2$ 未知)。
- 如何估计 $M, A, \phi(0 \leq \phi < 2\pi)$?

- **解** 模型是非线性的，设法转换为线性。
- 易见

$$y_j = M + A\cos\phi \cdot \cos(\omega t_j) - A\sin\phi \cdot \sin(\omega t_j) + \varepsilon_j$$

- 记

$$\begin{aligned} x_j &= \cos(\omega t_j), & z_j &= \sin(\omega t_j) \\ \beta &= A\cos\phi, & \gamma &= -A\sin\phi \end{aligned}$$

- 则

$$y_j = M + \beta x_j + \gamma z_j + \varepsilon_j,\ (j = 1, 2, \ldots, n)$$

化为线性模型。

- 计算正规方程中各项：

$$l_{11} = \sum_{j=1}^{n} (x_j - \bar{x})^2, \qquad l_{22} = \sum_{j=1}^{n} (z_j - \bar{z})^2$$

$$l_{12} = \sum_{j=1}^{n} (x_j - \bar{x})(z_j - \bar{z})$$

$$l_{1y} = \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}), \qquad l_{2y} = \sum_{j=1}^{n} (z_j - \bar{z})(y_j - \bar{y})$$

- 解正规方程得

$$\hat{\beta} = \frac{l_{22}l_{1y} - l_{12}l_{2y}}{l_{11}l_{22} - l_{12}^2}, \qquad \hat{\gamma} = \frac{-l_{12}l_{1y} + l_{11}l_{2y}}{l_{11}l_{22} - l_{12}^2}$$

$$M = \bar{y} - \hat{\beta}\bar{x} - \hat{\gamma}\bar{z}$$

- 反推得到原始模型参数估计

$$\hat{A} = \sqrt{\hat{\beta}^2 + \hat{\gamma}^2} \qquad \hat{\phi} = \mathrm{Arg}(\hat{\beta} - i\hat{\gamma})$$

(这里 i 表示虚数单位，$\hat{\phi}$ 是平面直角坐标系中坐标为 $(\hat{\beta}, \hat{\gamma})$ 的点的辐角)

- 检验 y 与 t 是否有指定的非线性关系，可检验 $H_0 : A = 0$，等同于检验

$$H_0 : \beta = \gamma = 0$$

- 仍使用统计量

$$F = \frac{U/2}{Q/(n-3)}$$

取 F$(2, n-3)$ 的右侧 α 水平临界值 λ，当且仅当 $F > \lambda$ 时拒绝 H_0，认为回归方程显著。

- 实际中 t_j 一般是等间隔的，

$$t_j = \frac{j-1}{n},\ j = 1, 2, \ldots, n$$

且 $\omega = 2\pi$(周期为 1)，常用 $n = 24$ 或 $n = 12$。

- 这时公式可以化简：

$$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} \cos(\omega t_j) = 0$$

$$\sum_{j=1}^{n} z_j = \sum_{j=1}^{n} \sin(\omega t_j) = 0$$

$$\sum_{j=1}^{n} x_j z_j = \sum_{j=1}^{n} \cos(\omega t_j) \sin(\omega t_j) = 0$$

$$\sum_{j=1}^{n} x_j^2 = \sum_{k=0}^{n-1} \frac{1 + \cos(2k\theta)}{2} = \frac{n}{2} \qquad \left(\theta = \frac{2\pi}{n}\right)$$

$$\sum_{j=1}^{n} z_j^2 = \frac{n}{2}$$

- 于是得

$$\begin{aligned}
\hat{M} &= \bar{y} \\
\hat{\beta} &= \frac{1}{n}\sum_{j=1}^{n} x_j y_j \\
\hat{\gamma} &= \frac{1}{n}\sum_{j=1}^{n} z_j y_j \\
F &= \frac{n\hat{A}^2/2}{Q/(n-3)}
\end{aligned}$$

本节目录

二值因变量的问题

- 经典线性回归分析中因变量和自变量都是连续取值的。
- 实际工作中经常需要处理因变量为二分类值的情况。
- 比如，x 表示一个家庭年收入，$Y=1$ 表示该家庭在某段时间购买某种耐用消费品（如汽车），$Y=0$ 表示不购买。
- 研究 $P(Y=1)$ 与 x 的关系。
- 更一般地，若随机变量 Y 只取值 0 或 1，有若干个变量 $x_1, x_2, \ldots, x_k$ 影响 Y 的取值，关心 $p=P(Y=1)$ 如何依赖于 $x_1, x_2, \ldots, x_k$。

优比和 logit 函数

- 对 $0 < p < 1$，有 $\frac{p}{1-p} \in (0, \infty)$ 为 p 的严格单调递增函数，$\frac{p}{1-p}$ 叫做发生比或优比 (odds ratio)。
- 定义函数

$$\text{logit}(p) = \ln \frac{p}{1-p}, \quad 0 < p < 1$$

则 $\text{logit}(p) \in (-\infty, \infty)$ 是 p 的严格单调递增函数, 叫做 logit 函数。

逻辑斯蒂回归模型

- 设因变量和自变量间的关系为

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{3.1}$$

其中 $p = P(Y = 1)$，$\beta_0, \beta_1, \ldots, \beta_k$ 是常数，这时称二分类变量 Y 与自变量 $x_1, x_2, \ldots, x_k$ 的关系符合逻辑斯蒂回归模型。

- 易见 (3.1) 等同于

$$P(Y = 1|x_1, x_2, \ldots, x_k) = \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}$$

逻辑斯蒂回归参数估计

- 模型 (3.1) 中的常数 $\beta_0, \beta_1, \ldots, \beta_k$ 通常是未知的，需要从数据中估计。这个模型中没有方差项。
- 下面只考虑 $k = 1$，即只有一个自变量的情形，用 x 表示 x_1。
- (3.1) 化为

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x \tag{3.2}$$

- 令 $p(x) = P(Y = 1|x)$，则

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3.3}$$

- 参数估计可以用最大似然法和最小二乘法。

最大似然估计

- 设数据为 $(x_i, y_i), i = 1, 2, \ldots, n$。
- 则

$$P(Y = y_i|x_i) = [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}$$

- 观测值 $(x_i, y_i), i = 1, 2, \ldots, n$ 对应的似然函数为

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n}[p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}$$

- 对数似然函数为

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + e^{\beta_0+\beta_1 x_i})$$

- 令一阶偏导数都等于零的似然方程组

$$\sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right) x_i = 0$$

- 若 $(\hat{\beta}_0, \hat{\beta}_1)$ 是似然方程组的根且 $x_1, x_2, \ldots, x_n$ 不全相等，则似然方程组的根是惟一的，而且 $(\hat{\beta}_0, \hat{\beta}_1)$ 是 $L(\beta_0, \beta_1)$ 的最大值点从而是模型参数的最大似然估计。
- 可以证明 $\ln L(\beta_0, \beta_1)$ 是二元严格凹函数。
- 似然方程组有时无解，如所有 y_i 都等于 1 时。

加权最小二乘估计

- 数据有特殊要求。
- 设 $x = x_i$ 时共有 n_i 次观测，n_i 较大，其中事件 $\{Y = 1\}$ 发生了 γ_i 次 $(i = 1, 2, \ldots, m)$（$x_1, x_2, \ldots, x_m$）两两不同）。
- 用

$$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5} \tag{3.4}$$

作为 $\ln \frac{p(x_i)}{1-p(x_i)}$ 的估计值 $(i = 1, 2, \ldots, m)$。
- 令

$$\nu_i = \frac{(n_i + 1)(n_i + 2))}{n_i(\gamma_i + 1)(n_i - \gamma_i + 1)} \quad (i = 1, 2, \ldots, m) \tag{3.5}$$

$$\tilde{Q}(\beta_0, \beta_1) = \sum_{i=1}^{m} \frac{1}{\nu_i}(z_i - \beta_0 - \beta_1 x_i)^2$$

- 使 $\tilde{Q}(\beta_0,\beta_1)$ 达到最小值的 $\tilde{\beta}_0,\tilde{\beta}_1$ 称为 β_0,β_1 的加权最小二乘估计。
- 可以证明加权最小二乘估计存在且惟一。
- 令两个一阶偏导数都等于零的方程组

$$\beta_0\sum_{i=1}^{m}\frac{1}{\nu_i}+\beta_1\sum_{i=1}^{m}\frac{x_i}{\nu_i}=\sum_{i=1}^{m}\frac{z_i}{\nu_i}$$

$$\beta_0\sum_{i=1}^{m}\frac{x_i}{\nu_i}+\beta_1\sum_{i=1}^{m}\frac{x_i^2}{\nu_i}=\sum_{i=1}^{m}\frac{x_iz_i}{\nu_i}$$

- 记

$$l_1 = \sum_{i=1}^{m} \frac{1}{\nu_i}, \qquad l_2 = \sum_{i=1}^{m} \frac{x_i}{\nu_i}$$

$$l_3 = \sum_{i=1}^{m} \frac{x_i^2}{\nu_i} \qquad l_4 = \sum_{i=1}^{m} \frac{x_i z_i}{\nu_i}$$

$$l_5 = \sum_{i=1}^{m} \frac{z_i}{\nu_i}$$

- 则

$$\tilde{\beta}_0 = \frac{l_5 l_3 - l_2 l_4}{l_1 l_3 - l_2^2} \tag{3.6}$$

$$\tilde{\beta}_1 = \frac{l_1 l_4 - l_2 l_5}{l_1 l_3 - l_2^2} \tag{3.7}$$

加权最小二乘法的理由

- 应该用 $\frac{\gamma_i}{n_i-\gamma_i}$ 作为 $\frac{p(x_i)}{1-p(x_i)}$ 的估计，为避免分子和分母出现零，做连续型修正变成 $\frac{\gamma_i+0.5}{n_i-\gamma_i+0.5}$。
- 可以证明，

$$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$$

近似服从正态分布

$$\mathsf{N}\left(\ln \frac{p(x_i)}{1-p(x_i)}, \frac{1}{n_i p(x_i)[1-p(x_i)]}\right)$$

- 利用 (3.2)，有

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i = 1, 2, \ldots, m$$

其中 ε 近似服从 $\mathsf{N}(0, \nu_i)$。

- 令

$$\tilde{\varepsilon}_i = \frac{1}{\sqrt{\nu_i}}\varepsilon_i$$

- 则

$$\frac{1}{\sqrt{\nu_i}}z_i = \frac{1}{\sqrt{\nu_i}}(\beta_0 + \beta_1 x_i) + \tilde{\varepsilon}_i, i = 1, 2, \ldots, n$$

其中 $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_n$ 的方差相等，仿照最小二乘法思想令

$$\sum_{i=1}^{m}\left[\frac{1}{\sqrt{\nu_i}}z_i - \frac{1}{\sqrt{\nu_i}}(\beta_0 + \beta_1 x_i)\right]^2$$

达到最小，即 $\tilde{Q}(\beta_0, \beta_1)$ 达到最小。

例 3.1（社会调查）

- **例 3.1（社会调查）** 一个人在家是否害怕生人来？
- 研究人的文化程度对此问题的影响。
- 因变量 $Y = 1$ 表示害怕，0 表示不害怕。
- 自变量 x 是文化程度，$x_1 = 0$ 表示文盲，$x_2 = 1$ 表示小学，$x_3 = 2$ 表示中学，$x_4 = 3$ 表示大专以上。
- 根据一项社会调查有如下数据：

自变量 (x)	不害怕人数	害怕人数
0	11	7
1	45	32
2	664	422
3	168	72

- 用逻辑斯蒂回归模型分析。用 $p(x)$ 表示文化程度为 x 的人害怕生人的概率。设模型

$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

- 用加权最小二乘法估计 β_0, β_1。
- 计算得 $z_1 = -0.3847$, $z_2 = -0.3269$, $z_3 = -0.4515$, $z_4 = -0.8425$, $\nu_1 = 0.2199$, $\nu_2 = 0.0527$, $\nu_3 = 0.00387$, $\nu_4 = 0.0197$。
- 用 (3.6) 和 (3.7) 得 $\tilde{\beta}_0 = 0.013$, $\beta_1 = -0.25$。
- 回归方程为

$$\ln \frac{p(x)}{1-p(x)} \approx 0.013 - 0.25x$$

- 可见文化程度越高，害怕生人的概率越低。

在统计软件中计算逻辑斯蒂回归

- 一般用统计软件计算逻辑斯蒂回归。
- 如上例的 R 程序：

```
x <- 0:3
n1 <- c(11, 45, 664, 168)
n0 <- c(7, 32, 422, 72)
y <- cbind(n1, n0)
glm1 <- glm(y ~ x, family=binomial)
print(summary(glm1))
```