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

第七讲 简单线性回归模型（续）

2020.3.11

$$R^2 = \frac{VAR(\hat{y})}{VAR(y)}, \ VAR: 方差或样本方差$$

总体复相关系数平方

$$R^2_{\text{总体}} = \frac{\text{var}\left(\overset{\vee}{y}\right)}{\text{var}(y)} = \frac{\text{var}\left(E(y\,|\,x)\right)}{\text{var}(y)}$$

正交分解：$y = E(y\,|\,\mathbf{x}) + y - E(y\,|\,\mathbf{x})$

方差分解：$\text{var}(y) = \text{var}\left(E(y\,|\,\mathbf{x})\right) + E\left(\text{var}(y\,|\,\mathbf{x})\right)$

x解释y的比例：$R^2 = \text{var}\left(E(y\,|\,\mathbf{x})\right) / \text{var}(y)$

简单线性模型/不相关化(总体版本)：$E(y\,|\,x) = a + bx$

正交分解：$y = a + bx + \varepsilon$，【ε与x独立】

方差分解：$\sigma_y^2 = \text{var}(y) = \text{var}(a+bx) + \sigma^2 = b^2\sigma_x^2 + \sigma^2 = \rho^2\sigma_y^2 + \sigma^2$

$$R^2 = \frac{\text{var}(a+bx)}{\text{var}(y)} = \rho_{xy}^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}} = \frac{\sigma_{yx}\sigma_{xx}^{-1}\sigma_{xy}}{\sigma_{yy}} = \frac{\sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx}}{\sigma_{xx}}$$ (参见第4讲P6)

样本复相关系数平方

简单线性模型平方和分解(样本版本):

正交分解：$y_i = \hat{y}_i + e_i = \hat{a} + \hat{b}x_i + e_i,\ \ i = 1,\ldots,n$ 【$\sum \hat{y}_i e_i = 0$】

两边求平方和：$s_{yy} = SS_{总} = \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^{\ 2} = s_{\hat{y}\hat{y}} + s_{ee}$

$\hat{=} SS_{回} + \mathrm{RSS}$

注：样本方差$s_y^2 = s_{yy}/(n-1), s_{\hat{y}}^2 = s_{\hat{y}\hat{y}}/(n-1)$

由$\hat{b} = s_{xy}/s_{xx},\ \ \hat{a} = \bar{y} - \hat{b}\bar{x} \Rightarrow$

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

所以

$$SS_{回} = \hat{b}^2 s_{xx} = s_{xy}^2 / s_{xx} = r_{xy}^2 s_{yy}$$

$$RSS = s_{yy} - s_{xy}^2 s_{xx} = s_{yy} - s_{yx} s_{xx}^{-1} s_{xy} \widehat{=} s_{yy \bullet x} = (1 - r_{xy}^2) s_{yy}$$

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

$$R^2_{样本} = \frac{\hat{y}的样本方差}{y的样本方差}$$

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

误差方差估计

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$\varepsilon_i$的预测（不是估计）

σ^2的“LS”估计取为:

$$\hat{\sigma}^2 = \frac{1}{n-2} RSS = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$$

- 虽然$\hat{\sigma}^2$不是由最小二乘法（极小化误差平方和）直接得到的，但通常也称它为LS估计。

- 为什么RSS除以$n-2$? 样本量为n, 估计了两个参数。
  这样定义的$\hat{\sigma}^2$是σ^2的无偏估计。

- $\hat{\sigma}^2 = \dfrac{n-1}{n-2}(1-r_{xy}^2)\ s_y^2$【回忆误差方差$\sigma^2 = (1-\rho_{xy}^2)\sigma_y^2$】

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

其中$s_y^2 = s_{yy}/(n-1)$为$y_1,\ldots,y_n$的样本方差。

总结：简单回归模型的参数及其LS估计

模型(总体)：$y = a + bx + \varepsilon$,

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

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

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

参数	估计
(1) $b = \mathrm{cov}(x, y) / \mathrm{var}(x) = \rho\sigma_y / \sigma_x$ (2) $a = \mu_y - b\mu_x$, (3) $\sigma^2 = (1-\rho^2)\sigma_y^2$	(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$e_i = y_i - (\hat{a} + \hat{b}x_i)$ $(e_1, ..., e_n)^T \perp (x_1, ..., x_n)^T$
$R^2 = \dfrac{\mathrm{var}(a + bx)}{\mathrm{var}(y)} = \rho^2$	$R^2 = \dfrac{\text{'var'}(\hat{y})}{\text{'var'}(y)} = \dfrac{s_{\hat{y}\hat{y}}}{s_{yy}} = r^2$

最小二乘估计的性质

命题3（无偏性）. $E(\hat{b}) = b, E(\hat{a}) = a.$

证明: $\hat{b} = s_{xy} / s_{xx}$

$$E(\hat{b} \mid \mathbf{x}) = E\left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \,\middle|\, \mathbf{x} \right) = \frac{\sum (x_i - \bar{x}) E(y_i \mid x_i)}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(a + bx_i + E(\varepsilon_i \mid x_i))}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(bx_i)}{\sum (x_i - \bar{x})^2} = b \Rightarrow E(\hat{b}) = b$$

由ε与x独立, $E(\varepsilon_i \mid x_i) = E(\varepsilon_i) = 0$

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

注：无偏性需要条件$E(\boldsymbol{\varepsilon}) = 0$以及$\boldsymbol{\varepsilon}$与$\mathbf{x}$的独立性，但不需要方差齐性条件。

命题4（估计的方差和方差的估计）.

(1) 给定所有自变量**x**条件下, $\mathrm{var}(\hat{b}\,|\,\mathbf{x}) = \dfrac{\sigma^2}{s_{xx}}$.

(2) $E(\hat{\sigma}^2) = \sigma^2$.

证明：$\hat{b} = s_{xy} / s_{xx}$, 因为 $\mathrm{var}(y_i \,|\, x_i) = \sigma^2$,

$$\mathrm{var}(\hat{b}\,|\,\mathbf{x}) = \mathrm{var}\left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \,\middle|\, \mathbf{x} \right)$$

$$= \frac{\sum (x_i - \bar{x})^2 \,\mathrm{var}(y_i \,|\, x_i)}{\left[\sum (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{s_{xx}}$$

下面的计算都是在给定$x_1,...,x_n$的条件下：

(2) 下面证明$\mathrm{E(RSS)}=(n-2)\sigma^2$.因为$SS_{回}=\hat{b}^2 s_{xx}, RSS=s_{yy}-\hat{b}^2 s_{xx}$

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

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

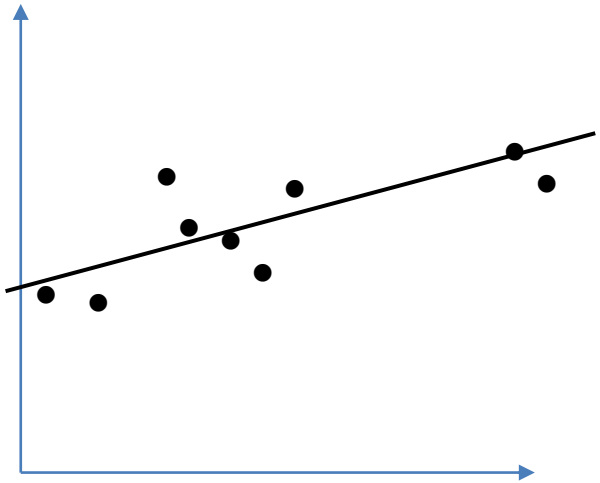
$$=\sum_{i=1}^{n}\left(\mathrm{var}(y_i)+(Ey_i)^2\right)-n(\mathrm{var}(\bar{y})-(E\bar{y})^2$$

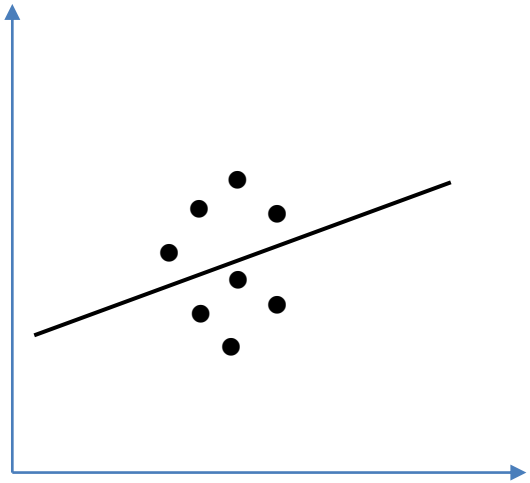
$$=n\sigma^2+\sum_{i=1}^{n}\left(a+bx_i\right)^2-n(\sigma^2/n)-n(a+b\bar{x})^2=(n-1)\sigma^2+b^2 s_{xx}$$

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

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

命题4表明，自变量的样本方差$s_x^2 = s_{xx}/(n-1)$越大，斜率估计的方差越小。





右图自变量方差比左图小，

回归直线不容易估计准确(方差大)。

LS估计的方差的估计

$$\mathrm{var}(\hat{b} \mid \mathbf{x}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

⇓ **“plug in”** 未知参数σ^2 的估计

$$\widehat{\mathrm{var}}(\hat{b} \mid \mathbf{x}) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

$$se(\hat{b}) = \widehat{sd}(\hat{b} \mid \mathbf{x}) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

命题5. (Gauss - Markov定理).
所有b的线性无偏估计中，LS估计的方差最小。

$$\mathbf{1}=(1,1,...,1)^{\mathsf{T}},$$
$$\mathbf{x}=(x_1,...,x_n)^{\mathsf{T}},$$
$$\mathbf{y}=(y_1,...,y_n)^{\mathsf{T}}$$
$$\mathbf{y}=\mathbf{1}a+\mathbf{x}b+\boldsymbol{\varepsilon}$$

证明：设$\mathbf{u}^{\mathsf{T}}\mathbf{y}$是$b$的任一线性无偏估计，即

$$b=\mathrm{E}(\mathbf{u}^{\mathsf{T}}\mathbf{y}\,|\,\mathbf{x})=\mathbf{u}^{\mathsf{T}}(\mathbf{1}a+\mathbf{x}b)=a\mathbf{u}^{\mathsf{T}}\mathbf{1}+b\mathbf{u}^{\mathsf{T}}\mathbf{x},$$

对任意a,b成立。 所以$\mathbf{u}^{\mathsf{T}}\mathbf{1}=0$，$\mathbf{u}^{\mathsf{T}}\mathbf{x}=1$

我们要证：$\mathrm{var}(\mathbf{u}^{\mathsf{T}}\mathbf{y}\,|\,\mathbf{x})=\sigma^2\mathbf{u}^{\mathsf{T}}\mathbf{u}\geq\ \mathrm{var}(\hat{b}\,|\,\mathbf{x})=\sigma^2/s_{xx}$,

即要证：$\mathbf{u}^{\mathsf{T}}\mathbf{u}\cdot s_{xx}\geq 1$

由Cauchy不等式：

$$\mathbf{u}^{\mathsf{T}}\mathbf{u}\cdot s_{xx}=\mathbf{u}^{\mathsf{T}}\mathbf{u}\cdot(\mathbf{x}-\mathbf{1}\bar{x})^{\mathsf{T}}(\mathbf{x}-\mathbf{1}\bar{x})\geq\left(\mathbf{u}^{\mathsf{T}}(\mathbf{x}-\mathbf{1}\bar{x})\right)^2=(\mathbf{u}^{\mathsf{T}}\mathbf{x})^2=1$$

注：注意到自变量的样本方差为$s_x^2 = s_{xx}/(n-1)$, 对任一(线性)无偏估计$\tilde{b}$，

$$\mathrm{var}(\tilde{b} \mid \mathbf{x}) \geq \sigma^2 / s_{xx},$$

$$\mathrm{var}(\tilde{b} \mid \mathbf{x}) \times s_x^2 \geq \frac{1}{n-1}\sigma^2$$

这说明b的估计和x的测量不可能同时非常精确

(类似于物理学的测不准原理/不确定性准则）。