课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第六讲 简单线性回归模型

2022.10.9

此前我们介绍了线性回归的总体模型。以后我们主要处理样本版本，基本工具是最小二乘。

简单线性回归模型

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）。

简单线性模型（数据模型）

假设独立样本$(x_i, y_i), i = 1,2,..., n$，来自于总体模型

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

即$(x_i, y_i), i = 1,2,..., n$，满足模型

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

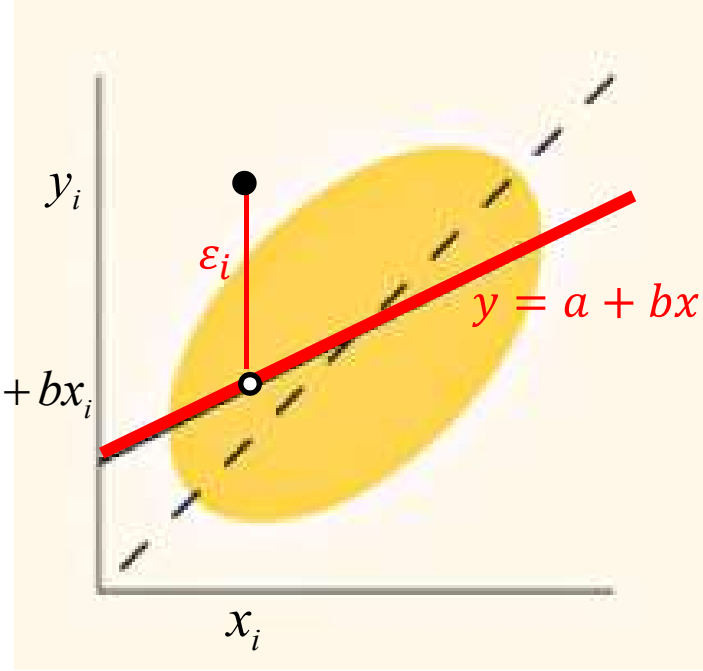
最小二乘法

最小二乘

以x_i的线性函数预测 y_i的误差： $\varepsilon_i = y_i - a - bx_i = y_i - E(y_i \mid x_i)$

极小化误差平方和称为最小二乘法(Least Squares, LS)：

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



为什么不是极小化数据点到直线的垂直距离？

参见后面的对称回归

命题1. 最小二乘估计

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

其中 $s_{xx}=\sum(x_i-\bar{x})^2$，$s_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum(x_i-\bar{x})y_i.$

证：误差平方和$\sum_{i=1}^{n}(y_i-a-bx_i)^2$对$a,b$求导，得正则方程：

$$\begin{cases}\Sigma\varepsilon_i=\sum(y_i-a-bx_i)=0\\ \Sigma x_i\varepsilon_i=\sum x_i(y_i-a-bx_i)=0\end{cases}$$

由第一个方程得$a=\bar{y}-b\,\bar{x}$,代入第二个方程得:

$$\sum x_i(y_i-\bar{y}-b(x_i-\bar{x}))=0$$

$\Rightarrow$ LS估计 $\hat{b}=\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$，$\hat{a}=\bar{y}-\hat{b}\,\bar{x}$,

回忆：$b=\Sigma_{xx}^{-1}\Sigma_{xy}$，$a=\mu_y-b\mu_x\Rightarrow$矩估计$\hat{b}=\frac{s_{xy}}{s_{xx}}$，$a=\bar{y}-\hat{b}\bar{x}$,与LS估计相同。

LS估计的均值和方差

命题1. (1) $\hat{a},\hat{b}$是无偏估计，即$E(\hat{a})=a,\ \ E(\hat{b})=b.$

(2) 给定所有自变量 $\mathbf{x}=(x_1,...,x_n)^{\mathsf{T}}$ 条件下，

$$\mathrm{var}\left(\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix}\middle|\mathbf{x}\right)=\begin{pmatrix}\sigma^2/n+\bar{x}^2\sigma^2/s_{xx} & -\bar{x}\sigma^2/s_{xx}\\ -\bar{x}\sigma^2/s_{xx} & \sigma^2/s_{xx}\end{pmatrix}。$$

证明: (1) 因为ε_i与x_i独立，所以$E(\varepsilon_i\mid x_i)=E(\varepsilon_i)=0,$

所以$E(y_i\mid x_i)=E(a+bx_i+\varepsilon_i\mid x_i)=a+bx_i+E(\varepsilon_i\mid x_i)=a+bx_i$

$$\Rightarrow E(\hat{b}\mid\mathbf{x})=E\big(\Sigma(x_i-\bar{x})y_i\mid\mathbf{x}\big)/s_{xx}=\Sigma(x_i-\bar{x})E(y_i\mid x_i)/s_{xx}$$

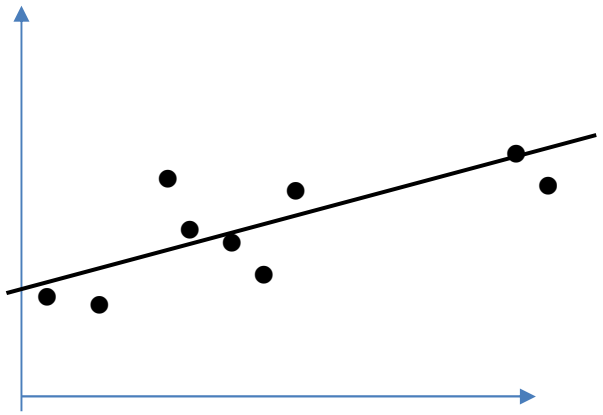
$$=\Sigma(x_i-\bar{x})(a+bx_i)/s_{xx}=b\Sigma(x_i-\bar{x})x_i/s_{xx}=b\Rightarrow E(\hat{b})=b$$

另外, $E(\hat{a}\mid\mathbf{x})=E(\bar{y}-\hat{b}\bar{x}\mid\mathbf{x})=a+b\bar{x}-E(\hat{b}\mid\mathbf{x})\bar{x}=a.$

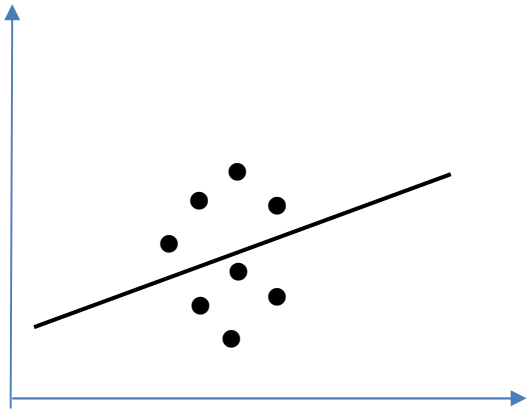
(2) 因为 $\mathrm{var}(y_i\mid x_i)=\sigma^2$,所以

$$\mathrm{var}(\hat{b}\mid\mathbf{x})=\mathrm{var}\left(\frac{\Sigma(x_i-\bar{x})y_i}{\Sigma(x_i-\bar{x})^2}\middle|\mathbf{x}\right)=\frac{\Sigma(x_i-\bar{x})^2\,\mathrm{var}(y_i\mid x_i)}{\left[\Sigma(x_i-\bar{x})^2\right]^2}=\frac{\sigma^2}{s_{xx}},\text{其它略。}$$

注：*LS*估计的方差公式表明自变量的样本方差$s_x^2 = s_{xx}/(n-1)$越大，LS估计的方差越小。例如，右图的自变量方差比左图小（较为集中），右图的回归直线更难以确定（参数估计的方差较大）。



自变量方差较大，
斜率估计稳定/精确



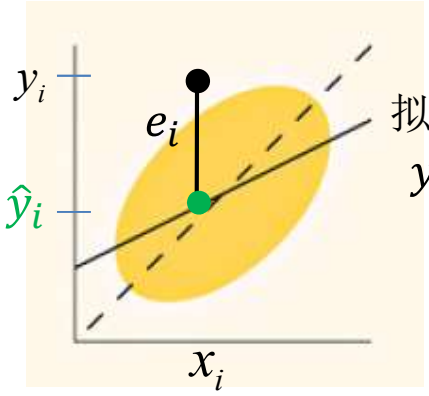
自变量方差较小，
斜率估计不稳定

拟合值与残差

拟合回归直线：$y = \hat{a} + \hat{b}x$

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

残差：$e_i = y_i - \hat{y}_i.$



正则方程：$\sum e_i = 0,\ \sum x_i e_i = 0$

由此得到如下事实：

(1) 残差与拟合值不相关：$\sum \hat{y}_i e_i = 0$

(2) 残差的均值：$\bar{e} = \sum e_i / n = 0$

(3) 拟合值的样本均值：$\bar{\hat{y}} = \sum \hat{y}_i / n = \bar{y}$

$\leftarrow$ $\mathbf{e} \perp \mathbf{1}, \mathbf{e} \perp \mathbf{x}$

$\leftarrow$ $\mathbf{e} \perp \hat{\mathbf{y}}$

向量记号：

$\mathbf{1} = (1,...,1)^{\mathsf{T}},$

$\mathbf{x} = (x_1,...,x_n)^{\mathsf{T}},$

$\hat{\mathbf{y}} = (\hat{y}_1,...,\hat{y}_n)^{\mathsf{T}},$

$\mathbf{e} = (e_1,...,e_n)^{\mathsf{T}}$

验证：(1)$\sum \hat{y}_i e_i = \sum (\hat{a} + \hat{b}x_i) e_i = \hat{a}\sum e_i + \hat{b}\sum x_i e_i = 0$

(3)$\sum \hat{y}_i = \sum y_i + \sum e_i = \sum y_i$

平方和

定义：$a_1,\ldots,a_n$的平方和 $s_{aa}=\sum(a_i-\bar{a})^2=\|\mathbf{a}-\bar{a}\mathbf{1}\|^2$

(i) 总平方和：　　响应的平方和

$$SS_{总}=s_{yy}=\sum(y_i-\bar{y})^2=\|\mathbf{y}-\bar{y}\mathbf{1}\|^2$$

(ii) 回归平方和：拟合值的平方和

$$SS_{回}=s_{\hat{y}\hat{y}}=\sum(\hat{y}_i-\bar{y})^2=\|\hat{\mathbf{y}}-\bar{y}\mathbf{1}\|^2$$

(iii) 残差平方和：残差的平方和

$$RSS=s_{ee}=\sum e_i^2=\|\mathbf{e}\|^2$$

$\bar{\hat{y}}=\bar{y}$

$\bar{e}=0$

命题2.

(a) $SS_{回}=s_{xy}^2/s_{xx}$

(b) $RSS=s_{yy}-s_{xy}^2/s_{xx}$

(c) 平方和分解$SS_{总}=SS_{回}+RSS$

证：（a）

$$SS_{回}=\sum(\hat{y}_i-\bar{y})^2=\sum(\hat{a}+\hat{b}x_i-\bar{y})^2=\sum(\bar{y}-\hat{b}\bar{x}+\hat{b}x_i-\bar{y})^2\hat{b}^2s_{xx}=s_{xy}^2/s_{xx}$$

（b）$RSS=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{a}-\hat{b}x_i)^2=\sum(y_i-\bar{y}+\hat{b}\bar{x}-\hat{b}x_i)^2$

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

$=s_{yy}+\hat{b}^2s_{xx}-2\hat{b}s_{xy}=s_{yy}-s_{xy}^2/s_{xx}$

↑以向量形式更容易表述：

(b) $\mathbf{y}=\hat{\mathbf{y}}+\mathbf{e}\Rightarrow\mathbf{y}-\bar{y}\mathbf{1}=(\hat{\mathbf{y}}-\bar{y}\mathbf{1})+\mathbf{e},$

其中 $\mathbf{e}\perp\mathbf{1},\mathbf{e}\perp\hat{\mathbf{y}}\Rightarrow\mathbf{e}\perp\hat{\mathbf{y}}-\bar{y}\mathbf{1},$

所以 $\|\mathbf{y}-\bar{y}\mathbf{1}\|^2=\|\hat{\mathbf{y}}-\bar{y}\mathbf{1}\|^2+\|\mathbf{e}\|^2$

$SS_{总}=SS_{回}+RSS$

决定系数
coefficient of determination

（样本版本的）决定系数R^2定义为自变量所能解释的响应变量总平方和的百分比：

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

命题3. 对于简单线性回归模型 $y_i=a+bx_i+\varepsilon_i,\varepsilon_i\sim(0,\sigma^2),i=1,\ldots,n$

$$R^2=r_{xy}^2\quad\left(=s_{xy}^2/s_{xx}s_{yy}=\frac{s_{yx}s_{xx}^{-1}s_{xy}}{s_{yy}}\right)$$

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

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

证：因为$SS_{回}=s_{xy}^2/s_{xx}$，$SS_{总}=s_{yy}$

$$R^2=\frac{SS_{回}}{s_{yy}}=\frac{s_{xy}^2}{s_{xx}s_{yy}}=\left(r_{xy}\right)^2$$

误差方差的估计

因为$\sigma^2 = \mathrm{var}(\varepsilon_i) = E(\varepsilon_i^2)$，其中$\varepsilon_i = y_i - a - bx_i$，而$e_i = y_i - \hat{a} - \hat{b}x_i$可看作是r.v. ε_i的预测，我们以e_i，$i = 1,2,...,n$的样本方差估计σ^2。

误差方差的估计

σ^2的“LS”估计取为：$\hat{\sigma}^2 = \frac{1}{n-2} RSS = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

惯例：

- 虽然$\hat{\sigma}^2$不是由最小二乘法直接 得到的，但通常也称之 为LS估计。
- 为什么除以$n-2$而不是$n-1$? 因为估计了两个参数a和b。
- 记 $s_y^2 = s_{yy}/(n-1)$ 为$y_1,...,y_n$的样本方差,则

$$\hat{\sigma}^2 = \frac{1}{n-2} RSS = \frac{1}{n-2}(1-r_{xy}^2)\ s_{yy} = \frac{n-1}{n-2}(1-r_{xy}^2)\ s_y^2 .$$

(对比参数关系： $\sigma^2 = (1-\rho_{xy}^2)\sigma_y^2$)

残差及残差平方和与响应变量有关，我们将它们表示成误差ε_i的函数（$\varepsilon_i, i = 1, \ldots, n$ iid均值为0，容易处理）

引理1.（1）$e_i = (\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x}) s_{x\varepsilon} / s_{xx}.$

（2）$RSS = s_{yy} - s_{xy}^2 / s_{xx} = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2 / s_{xx}$

证明:(1) $\hat{b} = \dfrac{\Sigma(x_i - \bar{x}) y_i}{\Sigma(x_i - \bar{x})^2} = \dfrac{\Sigma(x_i - \bar{x})(a + bx_i + \varepsilon_i)}{\Sigma(x_i - \bar{x})^2} = b + s_{x\varepsilon} / s_{xx},$

所以$e_i = y_i - \hat{y}_i = a + bx_i + \varepsilon_i - (\hat{a} + \hat{b}x_i) = a + bx_i + \varepsilon_i - (\bar{y} - \hat{b}\bar{x} + \hat{b}x_i)$

$= a + bx_i + \varepsilon_i - [a + b\bar{x} + \bar{\varepsilon} + (b + s_{x\varepsilon} / s_{xx})(x_i - \bar{x})]$

$= (\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x}) s_{x\varepsilon} / s_{xx}.$

（2）$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left((\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x}) s_{x\varepsilon} / s_{xx} \right)^2$

$= \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 - s_{x\varepsilon}^2 / s_{xx} = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2 / s_{xx}.$

误差方差估计的无偏性

命题4. $\hat{\sigma}^2$是σ^2的无偏估计，即 $E(\hat{\sigma}^2)=\sigma^2$。

证明1：由引理1(2), $RSS=s_{\varepsilon\varepsilon}-s_{x\varepsilon}^2/s_{xx}$。显然$E(s_{\varepsilon\varepsilon})=(n-1)\sigma^2$,

$$Es_{x\varepsilon}^2=E\left(\sum_{i=1}^n(x_i-\bar{x})\varepsilon_i\right)^2=\sum_{i=1}^n E(x_i-\bar{x})^2\varepsilon_i^{\,2})=s_{xx}\sigma^2$$

$$\Rightarrow E(RSS)=E(s_{\varepsilon\varepsilon})-Es_{x\varepsilon}^2/s_{xx}=(n-2)\sigma^2。$$

证明2：因为 $SS_{回}=\hat{b}^2s_{xx}, RSS=s_{yy}-\hat{b}^2s_{xx}$，给定$x_1,...,x_n$的条件下

(i) $E(\hat{b}^2)=\mathrm{var}(\hat{b})+\left(E(\hat{b})\right)^2=\sigma^2/s_{xx}+b^2$

(ii) $E(s_{yy})=E\sum_{i=1}^n(y_i-\bar{y})^2=\sum_{i=1}^n Ey_i^{\,2}-nE(\bar{y})^2$

$$=\sum_{i=1}^n\left(\mathrm{var}(y_i)+(Ey_i)^2\right)-n\left(\mathrm{var}(\bar{y})+(E\bar{y})^2\right)$$

$$=n\sigma^2+\sum_{i=1}^n\left(a+bx_i\right)^2-n\left(\sigma^2/n+(a+b\bar{x})^2\right)=(n-1)\sigma^2+b^2s_{xx},$$

$$\Rightarrow E(RSS)=E(s_{yy})-E(\hat{b}^2s_{xx})=(n-1)\sigma^2+b^2s_{xx}-s_{xx}(\sigma^2/s_{xx}+b^2)$$

$=(n-2)\sigma^2$, 所以$E(\hat{\sigma}^2\,|\,\mathbf{x})=\sigma^2, E(\hat{\sigma}^2)=\sigma^2$