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

第八讲 简单线性模型的统计推断

2020.3.13

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

预备：正态模型下的统计推断

例1. 假设$y_1,...,y_n$ iid $\sim (a,\sigma^2)$, 可表示为线性模型的形式

$$y_i = a+\varepsilon_i,\ \varepsilon_i,\ \ i=1,...,n\ \ iid \sim (0,\sigma^2)$$

是最简单的线性模型.

最小二乘: $\min \sum (y_i - a)^2$,

对a求导 $\Rightarrow \sum (y_i - a) = 0 \Rightarrow LS$ 估计 $\hat{a} = \bar{y}$(样本均值)

$RSS = \sum (y_i - \hat{a})^2 = \sum (y_i - \bar{y})^2$

$\hat{\sigma}^2 = RSS/(n-1) = s^2$(样本方差)

正态假设下，假设$y_1,...,y_n$ iid $\sim (a,\sigma^2)$,则

(1) $\bar{y} \sim N(a,\sigma^2/n)$, (2) $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, (3) $\bar{y}$与s^2独立.

(4) $\sqrt{n}(\bar{y}-a)/s \sim t_{n-1}$

证明(2)–(3):

$$(n-1)s^2=\sum(y_i-\bar{y})^2=\sum(y_i-a-(\bar{y}-a))^2=\sum\varepsilon_i{}^2-n\bar{\varepsilon}^2$$

记$\boldsymbol{\varepsilon}=(\varepsilon_1,...,\varepsilon_n)^{\mathsf{T}}\sim N(0,\sigma^2 I_n)$

令A为一个n阶正交矩阵, $A^{\mathsf{T}}A=AA^{\mathsf{T}}=I_n$, 其第一行为

$$\left(1/\sqrt{n},....,1/\sqrt{n}\right)$$

令$\mathbf{z}=A\boldsymbol{\varepsilon}$, 则$\mathbf{z}=A\boldsymbol{\varepsilon}\sim N(0,\sigma^2 I_n)$, 且$\|\mathbf{z}\|^2=\|\boldsymbol{\varepsilon}\|^2$,

$$z_1=(\varepsilon_1+...+\varepsilon_n)/\sqrt{n}=\sqrt{n}\bar{\varepsilon}.$$

$$(n-1)s^2=\sum_{i=1}^{n}\varepsilon_i{}^2-n\bar{\varepsilon}^2=\sum_{i=1}^{n}z_i{}^2-z_1^2=\sum_{i=2}^{n}z_i{}^2\sim\sigma^2\chi^2_{n-1},$$

s^2只与$z_2,...,z_n$有关, 而$\bar{y}=a+\bar{\varepsilon}$只与$z_1$有关, 故$\bar{y}$与$s^2$独立.

简单正态线性模型的统计推断

大多数软件假设模型中误差项服从正态分布，以后称之为正态线性模型.

我们仅考虑 b 的推断

命题6. 假设模型$y_i = a + bx_i + \varepsilon_i, \varepsilon_1, ..., \varepsilon_n \ \text{iid} \sim N(0, \sigma^2)$,则给定$\mathbf{x} = (x_1, ..., x_n)^{\mathsf{T}}$的条件下

(1) $\hat{b} \sim N(b, \sigma^2 / s_{xx})$

(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) $\dfrac{\hat{b} - b}{\sqrt{\text{var}(\hat{b})}} = \dfrac{\hat{b} - b}{\sqrt{\hat{\sigma}^2 / s_{xx}}} \sim t_{n-2}$

证明：(1) $y_i \mid x_i \sim N(a+bx_i, \sigma^2) \Rightarrow$ 给定$x_1,...,x_n$时，

$$\hat{b} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \sim N(b, \sigma^2 / s_{xx}).$$

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

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

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

$$= \| \boldsymbol{\varepsilon} \|^2 - (\mathbf{u}^{\mathsf{T}} \boldsymbol{\varepsilon})^2 - (\mathbf{v}^{\mathsf{T}} \boldsymbol{\varepsilon})^2,$$

其中$\mathbf{u}^{\mathsf{T}} = (1/\sqrt{n}, ..., 1/\sqrt{n}),$

$\mathbf{v}^{\mathsf{T}} = \left((x_1 - \bar{x}) / \sqrt{s_{xx}}, ..., (x_n - \bar{x}) / \sqrt{s_{xx}} \right)$

$$RSS = \|\boldsymbol{\varepsilon}\|^2 - (\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon})^2 - (\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon})^2$$

令$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I_n)$, 令$\mathbf{z} = A\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$,

其中A是一个正交矩阵, 其第一二行分别为$\mathbf{u}^{\mathsf{T}}, \mathbf{v}^{\mathsf{T}}$:

$$A = \begin{pmatrix} \mathbf{u}^{\mathsf{T}} \\ \mathbf{v}^{\mathsf{T}} \\ * \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{x_1 - \bar{x}}{\sqrt{s_{xx}}} & \frac{x_2 - \bar{x}}{\sqrt{s_{xx}}} & \cdots & \frac{x_n - \bar{x}}{\sqrt{s_{xx}}} \\ * & * & \cdots & * \end{pmatrix}$$

则 $z_1 = \mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon} = \sqrt{n}\bar{\varepsilon}, \quad z_2 = \mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon} = \sum_{i=1}^{n} (x_i - x)\varepsilon_i / \sqrt{s_{xx}},$

因为A是正交矩阵, $\|\mathbf{z}\|^2 = \|\boldsymbol{\varepsilon}\|^2$,

$$RSS = \|\boldsymbol{\varepsilon}\|^2 - (\mathbf{u}^{\mathsf{T}}\boldsymbol{\varepsilon})^2 - (\mathbf{v}^{\mathsf{T}}\boldsymbol{\varepsilon})^2 = \|\mathbf{z}\|^2 - z_1^{\,2} - z_2^{\,2} = \sum_{i=3}^{n} z_i^{\,2}$$

$$(n-2)\hat{\sigma}^2/\sigma^2 = RSS/\sigma^2 = \sum_{i=3}^{n} z_i^{\,2}/\sigma^2 \sim \chi^2_{n-2}\text{，且与}z_1, z_2\text{ 独立.}$$

而$\hat{a}, \hat{b}$仅与z_1, z_2 有关:

$$\hat{b} = b + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2} = b + z_2/\sqrt{s_{xx}},$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = a + b\bar{x} + \bar{\varepsilon} - \hat{b}\bar{x} = a + z_1/\sqrt{n} - z_2\bar{x}/\sqrt{s_{xx}},$$

所以$\hat{\sigma}^2$ 与$(\hat{a}, \hat{b})$独立。

验证：$RSS = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2 / s_{xx}$

$$(a)\ \ y_i - \bar{y} = a + bx_i + \varepsilon_i - (a + b\bar{x} + \bar{\varepsilon}) = b(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}),$$

$$(b)\ \text{由}\ \hat{a} = \bar{y} - \hat{b}\bar{x} \Rightarrow \hat{y}_i - \bar{y} = \hat{a} + \hat{b}x_i - \bar{y} = \hat{b}(x_i - \bar{x})$$

$$\text{而}\hat{b} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(a + bx_i + \varepsilon_i)}{\sum (x_i - \bar{x})^2} = b + s_{x\varepsilon} / s_{xx},$$

$$\Rightarrow \hat{y}_i - \bar{y} = \hat{b}(x_i - \bar{x}) = b(x_i - \bar{x}) + (x_i - \bar{x}) s_{x\varepsilon} / s_{xx}$$

$$\text{所以}RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y} - (\hat{y}_i - \bar{y})\right)^2$$

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

(3). 由 ⑴,(2)知：

$$\frac{\sqrt{s_{xx}}\left(\hat{b}-b\right)}{\sigma}\sim N(0,1),\quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim \chi^2_{n-2},$$

且两者独立，所以

$$\frac{\dfrac{\sqrt{s_{xx}}\left(\hat{b}-b\right)}{\sigma}}{\sqrt{\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}\Big/(n-2)}}=\frac{\sqrt{s_{xx}}\left(\hat{b}-b\right)}{\hat{\sigma}}\sim t_{n-2}$$

plug-in(代入)方差估计

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

代入σ^2的估计

Wald test

$$\frac{\hat{b}-b}{\sqrt{\widehat{\mathrm{var}}(\hat{b})}} = \frac{\hat{b}-b}{\hat{\sigma}/\sqrt{s_{xx}}} \sim t_{n-2}$$

置信区间

基于事实：$\dfrac{\sqrt{s_{xx}}\left(\hat{b}-b\right)}{\hat{\sigma}} \sim t_{n-2}$

$$1-\alpha = P\left(\left|\frac{\sqrt{s_{xx}}\left(\hat{b}-b\right)}{\hat{\sigma}}\right| \le t_{n-2}(\alpha/2)\right)$$

$$= P\left(\hat{b}-\frac{\hat{\sigma}}{\sqrt{s_{xx}}}t_{n-2}(\alpha/2) \le b \le \hat{b}-\frac{\hat{\sigma}}{\sqrt{s_{xx}}}t_{n-2}(\alpha/2)\right)$$

(1) b的$(1-\alpha)100\%$置信区间：$\left[\hat{b} \mp \dfrac{\hat{\sigma}}{\sqrt{s_{xx}}}t_{n-2}(\alpha/2)\right]$

(2) $H_0: b = b_0$ （b_0为已知常数，通常$b_0 = 0$）

t-检验统计量：$t = \dfrac{\sqrt{s_{xx}}(\hat{b}-b_0)}{\hat{\sigma}} \overset{H_0}{\sim} t_{n-2}$

回归系数的显著性 *t*-检验：

$$H_0: b = 0, \quad t = \frac{\hat{b}}{\sqrt{\hat{\sigma}^2 / s_{xx}}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2}$$

$|t| \geq t_{n-2}(\alpha/2)$时拒绝原假设。

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

$$t = \frac{\hat{b}}{\sqrt{\hat{\sigma}^2 / s_{xx}}} = \frac{s_{xy}/s_{xx}}{\sqrt{\frac{1}{n-2}\frac{s_{yy} - s_{xy}^2/s_{xx}}{s_{xx}}}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

两样本*t*-检验是回归系数显著性检验的特殊情形

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

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

应用线性模型：$y_i = a + bx_i + \varepsilon_i \quad (a=\mu_2, b=\mu_1-\mu_2)$

容易验证：$t = \dfrac{\hat{b}}{\sqrt{\hat{\sigma}^2 / s_{xx}}} = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{(n_1^{-1}+n_2^{-1})s^2}}$

非正态假设下的显著性检验

z-检验：

不假设误差服从正态分布的情形下，原假设$H_0: b = 0$成立时，

$$t = \frac{\hat{b}}{\sqrt{\hat{\sigma}^2 / s_{xx}}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \approx z = \sqrt{n}r \xrightarrow{\mathrm{d}} N(0,1),\ n \to \infty$$

$|t| \geq z_{\alpha/2}$，或$|z| \geq z_{\alpha/2}$时否定原假设。

卡方检验：

卡方检验统计量$z^2 = nr^2$在原假设H_0： $b = 0$下近似服从χ_1^2， $n \to \infty$.

当$z^2 \geq \chi_1^2(\alpha)$时，在$\alpha$水平下拒绝原假设.

结果的解释: 因果还是关联?

$y = a + bx + \varepsilon$， x 与 ε 是否独立？

- **随机化控制试验或天然试验，** x 与 ε 独立

  结论：因果， ***x*** **每增加一个单位，** ***y*** **的期望增加** ***b*** **个单位。** **Key: 同一个对象。**

- **观察研究, 自变量与误差几乎不可能独立，**

  结论：关联， 如果一个研究对象的 ***x*** **比另外一个大1个单位，则其** ***y*** **平均大** ***b*** **个单位。Key:不同的对象。**

例如，分析2001年人口抽样调查数据，得到妻子教育水平（上学的年数）与丈夫教育水平的回归方程如下：

$$\text{WifeEdLevel} = 5.60 + 0.57 \times \text{HusbandEdLevel} + \text{residual}$$

如果公司送王先生到大学在职培养一年，你是否预期王太太的教育水平会上升0.57年？若不是，0.57的含义是什么？

这是观察研究而非试验，结果是关联而不是因果：

b=0.57的含义是：如果该研究中某人 比另外一个人 多上一年学，那么这个人的妻子的比另外一人的妻子多上（平均意义上）0.57年学。