回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第六讲 简单线性回归模型

2020.3.6

$$y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \ \text{iid} \sim (0, \sigma^2)$$

简单线性模型

简单线性模型只有一个自变量，因为没有控制协变量，所以一般不用来研究响应和自变量之间的因果性，而是用于描述两个变量之间的相依/关联关系（对应于相关系数）。

此前我们介绍的是总体模型，现在我们开始介绍基于样本数据对总体模型进行统计分析。

假设数据$(x_i, y_i), i = 1,2,...,n$，来自于总体模型

$$y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2), \varepsilon \text{与} x \text{独立}$$

即$(x_i, y_i), i = 1,2,...,n$，满足模型

$$y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \ \text{iid} \sim (0, \sigma^2), \varepsilon_i \text{与} x_i \text{独立}$$

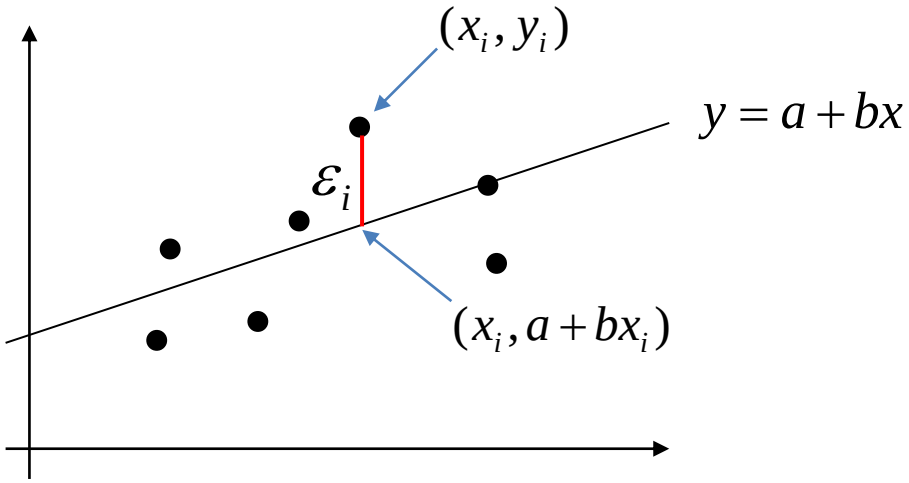
最小二乘法

最小二乘法：

最佳的直线使得误差平方和达到最小：

$$\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

极小化误差平方和称为最小二乘法(Least Squares, LS).



误差平方和对a,b求导，得正则方程：

$$\begin{cases}\sum \varepsilon_i = \sum (y_i - a - bx_i) = 0 & \text{对应于条件} E(\varepsilon) = 0 \\ \sum x_i \varepsilon_i = \sum x_i (y_i - a - bx_i) = 0 & \text{对应于条件} E(x\varepsilon) = 0\end{cases}$$

换言之，正则方程可看作是矩估计方程

令总体矩$E(\varepsilon) = (\varepsilon_1 + \ldots + \varepsilon_n)/n$样本矩：

$$\varepsilon_1 + \ldots + \varepsilon_n = 0$$

令总体矩$E(x\varepsilon) = (x_1\varepsilon_1 + \ldots + x_n\varepsilon_n)/n$样本矩：

$$x_1\varepsilon_1 + \ldots + x_n\varepsilon_n = 0$$

正则方程是合理的：它们与模型假设相容

正则方程：$\begin{cases}\sum \varepsilon_i = \sum (y_i - a - bx_i) = 0 \\ \sum x_i \varepsilon_i = \sum x_i (y_i - a - bx_i) = 0\end{cases}$

第一个方程 $\Rightarrow a = \bar{y} - b\bar{x},$

代入第二个方程 $\Rightarrow \sum x_i \left[y_i - \bar{y} - b(x_i - \bar{x}) \right] = 0$

LS估计：

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$$\hat{b} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \overset{\text{记为}}{=} s_{xy}/s_{xx}$$

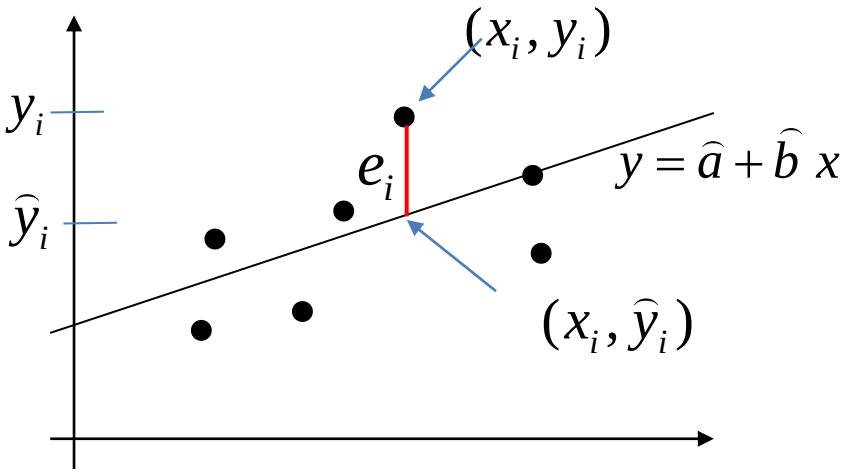
记号：$s_{xx} = \sum (x_i - \bar{x})^2$

$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i$

拟合值和残差

拟合值：$\hat{y}_i = \hat{a} + \hat{b}x_i$

残差：$e_i = y_i - \hat{y}_i$



残差平方和RSS (Residual Sum of Squares)：

$$RSS = \sum e_i^{\,2} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{a} - \hat{b}x_i)^2$$

命题1. $\sum e_i = 0, \quad \sum e_i x_i = 0, \quad \sum e_i \hat{y}_i = 0,$

证：前两个正是正则方程:

$$\sum e_i = \sum (y_i - \hat{a} - \hat{b}x_i) = 0, \quad \sum x_i e_i = \sum x_i (y_i - \hat{a} - \hat{b}x_i) = 0$$

所以 $\sum e_i \hat{y}_i = \sum e_i (\hat{a} + \hat{b}x_i) = \hat{a}\sum e_i + \hat{b}\sum e_i x_i = 0$

注1. 因为所有残差之和为0，$\bar{e} = \sum e_i / n = 0$, 故 残差平方和 $RSS = \sum e_i^2 = \sum (e_i - \bar{e})^2$ 度量了所有残差的变化程度。

注2. 拟合值的样本均值等于$\bar{y}$:

$$0 = \sum e_i = \sum (y_i - \hat{y}_i) \Rightarrow \sum \hat{y}_i / n = \sum y_i / n = \bar{y}$$

$$\mathrm{var}(y)=\mathrm{var}\big(E(y\mid\mathbf{x})\big)+E\big(\mathrm{var}(y\mid\mathbf{x})\big)$$

复相关系数平方

我们对总平方和s_{yy}进行分解：

$$s_{yy}=\sum(y_i-\bar{y})^2=\sum(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2+2\sum(\hat{y}_i-\bar{y})e_i$$

$$=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2 \;\widehat{=}\; s_{\hat{y}\hat{y}}+\mathrm{RSS}$$

$\sum(\hat{y}_i-\bar{y})^2=s_{\hat{y}\hat{y}}$是拟合值$\{\hat{y}_i,i=1,..,n\}$的变差平方和，度量了回归直线上的拟合值的变化程度($s_{\hat{y}\hat{y}}/(n-1)$是拟合值的样本方差)。

$s_{\hat{y}\hat{y}}=\sum(\hat{y}_i-\bar{y})^2 \;\widehat{=}\; \mathrm{SS}_{回}$称为回归平方和

(*)式的平方和分解：

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$s_{yy} \quad = \quad s_{\hat{y}\hat{y}} \quad + \quad \mathrm{RSS}$$

通常写成：$\mathrm{SS}_{总} = \mathrm{SS}_{回} + \mathrm{SSE}$

解释为：总平方和或总方差 = 回归线上的变化(x变化导致的) + 偏离回归线的变化(x不能解释的）

注意各种记号：SS : Sum of Squares

$\mathrm{SS}_{总} = s_{yy}$，$\mathrm{SS}_{回} = s_{\hat{y}\hat{y}}$，$\mathrm{SSE} = RSS = s_{ee}$.

定义：复相关系数平方R^2定义为回归函数(或自变量)所能解释的响应变量总平方和的百分比：

$$R^2=\frac{\mathrm{SS}_{回}}{\mathrm{SS}_{总}}=\frac{s_{\hat{y}\hat{y}}}{s_{yy}}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

化简回归平方和：

$$\mathrm{SS}_{回}=s_{\hat{y}\hat{y}}=\sum(\hat{a}+\hat{b}x_i-(\hat{a}+\hat{b}\bar{x}))^2=\hat{b}^2s_{xx}=s_{xy}^2/s_{xx}=r_{xy}^2s_{yy}$$

$$\Rightarrow \mathrm{RSS}=s_{yy}-s_{xy}^2/s_{xx}=(1-r_{xy}^2)\ s_{yy}$$

命题2. 对于简单线性回归模型$y_i=a+bx_i+\varepsilon_i,\varepsilon_i\sim(0,\sigma^2),i=1,\ldots,n$

$$R^2=r_{xy}^2=s_{xy}^2/s_{xx}s_{yy}=\frac{s_{yx}s_{xx}^{-1}s_{xy}}{s_{yy}}$$

其中r_{xy}为$(x_i,y_i),i=1,\ldots,n$的样本相关系数. 另外，

$$\mathrm{SS}_{回}=r_{xy}^2s_{yy},\quad RSS=(1-r_{xy}^2)\ s_{yy}.$$