回归分析

定义

回归分析是研究变量间非确定性关系问题的一种统计分析方法。

回归分析解决的问题

1. 描述：用一定的数学表达式去刻画实际事物或现象的一些基本关系和规律，即建立反映变量间非确定关系的数学表达式。
2. 预测：根据实验数据，实现对已有变量间关系的回归分析，并根据在此给定自变量新取值的条件下，对相应因变量之取值进行统计界定的过程。
3. 控制：在利用实验数据实现对已有变量间关系回归分析的基础上，对于给定因边浪所希望大到的取值，推断出应该将相应的自变量限定或控制在什么范围内。

利用回归分析解决问题的基本思路

1. 根据试验数据或者问题背景，初步确定描述变量间关系的表达式（一般通过做散点图完成），即回归模型。
2. 求解回归模型的模型参数
3. 对模型的可信度进行统计检验
4. 用相应的可信模型解决实际问题

一元线性回归

回归模型

回归模型：

$$Y = a + bx + \epsilon$$

称为一元线性回归模型，其中 ϵ 为随机误差，可设

$$\epsilon \sim N(0, \sigma^2)$$

待定常数 a, b 是模型参数，即回归系数，且 a, b, σ^2 均不依赖于 x 。

确定回归系数

首先引入三个统计量：

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

极大似然法

具体步骤：

1. 写出似然函数：

$$L(x_1, x_2, \dots, x_n, a, b) = \prod_{i=1}^{n} f(x_i, a, b)$$

具体有：

$$L(a, b) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2)$$

2. 令似然函数对 a, b 的偏导数分别为 0，求解 a, b
3. 最后结果：

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - (\frac{1}{n}\sum_{i=1}^{n} x_i)\hat{b}$$

最小二乘法

记误差平方和为：

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

求解 a, b 令 Q 最小。同样令 Q 对 a, b 的偏导数为 0，求得的结果和极大似然法一致。

回归方程的显著性检验（检验可信度）

相关系数检验法

样本相关系数定义为：

$$|r| = \sqrt{1 - \frac{Q}{S_{yy}}}$$

其中 Q 为样本误差平方和，S_{yy} 为前面定义的统计量，也即数据整体波动平方和。

样本相关系数的绝对值接近 1，表明变量间线性关系密切；相关系数接近 0，表明变量间没有密切的线性关系。

构造检验统计量

$$t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \sim t(n-2)$$

原假设为 $H_0 : r = 0$，对于显著性水平 α，拒绝域为

$$|t| > t_{1-\frac{\alpha}{2}(n-2)}$$

F 检验法

原假设为 $H_0 : b = 0$，构造检验统计量为

$$F = \frac{(n-2)U}{Q} \sim F(1, n-2)$$

其中

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

拒绝域为

$$F > F_{1-\alpha}$$

关于依分布收敛性的证明，我们可证平方和分解公式：

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

且有自由度：

$$\frac{S_{yy}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{Q}{\sigma^2} \sim \chi^2(n-2), \quad \frac{U}{\sigma^2} \sim \chi^2(1)$$

用回归方程进行预测

有了可信的回归方程，即可根据变量 x 的取值，对相应 y 的取值进行预测，称为点预测。对给定的置信度 $1-\alpha$，可以通过求解在该置信度下预测变量 y 所可能取值的范围，来刻画预测的精度，称为区间预测。

点预测

线性模型：

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

给出的 $\hat{y}$ 就是 y 的无偏估计，可以作为 y 的点预测值。

区间预测

构造枢轴量为：

$$T = \frac{y - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

置信度为 $1-\alpha$ 的预测区间即为：

$$(\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}})$$

其中：

$$\hat{\sigma}^2 = \frac{(1-r^2)S_{yy}}{n-2} = \frac{Q}{n-2}$$

是 σ^2 的无偏估计。当 n 较大且 x_0 偏离 $\bar{x}$ 不远时，预测区间可以近似为：

$$(\hat{y}_0 \pm u_{1-\frac{\alpha}{2}}\hat{\sigma})$$

用回归方程进行控制

确定了可信的回归模型，即可通过 y 的取值来控制 x 的取值范围。不妨设我们要求 $y_1 < y < y_2$，的置信度为 $1-\alpha$。

由区间预测部分的结论可知，当 $x = x_0$ 时，对应的样本 y 值有 $1-\alpha$ 的概率落在区间：

$$(\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}})$$

或者说，样本 y 值分别有 $\alpha/2$ 的概率落在上述区间的两端外。于是我们试图找到 x_1，它使得样本 y 值落在 $(-\infty,\ y_1)$ 的概率为 $\frac{\alpha}{2}$，再找到 x_2，它使得样本 y 值落在 $(y_2, +\infty)$ 的概率为 $\frac{\alpha}{2}$ 也即有方程组：

$$\hat{b}x_1 + \hat{a} - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_1 - \bar{x})^2}{S_{xx}}} = y_1$$

$$\hat{b}x_2 + \hat{a} + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_2 - \bar{x})^2}{S_{xx}}} = y_2$$

由此解出 x_1, x_2。

多元线性回归

回归模型：

$$Y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \epsilon$$

常设 $\epsilon \sim N(0, \sigma^2)$，并有样本观测值为 $(x_{1t}, x_{2t}, x_{3t}, \ldots, x_{kt}; y_t),\ \ 1 \le t \le n$

最小二乘法：

$$\frac{\partial Q}{\partial b_0} = -2\sum_t (y_t - b_0 - \sum_j b_j x_{jt}) = 0$$

$$\frac{\partial Q}{\partial b_i} = -2\sum_t (y_t - b_0 - \sum_j b_j x_{jt})x_{it} = 0$$

设

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

$$b^T = \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_k \end{bmatrix}$$
$$y^T = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$$

可以用矩阵方程表示为：

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

即

$$X^T X\hat{b} = X^T y$$

于是可以解得

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