第一章 回归分析

在现实世界中存在大量的变量，它们有相互依存、相互制约的关系，一般分为两类：确定性关系与非确定性关系。如圆的面积与半径之间的关系；人的身高与体重之间的关系；产品价格与需求量之间的关系。

确定性关系：数学中用明确的函数来表示的关系。

相关关系：变量之间相互制约，却又没有达到可以相互确认的程度。

回归分析的研究对象：具有相关关系的随机变量。研究用一组变量（常称为自变量或预测变量）去预测另一组变量（常称为因变量或响应变量）。

研究方法包括最小二乘准则下的经典多元线性回归分析（MLR），提取自变量组主成分的主成分回归分析（PCR）等方法外，还有近年发展起来的偏最小二乘（PLS）回归方法。

(1) 线性回归：

一元线性回归、多元线性回归、多个因变量与多个自变量的回归。

(2) 回归诊断：

从数据推断回归模型基本假设的合理性、当基本假设不成立时如何对数据进行修正、判定回归方程拟合的效果、选择回归函数的形式。

(3) 回归变量的选择：

自变量选择的准则，逐步回归分析方法。

(4) 参数估计方法的改进：

岭回归、主成分回归、偏最小二乘法。

(5) 非线性回归：

一元非线性回归、分段回归、多元非线性回归。

(6) 含有定性变量的回归：

自变量含定性变量的情况，因变量是定性变量的情况。

1.1 一元线性回归

仅有一个自变量的回归称为一元回归, 有多个自变量的回归称为多元回归。一元线性回归模型为：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

其中，$\beta_0 + \beta_1 x$ 为线性变化部分，ε 为随机误差。通常假定 ε 满足：

$$E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2.$$

实际上，可以得到 n 组样本的观测值：

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

于是，

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

通常假定 ε_i 独立同分布，且满足：

$$E(\varepsilon_i) = 0,\ D(\varepsilon_i) = \sigma^2$$

1.1.1 回归分析的任务

1. 通过 n 组样本的观测值：$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 对 β_0 和 β_1 进行估计。
2. 对所假定的回归模型进行检验。

用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别表示两个参数的估计值，则有如下的一元线性经验回归方程：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

1.1.2 参数 β_0 和 β_1 的估计

利用最小二乘法。定义离差平方和：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

最优的参数 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 满足：

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\min_{\beta_0,\beta_1}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

由极值的条件可知，最优解满足：

$$\begin{cases}\left.\frac{\partial Q}{\partial\beta_0}\right|_{\beta_0=\hat{\beta}_0}=0\\ \left.\frac{\partial Q}{\partial\beta_1}\right|_{\beta_1=\hat{\beta}_1}=0\end{cases}$$

进而可得：

$$\begin{cases}\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}\\ \hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\end{cases}$$

其中，$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$。求最大似然估计的推导过程如下。

设总体 X 为连续分布，其分布密度族为：$\{f(x,\theta),\theta\in\Theta\}$，则有似然函数：

$$\mathcal{L}(\theta;x_1,x_2,\ldots,x_n)=\Pi_{i=1}^{n}f(x_i;\theta)$$

最优的参数 $\hat{\theta}$ 满足：

$$\mathcal{L}(\hat{\theta};x_1,x_2,\ldots,x_n)=\max_{\theta}\mathcal{L}(\theta;x_1,x_2,\ldots,x_n)$$

对于一元线性回归模型，假设 $\varepsilon_i\sim\mathcal{N}(0,\sigma^2)$，则有：

$$y_i\sim\mathcal{N}(\beta_0+\beta_1x_i,\sigma^2)$$

对应的密度函数为：

$$f_i(y_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[y_i-(\beta_0+\beta_1x_i)\right]^2\right\}$$

进而可得似然函数：

$$\mathcal{L}(\beta_0,\beta_1,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_i)\right]^2\right\}$$

等价转化为对数似然函数：

$$\ln(\mathcal{L})=-\frac{n}{2}(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_i)\right]^2$$

因此，参数 β_0 和 β_1 的最优估计值：

$$\begin{cases}\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}\\ \hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\end{cases}$$

进而可得方差 σ^2 的最优估计值：

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)\right]^2$$

该估计为有偏估计，实际中用如下的无偏估计：

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\left[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)\right]^2$$

1.1.3 一元线性回归中的假设检验

检验自变量和因变量之间是否存在如回归模型所假定的那种形式的相关关系，是回归分析的另一项重要工作。

实际上，在一元线性回归中，利用最小二乘法估计回归函数 $\beta_0+\beta_1 x$，也只是选出“相对的最好”，即对已有的 n 个观测点，找出总残差最小的那条拟合直线。如果变量之间确实存在线性相关关系，直线回归方程就能较好的反映变量的这种关系，否则没什么意义。

如何检验 X,Y 之间是否存在线性相关关系？

有一种直观的简易方法, 就是把 n 个观测点 $(x_i,y_i), i=1,2,\ldots,n$ 描在平面上，即绘制出散点图，从这些点的分布情况大体上判断是否可以用直线来表示 X,Y 之间的相关关系。注意到，该方法最大的缺点是缺乏定量的分析。下面介绍两种常用的关于线性相关关系的假设检验方法。

1. 基于离差平方和分解的检验

记总离差平方和为：

$$Q_T=\sum_{i=1}^{n}(y_i-\overline{y})^2.$$

它反映了因变量 Y 的观测值 $y_1,y_2,\ldots,y_n$ 总的离散程度。对其进行分解：

$$Q_T=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y})\right]^2.$$

即：

$$Q_T=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+(\hat{y}_i-\overline{y})^2.$$

等式右边的两项分别为：残差平方和 Q_E、回归平方和 Q_R。因此，

$$Q_T=Q_E+Q_R.$$

表明：观测值 $y_1,y_2,\ldots,y_n$ 的离散程度可分别由两个部分来描述。

定义：

$$F=\frac{Q_R}{Q_E/(n-2)}$$

则 F 的值越大，Y 与 X 之间的线性关系越显著。

2. 样本相关系数检验法

在概率论中，相关系数 ρ_{XY} 的大小反映 X 和 Y 的线性相关程度。类似地，在一元线性回归分析中，定义如下的相关系数：

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

可以看出，$|R|$ 的值越接近于 1, X 与 Y 之间的线性相关关系越显著。

1.2 多元线性回归

多元线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

其中，β_0 为回归常数，$\beta_1, \beta_2, \ldots, \beta_p$ 为回归系数，ε 为随机误差。通常假定 ε 满足：

$$E(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2$$

实际上，可以得到 n 组样本的观测值：

$$(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i),\ i = 1, 2, \ldots, n$$

于是，

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

简记为：

$$y = X\beta + \varepsilon$$

为了便于参数估计，做如下的假设。

(1) $x_1, x_2, \ldots, x_p$ 不是随机变量，是确定性变量。

(2) $r(X) = p + 1 < n$

(3) $E(\varepsilon_i) = 0,\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$

(4) $\varepsilon_i \sim N(0, \sigma^2)$，且 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 相互独立。

基于以上假设可知：

$$y \sim N(X\beta, \sigma^2 I_n)$$

1.2.1 参数 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ 的估计

最小二乘法来估计。

定义离差平方和：

$$Q(\beta_0, \beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

最优的参数满足：

$$Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p) = \min_{\beta_0, \beta_1, \beta_2, \ldots, \beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

由极值的条件可知，最优解满足：

$$\begin{cases} \left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = 0 \\ \left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = 0 \\ \left.\frac{\partial Q}{\partial \beta_2}\right|_{\beta_2=\hat{\beta}_2} = 0 \\ \cdots \\ \left.\frac{\partial Q}{\partial \beta_p}\right|_{\beta_p=\hat{\beta}_p} = 0 \end{cases}$$

进而可得：

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{i1} = 0 \\ \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{i2} = 0 \\ \cdots \\ \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) x_{ip} = 0 \end{cases}$$

用矩阵表示：

$$X^T (y - X\hat{\beta}) = 0$$

其中，

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

当 $X^T X$ 可逆时，可得参数 β 的最优估计：

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

最大似然估计如下。

由假设可知：$y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$。因此，似然函数为：

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}$$

等价转化为对数似然函数：

$$\ln(\mathcal{L}) = -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$

注意到，该值的最大等价于：

$$\min_{\beta_0,\beta_1,\beta_2,\ldots,\beta_p} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

这与最小二乘法的结果是一致的，即参数 β 的最优估计为：

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

1.3 偏最小二乘回归分析

偏最小二乘回归提供一种多对多线性回归建模的方法，特别当两组变量的个数很多，且都存在多重相关性，而样本量又较少时，用偏最小二乘回归建立的模型具有传统的经典回归分析等方法所没有的优点。主要有以下的特点。

(1) 能够在自变量存在严重多重相关性的条件下进行回归建模。

(2) 允许在样本点个数少于变量个数的条件下进行回归建模。

(3) 偏最小二乘回归在最终模型中将包含原有的所有自变量。

(4) 偏最小二乘回归模型更易于辨识系统信息与噪声。

(5) 在偏最小二乘回归模型中，每一个自变量的回归系数将更容易解释。

问题背景

考虑回归方程中存在如下两个问题时的情形。

(1) 自变量 $x_1, x_2, \ldots, x_p$ 的数目较多。

(2) 样本观测值的个数相对较少，即 $n < p$。

解决方案

由于 $n < p$，故 X^TX 为奇异矩阵，故最小二乘法无法解决该问题。主成分回归（PCR）通过求解 X^TX 的非零特征值，并将其按从大到小的顺序排列后得到：

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$$

记 α_i 的特征向量为：$\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ip}$。作如下的线性变换：

$$z_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \cdots + \alpha_{ip}x_p,\ i = 1, 2, \ldots, r$$

接下来，y 对 $z_1, z_2, \ldots, z_r$ 做回归。

1.3.1 方案的改进

PCR 方法的缺陷在于 z_i 的选取与 y 无关，只在 $x_1, x_2, \ldots, x_p$ 中选了部分 $z_1, z_2, \ldots, z_r$。偏最小二乘（PLS）对该方法做了改进：在找 $x_1, x_2, \ldots, x_p$ 的线性函数时，选取与 y 相关性较强的线性函数，即只考虑偏向与 y 有关的一部分。该方法的步骤如下。

首先，将数据 y 和 X 的中心化，得到均值为 0 的数据。然后，将 y 对 x_i 单独做回归：

$$\hat{y}(x_i) = \hat{\beta}_1 x_i = \frac{x_i^T y}{x_i^T x_i} x_i$$

加权求和：

$$\sum_{i=1}^{p} \omega_i \frac{x_i^T y}{x_i^T x_i} x_i$$

如取 $\omega_i = x_i^T x_i$，则有：

$$t_1 = \sum_{i=1}^{p} (x_i^T y) x_i$$

将 t_1 作为自变量，对 y 求回归，则有：

$$\hat{y}(t_1) = \frac{t_1^T y}{t_1^T t_1} t_1$$

并用该式预测 y，得到预测向量 $\hat{y}(t_1)$。将 x_i 作为自变量，对 t_1 求回归，得回归方程和预测值：

$$\hat{x}_i(t_1) = \frac{t_1^T x_i}{t_1^T t_1} t_1, \ i = 1, 2, \ldots, p$$

重复以上过程，求得：$t_2, t_3, \ldots, t_r$。利用 y 对 $t_2, t_3, \ldots, t_r$ 采用最小二乘做回归，经过变量转换，可得 y 对 $x_1, x_2, \ldots, x_p$ 的回归方程。

1.3.2 多个因变量和自变量的偏最小二乘

求解步骤：首先在自变量集中提出第一成分 u_1；同时在因变量集中也提取第一成分 v_1，并要求 u_1 与 v_1 相关程度达到最大。

然后建立因变量 $y_1, y_2, \ldots, y_p$ 与 u_1 的回归方程，如果回归方程已达到满意的精度，则算法中止。否则继续第二对成分的提取，直到能达到满意的精度为止。

若最终对自变量集提取 r 个成分 $u_1, u_2, \ldots, u_r$，偏最小二乘回归将通过建立 $y_1, y_2, \ldots, y_p$ 与 $u_1, u_2, \ldots, u_r$ 的回归式，然后再表示为与原自变量的回归方程式，即偏最小二乘回归方程式。

为了方便，不妨假定 p 个因变量 $y_1, y_2, \ldots, y_p$ 与 m 个自变量 $x_1, x_2, \ldots, x_m$ 均为标准化变量。自变量组和因变量组的标准化观测数据矩阵分别记为：

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

偏最小二乘回归分析建模的具体步骤如下。

（1）分别提取两变量组的第一对成分，并使之相关性达最大。假设从两组变量分别提出第一对成分为 u_1 和 v_1，u_1 是自变量集 $X = (x_1, x_2, \ldots, x_m)^T$ 的线性组合：

$$u_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1m}x_m$$

v_1 是因变量集 $Y = (y_1, y_2, \ldots, y_p)^T$ 的线性组合：

$$v_1 = \beta_{11}y_1 + \beta_{12}y_2 + \cdots + \beta_{1p}y_p$$

为了回归分析的需要，要求：u_1 和 v_1 各自尽可能多地提取所在变量组的变异信息；u_1 和 v_1 和的相关程度达到最大。

（2）建立 $y_1, y_2, \ldots, y_p$ 对 u_1 的回归及 $x_1, x_2, \ldots, x_p$ 对 u_1 的回归。

（3）用残差阵 A_1 和 B_1 代替 A 和 B 重复以上步骤。

（4）得到 p 个因变量的偏最小二乘回归方程式：

$$y_j = c_{j1}x_1 + c_{j2}x_2 + \cdots + c_{jm}x_m, \ j = 1, 2, \ldots, p$$

（5）交叉有效性检验。

1.3.3 偏最小二乘回归的 matlab 计算

命令：[XL, YL, XS, YS, BETA, PCTVAR, MSE, stats] = plsregress(X, Y, ncomp)

输入：X、Y、ncomp

X：标准化后的原始 X 数据，每列是一项指标。

Y：标准化后的原始 Y 数据，每列是一项指标。

ncomp：选取的主成分对数。

输出：XL, YL, XS, YS, BETA, PCTVAR, MSE, stat

XL：自变量的负荷量矩阵。

YL：自变量的负荷量矩阵。

XS：对应与主成分 u 的得分矩阵。

YS：对应与主成分 v 的得分矩阵。

BETA：最终的回归表达式系数矩阵。

PCTVAR：两行的矩阵，两行分别为自变量和因变量的贡献。

MSE：剩余标准差矩阵。

stats：返回 4 个值。

1.4 模型的应用

参见... 年数学建模... 题优秀论文。