2014.3.20

第十讲 简单线性回归模型的统计推断

最小二乘估计的性质

性质1（无偏性）. $E(\hat{b})=b, E(\hat{a})=a, E(\hat{\sigma}^2\mid\mathbf{x})=\sigma^2$

证明:(3) $\hat{b}=s_{xy}/s_{xx}=\dfrac{(\mathbf{x}-\mathbf{1}\bar{x})'\mathbf{y}}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}$,

$$\hat{\mathbf{y}}=\mathbf{1}\hat{a}+\mathbf{x}\hat{b}=\left(\frac{\mathbf{11}'}{n}+\frac{(\mathbf{x}-\bar{x}\mathbf{1})(\mathbf{x}-\mathbf{1}\bar{x})'}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}\right)\mathbf{y}$$

$$\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}=\left(I_n-\frac{\mathbf{11}'}{n}-\frac{(\mathbf{x}-\bar{x}\mathbf{1})(\mathbf{x}-\mathbf{1}\bar{x})'}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}\right)\mathbf{y}\overset{\text{记为}}{=}C\mathbf{y}$$

容易验证C是对称幂等矩阵，所以$(I-C)$也是，而且：

$\mathrm{tr}(C)=n-2$，$C\mathbf{1}=0$，$C\mathbf{x}=0$

则给定$\mathbf{x}$的条件下，

$$\mathrm{E}(\mathrm{RSS}\mid\mathbf{x})=E(\mathbf{e}'\mathbf{e}\mid\mathbf{x})=E(\mathbf{y}'C\mathbf{y}\mid\mathbf{x})$$
$$=(\mathbf{1}a+\mathbf{x}b)'C(\mathbf{1}a+\mathbf{x}b)+tr(C)\sigma^2$$
$$=(n-2)\sigma^2$$

所以$E(\hat{\sigma}^2\mid\mathbf{x})=\mathrm{E}(\mathrm{RSS}/(n-2)\mid\mathbf{x})=\sigma^2$。

注：事实上，$\mathrm{P}=I-C$是$\mathrm{L}(\mathbf{1},\mathbf{x})$对应的投影矩阵。

$\hat{\mathbf{y}}=\mathrm{P}\mathbf{y}$是$\mathbf{y}$在$\mathrm{L}(\mathbf{1},\mathbf{x})$上的投影。

注：LS与投影

最小二乘法极小化：

$$\min\sum_{i=1}^{n}(y_i-a-bx_i)^2=\min\|\mathbf{y}-\mathbf{1}a-\mathbf{x}b\|^2=\min_{\mathbf{u}\in L(\mathbf{1},\mathbf{x})}\|\mathbf{y}-\mathbf{u}\|^2$$

记$\mathbf{x}^{\perp}=\mathbf{x}-\mathrm{P}_1\mathbf{x}=\mathbf{x}-\mathbf{1}\bar{x}$

最优解为$\mathbf{u}=\hat{\mathbf{y}}\overset{\text{记}}{=}\mathrm{P}_{(\mathbf{1},\mathbf{x})}\mathbf{y}=(\mathrm{P}_1+\mathrm{P}_{\mathbf{x}^{\perp}})\mathbf{y}=\left(\dfrac{\mathbf{11}'}{n}+\dfrac{(\mathbf{x}-\mathbf{1}\bar{x})(\mathbf{x}-\mathbf{1}\bar{x})'}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}\right)\mathbf{y}$

$$=\mathbf{1}\frac{\mathbf{1}'\mathbf{y}}{n}+(\mathbf{x}-\mathbf{1}\bar{x})\frac{(\mathbf{x}-\mathbf{1}\bar{x})'\mathbf{y}}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}$$

$$=\mathbf{1}\left(\frac{\mathbf{1}'\mathbf{y}}{n}-\bar{x}\frac{(\mathbf{x}-\mathbf{1}\bar{x})'\mathbf{y}}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}\right)+\mathbf{x}\frac{(\mathbf{x}-\mathbf{1}\bar{x})'\mathbf{y}}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})}=\mathbf{1}\hat{a}+\mathbf{x}\hat{b}$$

其中$\hat{b}=\dfrac{(\mathbf{x}-\mathbf{1}\bar{x})'\mathbf{y}}{(\mathbf{x}-\mathbf{1}\bar{x})'(\mathbf{x}-\mathbf{1}\bar{x})},\hat{a}=\bar{y}-\bar{x}\hat{b}$

性质2(估计的方差). 给定x条件下,

$$\mathrm{cov}\left(\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix}\Bigg|_{\mathbf{x}}\right)=\begin{pmatrix}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)\sigma^2 & -\frac{\bar{x}}{s_{xx}}\sigma^2\\ -\frac{\bar{x}}{s_{xx}}\sigma^2 & \boxed{\frac{\sigma^2}{s_{xx}}}\end{pmatrix}$$

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

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

$$=\frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$,等等。

LS估计的方差的估计

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

$\Downarrow$ **"plug in"** 未知参数 σ^2 的估计

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

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

线性估计 : $\mathbf{y}$的线性组合

性质3(Gauss - Markov定理的特殊情况).
所有b的线性无偏估计中，LS估计的方差最小。

证明：设$u'\mathbf{y}$是b的任一线性无偏估计，

无偏性即 $b=\mathrm{E}(u'\mathbf{y}\,|\,\mathbf{x})=u'(\mathbf{1}a+\mathbf{x}b)=au'\mathbf{1}+bu'\mathbf{x}$,

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

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

由$1=(u'\mathbf{x})^2=(u'(\mathbf{x}-\mathbf{1}\bar{x}))^2\le u'u\cdot s_{xx}\Rightarrow u'u\ge 1/s_{xx}$.

正态模型下的统计推断

定理. 假设模型$y_i=a+bx_i+\varepsilon_i,\varepsilon_1,\ldots,\varepsilon_n$ iid $\sim \mathrm{N}(0,\sigma^2)$,则
给定$\mathbf{x}=(x_1,\ldots,x_n)'$的条件下

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

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

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

证明：(1) 正态变量的线性组合仍为正态

(2) 因为 $(n-2)\hat{\sigma}^2 = RSS = \|\mathbf{e}\|^2 = \mathbf{e}'\mathbf{e}$, 我们已知：

$$\mathbf{e} = \left(I_n - \frac{\mathbf{11}'}{n} - \frac{(\mathbf{x}-\bar{x}\mathbf{1})(\mathbf{x}-\bar{x}\mathbf{1})'}{(\mathbf{x}-\bar{x}\mathbf{1})'(\mathbf{x}-\bar{x}\mathbf{1})} \right)\mathbf{y} \ \hat{=}\ \mathrm{C}\mathbf{y}$$

其中C对称幂等, $\mathrm{trC} = \mathrm{n} - 2$, 及 $\mathrm{C}\mathbf{1} = 0, \mathrm{C}\mathbf{x} = 0$

$$\text{所以}\ \frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\mathbf{y}'\mathrm{C}\mathbf{y}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\mathrm{C}\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2_{n-2} \quad (\because \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n))$$

由于 $C\mathbf{x} = 0, C\mathbf{1} = 0$, 所以 $\hat{\sigma}^2$ 与 $(\hat{a}, \hat{b})$ 独立.

(3) 由 (1), (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

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

总结Plug-in: Normal → t

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

$\Downarrow$ "plug in" σ^2 的估计(df=n-2)

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

统计推断

基于事实：$\dfrac{\sqrt{s_{xx}}\left(\hat{b} - b\right)}{\hat{\sigma}} \sim t_{n-2}$，可构建回归系数b的$(1-\alpha)100\%$置信区间：

$$\left[\hat{b} - \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(1-\alpha/2),\ \hat{b} + \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(1-\alpha/2) \right]$$

$H_0 : b = 0$

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

$H_0 : b = b_0$ (b_0为已知常数)

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

一些注解

- 回归系数的显著性检验等于相关检验

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

- 两样本**t-**检验是回归分析的特殊情形

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

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

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

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

- 结果的解释：因果还是关联？

$$y=a+bx+\varepsilon, \quad x\text{与}\varepsilon\text{是否独立？}$$

- 对于随机化控制试验或天然试验，结果为因果关系：
x 每增加一个单位，***y*** 的期望增加 ***b*** 个单位。

- 对于观察研究，回归分析结果很可能只是关联关系：
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

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：

方程中的**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，那么平均意义上第一人的妻子的比另外一人能的妻子多上**0.57**年学。

例1. 胡克定律（试验研究）

建立模型：

$$L_i=a+bW_i+\varepsilon_i, \quad \varepsilon_i\sim(0,\sigma^2)\text{与}W_i\text{独立}$$

其中a代表弹簧原长度，b为弹性系数。

W Weight (kg)	L Length (cm)
0	439.00
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

$$\hat{a}=439.01cm, \quad \hat{b}=0.049cm/kg$$

$$\hat{\sigma}=0.0084$$

$$\text{t}=48.98, p\text{值}=1.04\times10^{-6}\text{，高度显著。}$$

结论：物体每增加**1kg**，弹簧伸长**0.049**厘米

（考虑到这是一个随机化试验，结论是因果关系）

R输出结果

例如W系数的估计$\hat{b}=0.004914$，其标准差$sd(\hat{b})=0.0001003$

$$t=\frac{\hat{b}}{sd(\hat{b})}=\frac{0.004914}{0.0001003}=49.98$$

Coefficients: $\hat{\beta}$ $sd(\hat{\beta})$ $t=\hat{\beta}/sd(\hat{\beta})$ p值

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.390e+02	6.076e-03	72254.92	< 2e-16 ***
W	4.914e-02	1.003e-03	48.98	1.04e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008395 on 4 degrees of freedom
Multiple R-squared: 0.9983, Adjusted R-squared: 0.9979
F-statistic: 2399 on 1 and 4 DF, p-value: 1.04e-06

误差标准差估计：$\hat{\sigma}=0.008395, df=n-2=4$

复相关系数平方：$R^2=SS_{回}/SS_{总}=(样本相关系数r)^2=0.9983$

F-statistic：这里 $=t^2$

多数情况下，简单线性模型的**GM**假设一般不满足，下面我们举例说明观察研究情况下模型用来

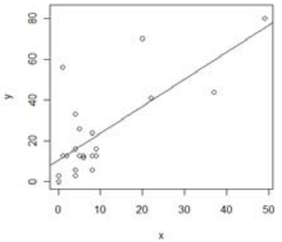
- 预测
- 描述变量关系， 发现关联规律

例2. (奥运主办国金牌预测). 为了预测北京奥运会中国金牌数目，首先需要确定预测变量(自变量)。

上一届奥运会的金牌数**(x)** 作为预测变量。

悉尼奥运会中国金牌数**32.**

历史数据：主办国金牌数**(y)**及其上届金牌数**(x)**。



拟合历史数据得回归方 程：$y=10.6+1.2x$

上一届中国金牌数 $x=32$，预测本届

$\hat{y}=10.6+1.2\times 32=52.$

主办国(y)	上届(x)
26	5
70	20
56	1
24	8
13	2
13	9
6	4
41	22
33	4
3	4
6	8
13	6
13	8
16	4
3	0
13	5
0	0
80	49
83	NA
12	6
13	1
44	37
16	9

例3. (标准化) 体重指数的定义 – 通过发现体重与身高的关系，消除身高的影响之后定义一个标准化的指标

如何计算一个体重指标，刻画体重是否正常或不合标准？

最简单的做法是将重量**W**作为指数：
假设**W**或者**log(W)** 服从正态分布**N(**μ，σ **2),** 重量处于群体**95%**置信区间定义为正常（比如：若**log(W)>**μ**+1.645**σ，定义为偏胖（体重超过**95%**的人）

但显然，不同身高、性别、年龄的人不具可比性。

即μ是若干因素的函数。

指数除了标准化之外，还应该具有普适性, 即需要考虑到上式中 影响 μ 的其它因素, 比 如身高 **H**。假设成年男性的体重服从

假设身高H的成年男性的体重服从

$$\log(W) \sim N\left(a + b\log(H),\ \sigma^2\right)$$

$$\log(W) = a + b\log(H) + \varepsilon,\ \varepsilon \sim N(0,\sigma^2)$$

$$标准化:\ z = \frac{\log(W) - (a + b\log(H))}{\sigma} \sim N(0,1)$$

例如，若$z > 1.645 \Leftrightarrow \frac{W}{H^b} > e^{a+1.645\sigma}$，定于为偏胖。

经验数据表明 $b \approx 2$

Body Mass Index : $BMI = \frac{W}{H^2}$ 单位kg/m^2

Category	BMI range – kg/m²	BMI Prime	Mass (weight) of a 1.8 metres (5 ft 11 in) person with this BMI.
Severely underweight	less than 16.0	less than 0.66	less than 51.8 kilograms (8.16 st; 114 lb)
Underweight	from 16.0 to 18.5	from 0.66 to 0.73	between 51.8 and 59.9 kilograms (8.16 and 9.43 st; 114 and 132 lb)
Normal	from 18.5 to 25	from 0.74 to 0.99	between 59.9 and 81.0 kilograms (9.43 and 12.76 st; 132 and 179 lb)
Overweight	from 25 to 30	from 1.0 to 1.19	between 81.0 and 97.2 kilograms (12.76 and 15.31 st; 179 and 214 lb)
Obese Class I	from 30 to 35	from 1.2 to 1.39	between 97.2 and 113.4 kilograms (15.31 and 17.86 st; 214 and 250 lb)
Obese Class II	from 35 to 40	from 1.4 to 1.59	between 113.4 and 129.6 kilograms (17.86 and 20.41 st; 250 and 286 lb)
Obese Class III	over 40	over 1.6	from 129.6 kilograms (20.41 st; 286 lb)

from wiki