课程主页：http://staff.ustc.edu.cn/~ynyang/2023

第六讲 简单线性回归模型

2023.10.20

第五讲：总体模型 $y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2)$

这一讲： 样本$(y_i, x_i), i = 1, \ldots, n,$来自于上述模型:

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

统计推断将在给定自变量$x_i's$的条件下进行。

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，而是仅用来描述两个变量之间的依赖关系和发现规律。

一般的多变量回归模型的统计推断将采用矩阵-向量表达和处理，缺点是矩阵和向量中的细节有时难以理解。

对于简单回归模型，既可以使用矩阵-向量语言处理，也可以使用初等代数运算。我们将使用初等代数，便于理解和容易操作，也为理解以后的矩阵-向量提供支持。

简单线性模型

二元随机向量(x, y)满足模型

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

-

回归效应

由上一讲命题2,我们知道简单模型的参数由x, y的均值和方差、协方差决定

记$\mu_x = E(x), \mu_y = E(y), \sigma_x^{\ 2} = \mathrm{var}(x) = \Sigma_{xx}, \sigma_y^{\ 2} = \mathrm{var}(y) = \Sigma_{yy},$
$\rho = \rho_{xy}, \Sigma_{xy} = \mathrm{cov}(x, y) = \rho\sigma_x\sigma_y$,则回归系数和误差方差由这些参数决定:

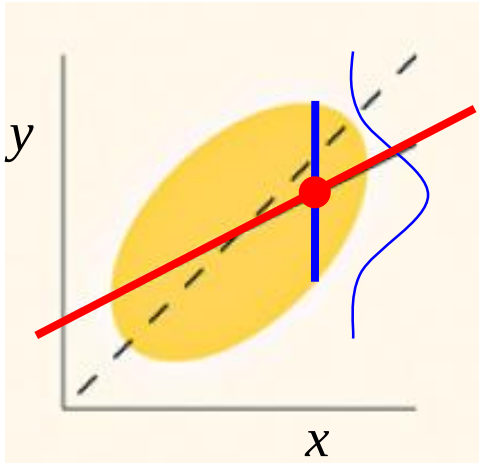
$$b = \frac{\Sigma_{xy}}{\Sigma_{xx}} = \rho\frac{\sigma_y}{\sigma_x}, \quad a = \mu_y - b\mu_x, \sigma^2 = (1-\rho^2)\sigma_y^2$$

所以回归函数或均值函数(实线):

$$E(y \mid x) = a + bx = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x),$$

其中的$\rho, |\rho| < 1$导致回归效应:
当x比其均值μ_x大k个标准差，即$x = \mu_x + k\sigma_x$时
$E(y \mid x) = \mu_y + \rho k\sigma_y$比其均值$\mu_y$大$\rho k < k$个标准差.

而x, y等比例增加的情况对应于上述方程中$\rho = 1$的情况(虚线).



关键：回归函数是条件期望

$E(y \mid x) = a + bx$

x, y地位不等。

直观解释：

假设(x, y)服从二元正态，虚线等于或接近椭圆的对称轴。

对给定的$x > \mu_x$,变量y服从图中所示的正态分布

该正态分布的中心(红点)：$E(y \mid x) = \mu_y + \rho \dfrac{\sigma_y}{\sigma_x}(x - \mu_x)$

它是蓝色线段的中点，在虚线下方。

最小二乘法(Least Squares)

简单线性模型（样本模型）

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

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

即$(x_i, y_i), i = 1,2,\ldots,n$，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

以x_i预测y_i的误差：

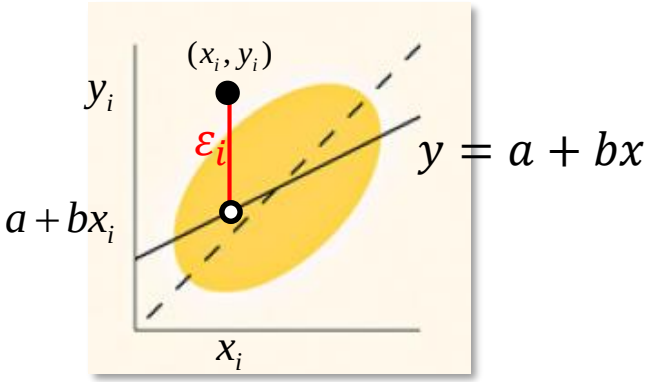
$$\varepsilon_i = y_i - a - bx_i = y_i - E(y_i \mid x_i)$$

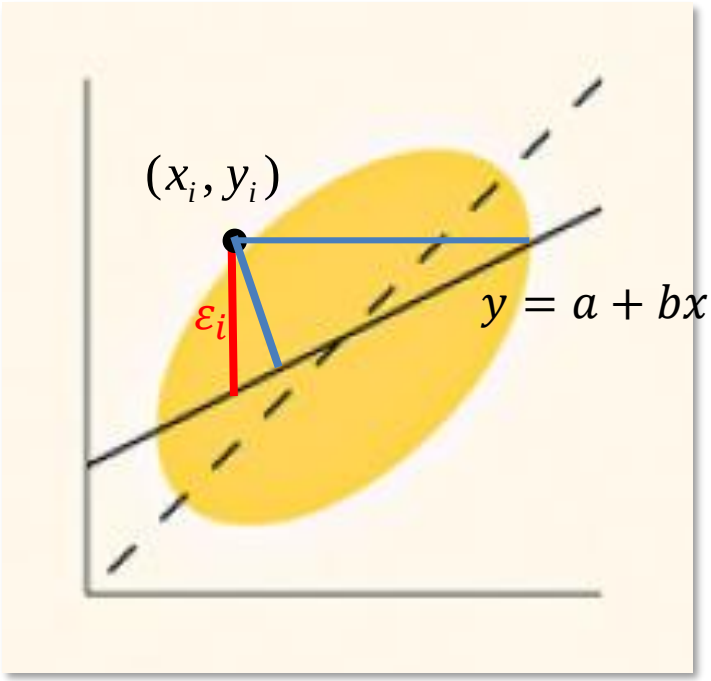
为了求解回归直线，即求解a, b,

最小二乘法（LS，Least Squares）

极小化误差平方和：

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





为什么是竖直方向的误差，而不是水平方向的误差，甚至点到直线的垂直距离？

这是因为x和y地位不同，线性模型以x的函数描述y：

$$y = a + bx + \text{error}$$

如果认为x和y地位对称，如果我们希望求解一条直线能最好地描述两者的关系，那么该直线应该写成对称形式：

$$ax + by = \text{c}$$

此时，可极小化点到直线的垂直距离，这称为对称回归（total least squares），此时的解为虚线。

记号：$s_{aa}=\sum(a_i-\bar{a})^2=\sum(a_i-\bar{a})a_i$，
$s_{ab}=\sum(a_i-\bar{a})(b_i-\bar{b})=\sum(a_i-\bar{a})b_i$.

命题1. 最小二乘问题$\min\limits_{a,b}\sum\limits_{i=1}^{n}(y_i-a-bx_i)^2$的最优解，即最小二乘估计

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

LS估计与矩估计完全相同（参见上一讲最后一页）。

证：误差平方和$\sum\limits_{i=1}^{n}(y_i-a-bx_i)^2$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 $\hat{b}=\dfrac{\sum x_i(y_i-\bar{y})}{\sum(x_i-\bar{x})x_i}=\dfrac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$，$\hat{a}=\bar{y}-\hat{b}\,\bar{x}$，

评注LS估计

LS估计是否符合直观？只考虑$\hat{b}$，形式上

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

$$= \frac{\sum(x_i - \bar{x})^2 \left(\frac{y_i}{x_i - \bar{x}}\right)}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})^2 \left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right)}{\sum(x_i - \bar{x})^2} \triangleq \Sigma w_i z_i$$

其中权重 $w_i = \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$，$\sum w_i = 1$，$z_i = \frac{y_i - \bar{y}}{x_i - \bar{x}}$ 或 $\frac{y_i}{x_i - \bar{x}}$

所以 $\hat{b}$ 是每个样本点处的局部斜率估计z_i的加权平均，斜率z_i较容易理解，但权重w_i为什么是上述形式？

引理：假设$z_1, \ldots, z_n$独立，$\sigma_i^2 = \mathrm{var}(z_i)$已知但不等，假设权重$w_i \geq 0$, $i = 1, .., n, \sum w_i = 1$，则$\frac{\sum z_i / \sigma_i^2}{\sum 1/\sigma_i^2}$是所有加权平均$\sum w_i z_i$中方差最小的，即

$$\mathrm{var}(\sum w_i z_i) = \sum w_i^2 \sigma_i^2 \geq \frac{1}{\sum 1/\sigma_i^2}.$$

证明：由CS不等式，$\mathrm{var}(\sum w_i z_i) = \sum w_i^2 \sigma_i^2 \geq \frac{1}{\sum 1/\sigma_i^2}$，$w_i \propto 1/\sigma_i^2$时等号成立。

考虑 b的加权平均估计 $\tilde{b} = \sum w_i z_i$，$w_i \geq 0$，$\sum w_i = 1$，因为

$$var(z_i|x_i) = var\left(\frac{y_i}{x_i - \bar{x}}|x_i\right) = \sigma^2/(x_i - \bar{x})^2,$$

因此最优权重$w_i \propto (x_i - \bar{x})^2/\sigma^2$，此时$\tilde{b} = \sum w_i z_i$ = LS估计。

以上讨论说明了LS估计基本符合直观，权重是最优选择（使得方差最小）。

其它选择

其它权重选择也是允许的，这依赖于具体问题背景，例如

$$\tilde{b}_{EV} = \Sigma(x_i - \bar{x})(y_i - \bar{y})^2/\Sigma(x_i - \bar{x})^2(y_i - \bar{y})$$

Eror-in-variable 模型的广义据估计(GMM)

$$\tilde{b}_{\text{IVLS}} = \Sigma(u_i - \bar{u})(y_i - \bar{y})/\Sigma(u_i - \bar{u})(x_i - \bar{x}).$$

GM假设第三条不成立时的工具变量估计

其中的u_i是外生变量/工具变量（2021诺贝尔经济）。

我们甚至可以考虑其它局部斜率，比如 $z_{ij} = \frac{y_i - y_j}{x_i - x_j}$，$i, j = 1, \ldots, n$， 并构造 b的估计 $\tilde{b} = \sum w_{ij} z_{ij}$，这称为U统计量，如何选择权重？有没有人研究过此类估计？

统计推断一般在自变量给定的条件下进行，在自变量给定时计算LS估计的方差。

LS估计的均值和方差

命题2. (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，

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

证明: (1) ε_i与x_i独立 $\Rightarrow E(y_i\mid x_i)=a+bx_i,$

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

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

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

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

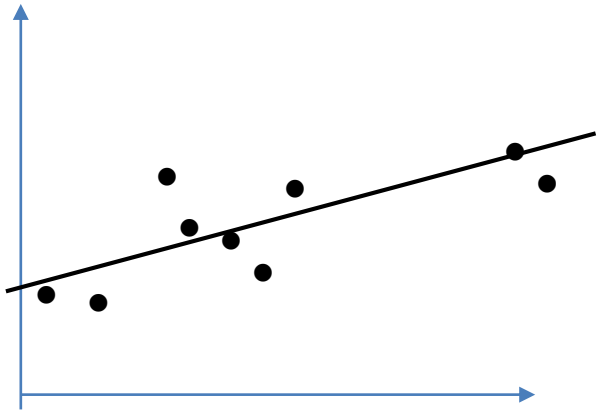
$$\mathrm{var}(\hat{b}\mid\mathbf{x})=\mathrm{var}\left(\left.\frac{\Sigma(x_i-\bar{x})y_i}{\Sigma(x_i-\bar{x})^2}\right|\mathbf{x}\right)=\frac{\Sigma(x_i-\bar{x})^2\,\mathrm{var}(y_i\mid x_i)}{\left[\Sigma(x_i-\bar{x})^2\right]^2}=\frac{\sigma^2}{s_{xx}},\text{其它略。}$$

注：如何理解命题2的方差结果？

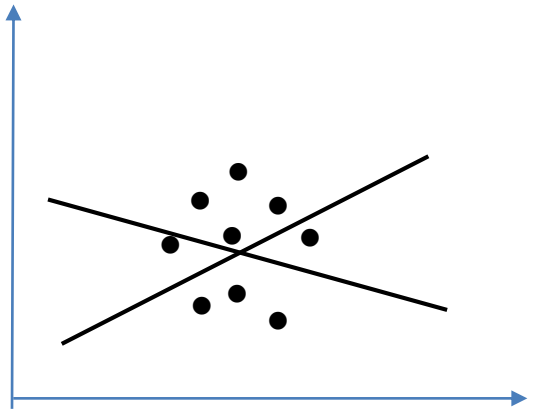
- 当$\bar{x}=0$时，$\hat{a},\hat{b}$不相关，直观解释？
- $\text{var}(\hat{b}\,|\,\mathbf{x})=\dfrac{\sigma^2}{s_{xx}}=\dfrac{\sigma^2}{(n-1)s_x^2}$，其中$s_x^2=s_{xx}/(n-1)$为自变量的样本方差。

这说明自变量方差越大，LS估计的方差越小,估计精度越高。

例如，右图的自变量方差较小，斜率难以估计($\hat{b}$的方差大)。



自变量方差较大，
斜率估计稳定/精确



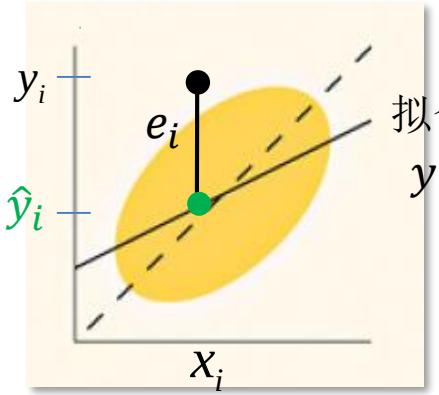
自变量方差较小，
斜率难以估计

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：$\sum e_i = 0,\ \sum x_i e_i = 0 \Rightarrow$

(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：

$\mathbf{1} = (1,\ldots,1)^{\mathrm{T}}$,

$\mathbf{x} = (x_1,\ldots,x_n)^{\mathrm{T}}$,

$\hat{\mathbf{y}} = (\hat{y}_1,\ldots,\hat{y}_n)^{\mathrm{T}}$,

$\mathbf{e} = (e_1,\ldots,e_n)^{\mathrm{T}}$

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)

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

(除以$n-1$就是响应$y_1,\ldots,y_n$的样本方差)

(ii) 回归平方和(拟合值的平方和)

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

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

(除以$n-1$就是拟合值$\hat{y}_1,\ldots,\hat{y}_n$的样本方差)

(iii) 残差平方和：

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

$\bar{e}=0$

(除以$n-1$就是残差$e_1,\ldots,e_n$的样本方差)

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：

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

命题3. $SS_{回} = s_{xy}^2/s_{xx}$, $RSS = s_{yy} - s_{xy}^2/s_{xx}$, $R^2 = r_{xy}^2$,
因此$SS_{总} = s_{yy} = SS_{回} + \mathrm{RSS}$。

证：$(a)\ SS_{回} = \sum(\hat{y}_i - \bar{y})^2 = \sum(\hat{a} + \hat{b}x_i - \bar{y})^2$

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

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

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

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

$(c)\ R^2 = SS_{回} / SS_{总} = (s_{xy}^2/s_{xx}) / s_{yy} = r^2.$

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，我们尝试基于$\{e_i,\ i = 1,2,...,n\}$估计σ^2。

误差方差的估计

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

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。

残差及残差平方和是$y_i's$的函数，也可表示成误差$\varepsilon_i's$的函数（引理1）。

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

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

注：从命题3知$RSS = s_{yy} - s_{xy}^2 / s_{xx}$，引理1(2）说明其中的$y_i = a + bx_i + \varepsilon_i$可以换成$\varepsilon_i$，也即RSS与自变量$x's$无关-这并不显然。

后面我们将从投影的角度来看，这是显然的（可以猜测到的）。

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$E(s_{\varepsilon\varepsilon})=(n-1)\sigma^2$,

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

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

证明2(基于表达$RSS=s_{yy}-s_{xy}^2/s_{xx}$证明无偏性，比基于$\varepsilon$'s计算要复杂一些)

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

总结：简单模型的参数及其LS估计

模型(总体)： $y = a + bx + \varepsilon$,

$\varepsilon \sim (0, \sigma^2)$, ε与x独立

模型(样本)： $y_i = a + bx_i + \varepsilon_i$,

$\varepsilon_i \sim (0, \sigma^2)$, ε_i与x_i独立

参数	估计
（1） $b = \mathrm{cov}(x, y) / \mathrm{var}(x) = \rho_{xy}\sigma_y / \sigma_x$ （2） $a = \mu_y - b\mu_x$, （3） $\sigma^2 = (1 - \rho_{xy}^2)\sigma_y^2$	(1) $\hat{b} = s_{xy} / s_{xx} = r_{xy} s_y / s_x$ (2) $\hat{a} = \bar{y} - \hat{b}\bar{x}$ (3) $\hat{\sigma}^2 = (1 - r_{xy}^2)s_y^2 \times (n-1)/(n-2)$
回归函数 : $a + bx$ 误差 : $\varepsilon = y - (a + bx)$ ε与x独立	拟合值： $\hat{y}_i = \hat{a} + \hat{b}x_i$ 残差$e_i = y_i - (\hat{a} + \hat{b}x_i)$ $(e_1, \ldots, e_n)^{\mathsf{T}} \perp (x_1, \ldots, x_n)^{\mathsf{T}}$
$R^2 = \dfrac{\mathrm{var}(a + bx)}{\mathrm{var}(y)} = \rho_{xy}^2$	$R^2 = \dfrac{SS_{回}}{SS_{总}} = \dfrac{s_{\hat{y}\hat{y}}}{s_{yy}} = r^2$

“LS估计的方差”的估计

Plug-in

$\text{var}(\hat{b}\,|\,\mathbf{x}) = \dfrac{\sigma^2}{s_{xx}}$中将$\sigma^2$的估计代入(plug-in)得：

$$\widehat{\text{var}}(\hat{b}\,|\,\mathbf{x}) = \frac{\hat{\sigma}^2}{s_{xx}},$$

标准差：$se(\hat{b}) = \sqrt{\widehat{\text{var}}(\hat{b}\,|\,\mathbf{x})} = \hat{\sigma}/\sqrt{s_{xx}}$,

截距项LS估计$\hat{a}$的方差估计类似得到（一般不关心）。

Wald检验

Wald检验方法是构造检验统计量常用的方法之一（另外两种常用方法是似然比检验，Score检验）。一般地，若参数θ的估计为$\hat{\theta}$,其标准差为$se(\hat{\theta})$,则$H_0:\theta=\theta_0$的Wald检验统计量定义为：

$$W = (\hat{\theta}-\theta_0)/se(\hat{\theta})。$$

对于简单模型的$H_0:b=b_0, b_0$已知，Wald检验统计量$W = \dfrac{\hat{b}-b_0}{se(\hat{b})} = \dfrac{\sqrt{s_{xx}}(\hat{b}-b_0)}{\hat{\sigma}}$

正态模型下的统计推断

对截距项一般没必要做统计推断，只有斜率b（x的效应）才是我们关心的，下面只考虑正态假设下b的统计推断（非正态情形有大样本检验或置信区间）。

命题5. 假设模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_1, ..., \varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$, 则

(1) $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$

(2) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, 且$\hat{\sigma}^2$与$(\hat{a}, \hat{b})$独立

(3) $\dfrac{\sqrt{s_{xx}}(\hat{b} - b)}{\hat{\sigma}} \sim t_{n-2}$

证明：(1) $y_i \mid x_i \sim N(a + bx_i, \sigma^2) \Rightarrow$ 给定$\mathbf{x}$时，$\hat{b} = \dfrac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \sim N(b, \sigma^2 / s_{xx})$,

所以 $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \mid_{\mathbf{x}} \sim N(0,1)$, 该分布与$\mathbf{x}$的具体值无关，

所以无条件地$\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0,1)$。

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

$$= \sum_{i=1}^{n}\varepsilon_i{}^2 - n\bar{\varepsilon}^2 - \left(\sum_{i=1}^{n}\left(\frac{x_i - x}{\sqrt{s_{xx}}}\right)\varepsilon_i\right)^2 \hat{=} \|\boldsymbol{\varepsilon}\|^2 - (\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon})^2 - (\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon})^2$$

其中$\mathbf{u}^{\mathsf{T}} = (1/\sqrt{n},...,1/\sqrt{n}), \mathbf{v}^{\mathsf{T}} = \left((x_1 - \bar{x})/\sqrt{s_{xx}},...,(x_n - \bar{x})/\sqrt{s_{xx}}\right)$，模长1，$\mathbf{u} \perp \mathbf{v}$。

由第二讲引理2(2)即可得证。

(3). 由(1),(2)知：$\mathrm{A} = \dfrac{\sqrt{s_{xx}}\left(\hat{b} - b\right)}{\sigma} \sim N(0,1)$，$\mathrm{B} = \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$，

且两者独立，由t分布的定义$\dfrac{\mathrm{A}}{\sqrt{\mathrm{B}/(n-2)}} = \dfrac{\sqrt{s_{xx}}(\hat{b} - b)}{\hat{\sigma}} \sim t_{n-2}$。

Wald检验

原假设$H_0: b = b_0$，b_0已知,

Wald检验统计量定义为:

$$t = (\hat{b} - b_0)/se(\hat{b}) = \sqrt{s_{xx}}(\hat{b} - b_0)/\hat{\sigma},$$

由命题5知，$t \overset{H_0}{\sim} t_{n-2}$. $|t| \geq t_{n-2}(\alpha/2)$时拒绝原假设.

绝大数情况下我们只关心$H_0: b = 0$,

Wald检验统计量定义为

$$t = \hat{b}/se(\hat{b}) = \sqrt{s_{xx}}\hat{b}/\hat{\sigma} \overset{H_0}{\sim} t_{n-2}$$

检验准则(水平α): $|t| \geq t_{n-2}(\alpha/2)$时拒绝原假设。

Wald检验即相关性检验

命题6. $H_0: b = 0$的检验统计量$t = \sqrt{s_{xx}}\hat{b}/\hat{\sigma} = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}}$。

证：由 $\hat{b} = s_{xy}/s_{xx}$，$\hat{\sigma}^2 = \dfrac{1}{n-2}RSS = \dfrac{1}{n-2}\left(s_{yy} - s_{xy}^2/s_{xx}\right)$，

$$t = \frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{\dfrac{1}{n-2}\left(s_{yy} - s_{xy}^2/s_{xx}\right)}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}。$$

注：$H_0: b = 0$的Wald检验$t_1 = \dfrac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}}$，而第二讲中我们知道

$H_0: \rho = 0$的检验$t_2 = \sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}}$。

因为简单回归模型中斜率$b = \rho\dfrac{\sigma_y}{\sigma_x}$, $H_0: b = 0 \Leftrightarrow H_0: \rho = 0$

所以$t_1 = t_2$并不太令人意外。

两样本t检验

两样本t-检验是回归系数显著性检验的特殊情形

$$y_1,\ldots,y_{n_1} \ \text{iid} \sim N(\mu_1,\sigma^2) \quad \leftarrow x_1,\ldots,x_{n_1}=1$$
$$y_{n_1},\ldots,y_{n_1+n_2} \ \text{iid} \sim N(\mu_2,\sigma^2) \quad \leftarrow x_{n_1+1},\ldots,x_{n_1+n_2}=0$$

给两组样本分别赋予标号$x_i=0,1$,两样本问题写成线性模型：

$$y_i = a + bx_i + \varepsilon_i \quad (a=\mu_2, b=\mu_1-\mu_2), \quad \varepsilon_i \ \text{iid} \sim N(0,\sigma^2)$$
$$H_0: b=0 \Leftrightarrow \mu_1=\mu_2$$

容易验证该模型$H_0: b=0$的检验统计量$t=\dfrac{\hat{b}}{\sqrt{\hat{\sigma}^2/s_{xx}}}$

等于两样本t检验统计量$\dfrac{\bar{y}_1-\bar{y}_2}{\sqrt{(n_1^{-1}+n_2^{-1})s^2}}$

模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_1, ..., \varepsilon_n \ \text{iid} \sim N(0, \sigma^2)$

斜率的置信区间

由命题5, $\frac{\sqrt{s_{xx}}(\hat{b}-b)}{\hat{\sigma}} \sim t_{n-2}$，故$b$的$(1-\alpha)100\%$置信区间可以取为：

$$\left[\hat{b} \mp \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(\alpha/2)\right] = \left[b: \hat{b} - \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(\alpha/2), \ \hat{b} + \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(\alpha/2)\right]$$

均值函数的置信带

均值函数/回归函数$m(x_0) = E(y|x = x_0) = a + bx_0, x_0 \in R$。

对于给定的x_0，$m(x_0)$的LS估计$\hat{m}(x_0) = \hat{a} + \hat{b}x_0$，可以证明：

$$\hat{m}(x_0) \sim N\left(m(x_0), \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\right)$$

与$\hat{\sigma}^2$独立，类似于命题5可证

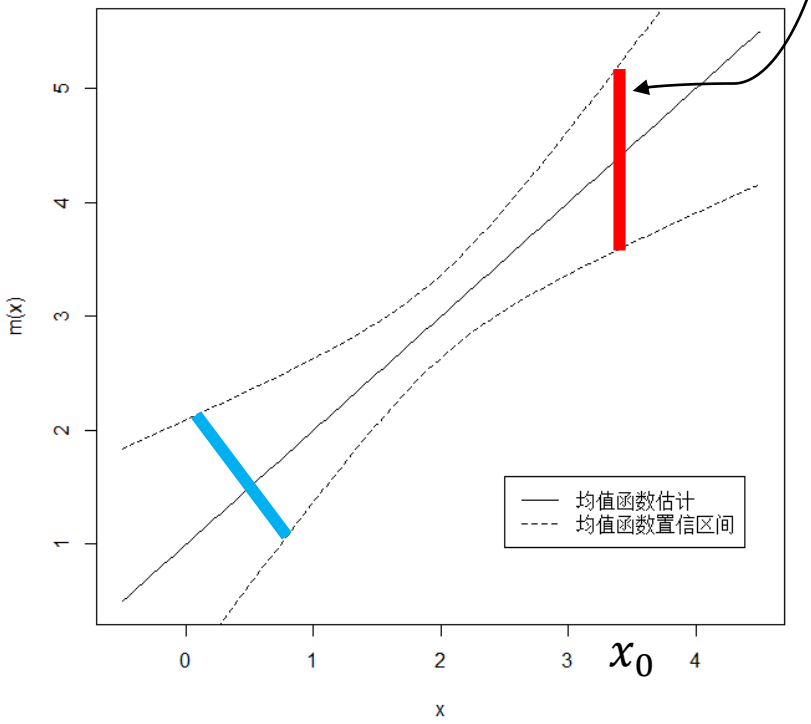
$$\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

基于该分布，我们得到$m(x_0)$的置信区间：

$m(x_0)$的置信水平$1-\alpha$的置信区间

$$\left[\hat{m}(x_0) \pm t_{n-2}(\alpha)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}}\right]$$

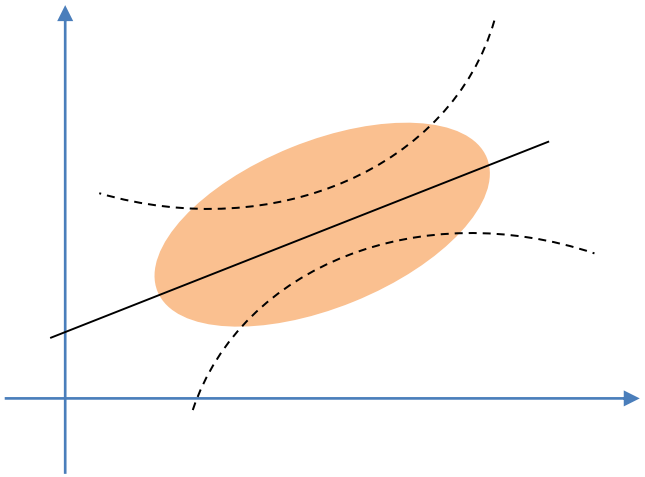
当x_0变化时，上述区间形成一个置信带



在y轴方向关于回归直线对称

在垂直于回归直线的方向并不对称。

练习：对于如下分布形状的数据，假设某人得到如下图所示的均值函数的估计（实线）及其置信带（虚线）。指出该图的两处错误。



错误的根源在于他对称地看待(x, y)，正确的方式是$y|x$