《生物实验设计》
第十一章　多元线性回归与多元相关分析

王超

广东药科大学

Email: wangchao@gdpu.edu.cn

2022-10-25

第十一章　多元线性回归与多元相关分析

一元与多元回归

- 因变量 y 在一个自变量 x 上的回归或相关，统称为一元回归或一元相关
- 在实际问题中，影响 y 的因素常常不只是一个，而是两个或两个以上
- 为了清楚了解因变量 y 和多个自变量 x 之间的关系，必须在一元回归与相关分析的基础上，进行多元回归与多元相关分析（复回归与复相关）

基本方法：

以多元线性回归模型为基础，根据最小二乘法建立正规方程，求解得出多元线性回归方程，并对回归方程和偏回归系数进行检验，作出回归方程的区间估计。

- 多元线性回归
  - 具一个因变量 y 与两个或两个以上自变量 x，且各自变量均为一次项的回归。
- 设自变量 $x_1, x_2, x_3, \ldots, x_m$ 与因变量 y 皆呈线性关系，则一个 m 元线性回归的数学模型可以表示为：

$$y_i = \mu_y + \beta_{y1}(x_1 - \mu_{x_1}) + \beta_{y2}(x_2 - \mu_{x_2}) + \cdots + \beta_{ym}(x_m - \mu_{x_m}) + \epsilon_i$$

  - i 代表第 i 个样本
  - $\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_m}$ 依次为 $y, x_1, x_2, \ldots, x_m$ 的总体平均数，其样本估计值为 $\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$
  - β_{y1} 为 $x_2, x_3, \ldots, x_m$ 固定不变时，x_1 每变动一个单位 y 平均变动的相应单位数，称为 x_1 对 y 的偏回归系数，记作 β_1，样本估计值记作 b_1，余下类推
  - ϵ_i 为随机误差，服从 $N(0, \sigma^2)$ 的正态分布，σ^2 为离回归方差

- 若令 $\alpha = \mu_y - \beta_1\mu_{x_1} - \beta_2\mu_{x_2} - \cdots - \beta_m\mu_{x_m}$，则多元线性回归的数学模型为：

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

- 于是，样本多元线性回归方程为：

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$

- 其中，a 为 α 的样本估计值，b 为 β 的样本估计值

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m$$

- 多元线性回归方程可根据最小二乘法建立
- 也就是求以下方程最小值

$$min(Q) = \sum (y - \hat{y})^2$$
$$= \sum [y - \bar{y} - b_1(x_1 - \bar{x}_1) - b_2(x_2 - \bar{x}_2) - \cdots - b_m(x_m - \bar{x}_m)]^2$$

- 简单形式：

$$min(Q) = \sum (Y - b_1 X_1 - b_2 X_2 - \cdots - b_m X_m)^2$$

其中

$$\begin{cases} Y = y - \bar{y} \\ X_1 = x_1 - \bar{x}_1 \\ \cdots \\ X_m = x_m - \bar{x}_m \end{cases}$$

- 求自变量和因变量关系的最小二乘法

$$min(Q) = \sum_{1}^{n}(y - \hat{y})^2 = \sum_{1}^{n}(y - a - bx)^2$$

- 根据极值定理，对 a 和 b 分别求导：

$$\frac{\partial Q}{\partial a} = -2\sum(y - a - bx) = 0, \frac{\partial Q}{\partial b} = -2\sum(y - a - bx)x = 0$$

- 整理得到：

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \end{cases}$$

- 使 $b_1, b_2, \ldots, b_m$ 的偏微分方程皆等于 0，就有

$$\begin{cases} \frac{\partial Q}{\partial b_1} = -2\sum(Y - b_1X1 - b_2X_2 - \cdots - b_mX_m)X_1 = 0 \\ \frac{\partial Q}{\partial b_2} = -2\sum(Y - b_1X1 - b_2X_2 - \cdots - b_mX_m)X_2 = 0 \\ \vdots \\ \frac{\partial Q}{\partial b_m} = -2\sum(Y - b_1X1 - b_2X_2 - \cdots - b_mX_m)X_m = 0 \end{cases}$$

- 整理后得到

$$\begin{cases} b_1\sum X_1^2 + b_2\sum X_1X_2 + \cdots + b_m\sum X_1X_m = \sum X_1Y \\ b_1\sum X_1X_2 + b_2\sum X_2^2 + \cdots + b_m\sum X_2X_m = \sum X_2Y \\ \vdots \\ b_1\sum X_1X_m + b_2\sum X_2X_m + \cdots + b_m\sum X_m^2 = \sum X_mY \end{cases}$$

- 因为 $X = x - \bar{x}$，所以可以表示为
  $\sum X_1^2 = SS_1, \sum X_m^2 = SS_m, \sum X_1 X_m = SP_{1m}$
- 得到如下方程组

$$\begin{cases} b_1 SS_1 + b_2 SP_{12} + \cdots + b_m SP_{1m} = SP_{1y} \\ b_1 SP_{12} + b_2 SS_2 + \cdots + b_m SP_{2m} = SP_{2y} \\ \vdots \\ b_1 SP_{1m} + b_2 SP_{2m} + \cdots + b_m SS_m = SP_{my} \end{cases}$$

- 该线性方程组可以用消元法去解，也可以用矩阵方法求解
- 矩阵方式解方程更加容易

- 以上线性方程组的矩阵形式表示为

$$\begin{bmatrix} SS_1 & SP_{12} & \dots & SP_{1m} \\ SP_{12} & SS_2 & \dots & SP_{2m} \\ \vdots & & & \\ SP_{1m} & SP_{2m} & \dots & SS_m \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} SP_{1y} \\ SP_{2y} \\ \vdots \\ SP_{my} \end{bmatrix}$$

- 系数矩阵用 A 表示，可以通过计算得出
- 未知元矩阵用 b 表示，是多元回归方程的偏回归系统组成
- 常数矩阵用 K 表示

- 以上矩阵可简写为

$$Ab = K$$

- 因为 $AA^{-1} = I$，是单位矩阵，所以

$$A^{-1} \times Ab = b = A^{-1} \times K$$

- 那么

$$b = A^{-1}K$$

- 求解逆矩阵的方法：
  - 初等变换法
  - 伴随矩阵法
  - 表解法

- 初等变换法
  - 写出增广矩阵 $A|E$，即矩阵 A 右侧放置一个同阶的单位矩阵，得到新的矩阵
    - 交换矩阵的某两行（列）
    - 以数 $k \neq 0$ 乘以矩阵的某一行（列）
    - 把矩阵的某一行（列）的 k 倍加到另一行（列）

$$A = \begin{bmatrix} 3 & 0 & 2 & 1 & 0 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

- **初等变换法**

$$A=\begin{bmatrix}3&0&2&1&0&0\\2&0&-2&0&1&0\\0&1&1&0&0&1\end{bmatrix}\sim\begin{bmatrix}1&0&4&1&-1&0\\0&1&1&0&0&1\\2&0&-2&0&1&0\end{bmatrix}$$

$$\sim\begin{bmatrix}1&0&4&1&-1&0\\0&1&1&0&0&1\\0&0&-10&-2&3&0\end{bmatrix}\sim\begin{bmatrix}10&0&40&10&-10&0\\0&10&10&0&0&10\\0&0&-10&-2&3&0\end{bmatrix}$$

$$\sim\begin{bmatrix}10&0&40&10&-10&0\\0&10&10&0&0&10\\0&0&-10&-2&3&0\end{bmatrix}\sim\begin{bmatrix}10&0&0&2&2&0\\0&10&0&-2&3&10\\0&0&10&2&-3&0\end{bmatrix}$$

- 伴随矩阵法
  - 余子式矩阵：为矩阵的每个元素
    - 不使用在本行与本列的元素
    - 计算剩下来的值的行列式
    - 把行列式的结果放进一个新的矩阵——余子式矩阵

$$A = \begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$$

- 伴随矩阵法

$$\begin{bmatrix} \mathbf{3} & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \sim \begin{vmatrix} 0 & -2 \\ 1 & 1 \end{vmatrix} \sim 0 \times 1 - (-2) \times 1 = 2$$

$$\begin{bmatrix} 3 & \mathbf{0} & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \sim \begin{vmatrix} 2 & -2 \\ 0 & 1 \end{vmatrix} \sim 2 \times 1 - (-2) \times 0 = 2$$

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & \mathbf{1} & 1 \end{bmatrix} \sim \begin{vmatrix} 3 & 2 \\ 2 & -2 \end{vmatrix} \sim 3 \times (-2) - 2 \times 2 = -10$$

- 伴随矩阵法行列式的计算

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

余子式矩阵

$$\begin{bmatrix} 2 & 2 & 2 \\ -2 & 3 & 3 \\ 0 & -10 & 0 \end{bmatrix}$$

代数余子式矩阵：a_{ij} 余子式的值乘以 $(-1)^{i+j}$ 就是 a_{ij} 的代数余子式的值

$$\begin{bmatrix} 2 & 2 & 2 \\ -2 & 3 & 3 \\ 0 & -10 & 0 \end{bmatrix} \sim \begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix} \xrightarrow{(-1)^{i+j}} \begin{bmatrix} 2 & -2 & 2 \\ 2 & 3 & -3 \\ 0 & 10 & 0 \end{bmatrix}$$

$$\xrightarrow{transpose} \begin{bmatrix} 2 & 2 & 0 \\ -2 & 3 & 10 \\ 2 & -3 & 0 \end{bmatrix}$$

- 伴随矩阵法
  - 伴随矩阵乘以 1/原始矩阵行列式得到逆矩阵

$$\begin{vmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{vmatrix} = 10$$

$$A^{-1} = \frac{1}{10}\begin{bmatrix} 2 & 2 & 0 \\ -2 & 3 & 10 \\ 2 & -3 & 0 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.2 & 0 \\ -0.2 & 0.3 & 1 \\ 0.2 & -0.3 & 0 \end{bmatrix}$$

R DEMO

```
matrix_A <- matrix(c(1.65640, -1.49800, -3.63000,
                     -1.49800, 36.89875, 83.80000,
                     -3.63000, 83.80000, 320.00000),
                   nrow = 3, ncol = 3)
inverse_matrix <- solve(matrix_A)
inverse_matrix
```

```
##              [,1]         [,2]          [,3]
## [1,] 0.626886834  0.02294788  0.001101772
## [2,] 0.022947876  0.06771345 -0.017472144
## [3,] 0.001101772 -0.01747214  0.007713016
```

- 求算偏回归系数建立多元线性回归方程
  - 解出系数矩阵 A 的逆矩阵
  - 由 A^{-1} 求出 b_i 和 a

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} SS_1 & SP_{12} & \dots & SP_{1m} \\ SP_{12} & SS_2 & \dots & SP_{2m} \\ \vdots & & & \\ SP_{1m} & SP_{2m} & \dots & SS_m \end{bmatrix}^{-1} \times \begin{bmatrix} SP_{1y} \\ SP_{2y} \\ \vdots \\ SP_{my} \end{bmatrix}$$

（一）多元线性回归方程的估计标准误

- 建立多元线性回归方程时，由于实际观测值 y 与多元回归方程的点估计值 $\hat{y}$ 存在差异，其差值的平方和称为多元回归方程的离回归平方和 Q_y
- Q_y 的自由度 $\mathrm{d}f = n - (m+1) = n - m - 1$
- 多元回归方程的估计标准误

$$s_y = \sqrt{\frac{Q_y}{n-m-1}}$$

（一）多元线性回归方程的估计标准误

- 多元回归中的 y 的离均差平方和可以分解为离回归平方和与回归平方和：

$$\begin{cases} SS_y = Q_y + U_y \\ U_y = b_1 SP_{1y} + b_2 SP_{2y} + \cdots + b_m SP_{my} \end{cases}$$

- Q 代表误差因素引起的平方和，也叫残差平方和
- U 代表由 x 变异引起 y 变异的平方和，也叫回归平方和

（二）多元线性回归方程的假设检验

假设：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$$
$$H_A : \beta_1, \beta_2, \ldots, \beta_m \neq 0$$

F 检验 $(\mathrm{d}f_1 = m, \mathrm{d}f_2 = n - m - 1)$

$$F = \frac{\frac{U_y}{m}}{\frac{Q_y}{n-m-1}}$$

（二）多元线性回归方程的假设检验

- 应注意两个问题：
  - 多元线性回归关系显著不排斥有更合理的多元非线性回归方程存在
  - 多元线性回归显著不排斥其中存在着与因变量 y 无线性关系的自变量
- 因此：
  - 有必要对各个偏回归系数逐个进行假设检验
    $H_A: \beta_1, \beta_2, \ldots, \beta_m \neq 0$
  - 只有当多元回归方程的偏回归系数均达到显著，多元回归的 F 值才有确定的意义

（三）偏回归系数的假设检验

- 偏回归系数的假设检验是分别计算各偏回归系数 b_i 来自 $\beta_i = 0$ 的总体的概率
- 假设

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

- 偏回归系数的假设检验方法有 t 检验和 F 检验两种

（三）偏回归系数的假设检验

- t 检验
  - 偏回归系数 b_i 的标准误为

$$s_{b_i} = s_y\sqrt{c_{ii}}$$

- 由于 $\frac{b_i-\beta_i}{s_{b_i}}$ 符合 $\mathrm{d}f = n-m-1$ 的 t 分布，所以在 $\beta_i = 0$ 的假设下，由

$$t = \frac{b_i}{s_{b_i}}$$

可知 b_i 抽自 β_i 的总体的概率

（三）偏回归系数的假设检验

- F 检验
  - 对于 U_y 来说，其中的每一个组成 U_i 称为 y 在 x_i 上的偏回归平方和

$$U_i = \frac{b_i^2}{c_i i}$$

  - U_i 是因为添入 x_i 后增加的回归部分平方和，具有 1 个自由度

$$F = \frac{U_i}{\frac{Q_y}{n-m-1}}$$

可知 b_i 抽自 $\beta_i = 0$ 的总体的概率

（三）偏回归系数的假设检验

- 值得注意的两个问题：
  1. 由于对各偏回归系数的 F 检验中分子自由度均为 1，故其平方根值等于相应的 t 值的绝对值

$$\sqrt{F}=\sqrt{\frac{U}{\frac{Q_y}{n-m-1}}}=\sqrt{\frac{b_i^2/c_{ii}}{s_y^2}}=\sqrt{\frac{b_i^2}{s_{b_i}^2}}=|t|$$

  2. 自变量间的相关会使 x_i 对 y 的作用发生改变
     - 在 m 元线性回归中，如果各自变量间没有相关，即 $r_{ij}=0$，则 $U_y=\sum_{i=1}^{m}U_i$
     - 如果各自变量间存在不同程度的相关，即 $r_{ij}\neq 0$，则 $U_y\neq\sum_{i=1}^{m}U_i$

- 多元相关是 m 个自变量和因变量的总相关
- 多元相关系数表示多个自变量与因变量总的密切程度的量值，以 R_y 表示
- 由于 m 个自变量对 y 的回归平方和为 U_y，U_y 占 y 平方和 SS_y 的比例越大，表明 y 和 m 个自变量的总相关越密切，因此定义 R_y 为

$$R_y = \sqrt{\frac{U_y}{SS_y}}$$

- 以上公式表示多元相关系数为多元回归平方和与总变异平方和之比的平方根
- 取值区间为 $[0, 1]$

- 多元相关系数的假设检验是用 F 检验，不能用 t 检验
- 检验的假设

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0$$

- F 值为

$$F = \frac{\mathrm{d}f_2 R^2}{\mathrm{d}f_1(1-R^2)}, \mathrm{d}f_1 = m, \mathrm{d}f_2 = n - m - 1$$

- 多元相关系数的显著性与多元回归方程的显著性一致，也就是说 R_y 显著，多元回归方程必显著
- 多元相关系数的平方称为决定系数，是多元回归平方占 y 的总变异平方和的比率

- 生物学研究中，任何两个变量间的相关经常受到其他变量的影响
- 为消除这些影响，使两个变量间的相关关系得到真实的反映，必须排除其他变量影响的情况下进行两个变量间的分析
- 排除其他变量影响下的两个变量间的相关分析称为偏相关分析
- 其他变量保持一定，表示指定的两个变量之间相关密切程度的量值称为偏相关系数

（一）偏相关系数的一般解法

- 计算由简单相关系数构成的相关矩阵 R
- 求其逆矩阵 R^{-1}
- 计算偏相关系数 r_{ij}

（一）偏相关系数的一般解法

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & & & \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \xrightarrow{transpose}$$

$$R^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & & & \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix}$$

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}$$

（二）偏相关系数的假设检验

- 偏相关系数的假设检验可以用 t 检验
- 检验的假设

$$H_0 : \rho_{ij} = 0$$
$$H_A : \rho_{ij} \neq 0$$

- t 值为

$$t = \frac{r_{ij}}{\sqrt{1 - r_{ij}^2}} \sqrt{n - m - 1}$$

（三）偏相关与简单相关的区别

- 简单相关系数没有排除其他变量的影响，其中混有其他变量的效应
  - 当其他变量与简单相关系数正相关时，混有正效应，简单相关系数会高于偏相关系数
  - 当其他变量与简单相关系数负相关时，混有负效应，简单相关系数会低于偏相关系数
- 偏相关系数与简单相关系数相比，能排除假象，反映变量间真实的相关密切程度
- 对于多变量资料，必须采用多元相关分析