《线性回归》
—线性回归(9)

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.04.09

主要内容：线性回归模型中的假设检验

1. 线性回归模型中的假设检验
   - 引论
   - 似然比检验
   - F-检验
   - 例子
   - 拟合优度检验
   - $F-$检验: 模型显著性检验

引论

这一章主要研究检验线性模型的线性假设的方法。

- 线性模型

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon$$

中的假设可以表示为下面的一般形式：

$$H_0 : \mathbf{A}\theta = \mathbf{c},$$

其中$\mathbf{A}$是$q \times p$的已知的矩阵，$\mathbf{c}$是已知的$q \times 1$的向量。
不同$\mathbf{A}$和$\mathbf{c}$的选择对应不同的假设：

- ✓ $\mathbf{A} = I_p$, $\mathbf{c} = \mathbf{0}$.
- ✓ $\mathbf{A} = (0, 0, \cdots, 0, 1, 0, \cdots, 0)$ [第j个位置为1，其余元素都是0, $\mathbf{c} = c_0$ 或者0]
- ✓ $\mathbf{A} = [\mathbf{0}_{r\times(p-r)}\ \mathbf{I}_{r\times r}]$, $\mathbf{c} = \mathbf{0}$

似然比检验

- 考虑线性模型

$$G : \mathbf{Y} = \mathbf{X}\theta + \epsilon, \epsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

其中$\mathbf{X}$是秩为p的$n \times p$的设计矩阵。

- 欲检验假设：

$$H_0 : \mathbf{A}\theta = \mathbf{c}, \text{其中}\mathbf{A}\text{是秩为}q\text{的}q \times p\text{矩阵}.$$

- 模型G的似然函数为：

$$L(\theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\theta\|^2\right\}.$$

似然比检验(续)

- θ和σ^2的MLE是：

$$\begin{aligned}\hat{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y,\\ \widehat{\sigma}^2 &= n^{-1}\|\mathbf{Y}-\mathbf{X}\hat{\theta}\|^2.\end{aligned}$$

- 故似然函数的最大值是：

$$L(\hat{\theta},\widehat{\sigma}^2) = (2\pi\widehat{\sigma}^2)^{-n/2}\exp(-n/2).$$

- 下一步是在H_0之下求出似然函数的最大值。即相当于在约束条件$\mathbf{A}\theta=\mathbf{c}$之下求似然函数的最大值。

似然比检验(续)

- 可用Lagrange乘子法实现. 定义

$$\begin{aligned} r(\theta, \sigma^2, \lambda) &= \log L(\theta, \sigma^2) + (\theta^T \mathbf{A}^T - \mathbf{c}^T)\lambda \\ &= C - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\theta\|^2 + (\theta^T \mathbf{A}^T - \mathbf{c}^T)\lambda. \end{aligned}$$

- 用常规方法可以求得最大值点：$\hat{\theta}_H$和$\widehat{\sigma}^2_H = \|\mathbf{Y} - \mathbf{X}\hat{\theta}_H\|^2/n$, 且最大值为：

$$L(\hat{\theta}_H, \widehat{\sigma}^2_H) = (2\pi\widehat{\sigma}^2_H)^{-n/2}\exp(-n/2). \tag{1}$$

- 由此可得检验假设H_0的似然比统计量：

$$\Lambda = \frac{\sup_{\mathbf{A}\theta=\mathbf{c},\sigma^2} L(\theta, \sigma^2)}{\sup_{\theta,\sigma^2} L(\theta, \sigma^2)} = \frac{L(\hat{\theta}_H, \widehat{\sigma}^2_H)}{L(\hat{\theta}, \widehat{\sigma}^2)} = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_H}\right)^{n/2}. \tag{2}$$

似然比检验(续)

- 根据假设检验的似然比原理，在Λ的值很小的时拒绝H_0.
- 要做假设检验，还需要知道λ的分布。
- 可以证明，在H_0成立时，

$$F := \frac{n-p}{q}\left(\Lambda^{-n/2} - 1\right) \sim F_{q,n-p}.$$

- 因此，当F的值很大时，拒绝H_0. 利用这个结果还计算p值。

F-检验：动机

- 如果要检验假设$H_0 : \mathbf{A}\theta = \mathbf{c}$，一个很自然的检验统计量是：$\mathbf{A}\hat{\theta} - \mathbf{c}$, 如果这个量的绝对值很大，则拒绝原假设$H_0$.
- 由于$\mathbf{A}\hat{\theta}$中的元素的精度[量纲]各不相同，不应该把它们看作是一致的。为考虑各个量的精度，我们可以以适当的方式将$\hat{\theta}$的精度加在检验统计量中, 例如，考虑二次型：

$$(\mathbf{A}\hat{\theta} - \mathbf{c})^T \left(\mathsf{Var}[\mathbf{A}\hat{\theta}] \right)^{-1} (\mathbf{A}\hat{\theta} - \mathbf{c}),$$

  其中，$\mathsf{Var}[\mathbf{A}\hat{\theta}] = \sigma^2 \mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$.
- 由于σ^2常未知，可用$S^2 = \mathsf{SSE}/(n-p)$代替之，得：

$$(\mathbf{A}\hat{\theta} - \mathbf{c})^T \left[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T \right]^{-1} (\mathbf{A}\hat{\theta} - \mathbf{c})/S^2.$$

- 注意：在Seber and Lee （2003）中, 用RSS表示残差平方和，但在本ppt中用SSE表示之。从上下文判断其含义。

主要结论：

- 一些记号：

$$\begin{aligned}\text{SSE} &= \|\mathbf{Y}-\mathbf{X}\hat{\theta}\|^2 = \|\mathbf{Y}-\widehat{\mathbf{Y}}\|^2 = (n-p)S^2,\\ \text{SSE}_H &= \|\mathbf{Y}-\mathbf{X}\hat{\theta}_H\|^2 = \|\mathbf{Y}-\widehat{\mathbf{Y}}_H\|^2.\end{aligned}$$

- $\hat{\theta}$和$\hat{\theta}_H$之间的关系：

$$\hat{\theta}_H = \hat{\theta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{c}-\mathbf{A}\hat{\theta}), \qquad (3)$$

  其中SSE$_H$是在约束条件$\mathbf{A}\theta = \mathbf{c}$之下$\|\mathbf{Y}-\mathbf{X}\theta\|^2$的最小值。

- 检验假设H_0的F统计量如下：

Theorem

(i)

$$\begin{aligned} SSE_H - SSE &= \|\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_H\|^2 \\ &= (\mathbf{A}\hat{\theta} - \mathbf{c})^T [\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1} (\mathbf{A}\hat{\theta} - \mathbf{c}). \end{aligned}$$

(ii)

$$\begin{aligned} & \mathbf{E}[SSE_H - SSE] \\ = \ & \sigma^2 q + (\mathbf{A}\theta - \mathbf{c})^T [\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1} (\mathbf{A}\theta - \mathbf{c}) \\ = \ & \sigma^2 q + [SSE_H - SSE]_{\mathbf{Y}=\mathbf{E}[\mathbf{Y}]}. \end{aligned}$$

(iii) 当H_0成立时，

$$
\begin{aligned}
F &= \frac{(\mathrm{SSE}_H - \mathrm{SSE})/q}{\mathrm{SSE}/(n-p)} \\
&= \frac{(\mathbf{A}\hat{\theta} - \mathbf{c})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\theta} - \mathbf{c})}{qS^2}
\end{aligned}
$$

服从$F_{q,n-p}$分布。

(iv) 当$\mathbf{c} = \mathbf{0}$, F可以表示为：

$$
F = \frac{n-p}{q} \cdot \frac{\mathbf{Y}^T(\mathbf{P} - \mathbf{P}_H)\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}}
$$

其中$\mathbf{P}_H$是对称的幂等矩阵，$\mathbf{P}_H\mathbf{P} = \mathbf{P}\mathbf{P}_H = \mathbf{P}_H$.

详细的证明见Seber and Lee (2003), Theorem 4.1 p. 100-102.

对于简单线性模型，上面的检验结果有比较简单的表达式。

简单线性模型中的检验

- 考虑简单线性模型
$$\mathbf{Y}_i = \theta_0 + \theta_1 x_i + \epsilon_i, (i = 1, \cdots, n).$$
- 我们欲检验假设:
$$H_0 : \theta_1 = c.$$
- $\mathbf{X} = (\mathbf{1}_n, \mathbf{x})_{n\times 2}$. $\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}$,
$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$,
$\mathbf{X}^T Y = \begin{pmatrix} \sum_{i=1}^n \mathbf{Y}_i \\ \sum_{i=1}^n x_i\mathbf{Y}_i \end{pmatrix}$.

- $\hat{\theta}$的显式表达式：$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$,

$$
\begin{aligned}
\hat{\theta}_0 &= \overline{\mathbf{Y}} - \hat{\theta}_1 \bar{x}, \\
\hat{\theta}_1 &= \frac{\sum_{i=1}^n \mathbf{Y}_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (\mathbf{Y}_i - \overline{\mathbf{Y}})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\
\widehat{Y}_i &= \hat{\theta}_0 + \hat{\theta}_1 x_i = \overline{\mathbf{Y}} + \hat{\theta}_1 (x_i - \bar{x}).
\end{aligned}
$$

- 假设$H_0 : \theta_1 = c$的检验统计量为：

$$
F = \frac{(\hat{\theta}_1 - \mathbf{c})^2}{S^2 / \sum_{i=1}^n (x_i - \bar{x})^2}. \tag{4}
$$

- 尝试推导上面检验统计量的分布。

- 如果$(\mathbf{X}, \mathbf{Y})$服从二元正态分布

$$N_2\left(\left(\begin{array}{c}\mu_{\mathbf{X}}\\ \mu_{\mathbf{Y}}\end{array}\right),\left(\begin{array}{cc}\sigma_{\mathbf{X}}^2 & \rho\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}\\ \rho\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}} & \sigma_{\mathbf{X}}^2\end{array}\right)\right),$$

  则有

$$\mathbf{E}[\mathbf{Y}|\mathbf{X}=x]=\mu_{\mathbf{Y}}+\rho\frac{\sigma_{\mathbf{Y}}}{\sigma_{\mathbf{X}}}(x-\mu_{\mathbf{X}})=\theta_0+\theta_1 x.$$

- 在一定的条件之下，$\mathbf{E}[\mathbf{Y}|\mathbf{X}=x]=\theta_0+\theta_1 x$关于$x$的线性假设是成立的。并非没有道理！
- 试将上面的结果推广高维多元正态分布的情形。

拟合优度检验(Goodness-of-fit Test)

- 参见Seber and Lee (2003), p. 115-116

$F-$检验：模型显著性性检验

- 主要内容详见Sober and Lee (2003), p. 110–113
- 对于线性模型

$$\mathbf{Y}_i = \theta_0 + x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + \epsilon_i, i = 1, \cdots, n, \epsilon_i \sim \ iid\, N(0, \sigma^2).$$

- 现在欲检验假设：

$$H_0 : \theta_1 = \cdots = \theta_p = 0.$$

这个零假设要检验的是所有的协变量是否对响应变量有影响。如零假设被拒绝，则在$x_1, \cdots, x_p$中至少有一个协变量对$\mathbf{Y}$有影响，到底是哪一个，还要做进一步的检验。如果另假设没有被拒绝，则没有证据说$x_1, \cdots, x_p$对$\mathbf{Y}$ 有影响。
- 上面的假设亦称为模型显著性假设。

模型显著性检验

- 模型的显著性检验相当于考察下面的全模型(model F)和在H_0成立的情况下的模型(model H)是否有差别：

$$\begin{aligned}\text{Model F: } \mathbf{Y}_i &= \theta_0 + x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + \epsilon_i, \\ & i = 1, \cdots, n, \epsilon_i \sim \ iid\ N(0, \sigma^2).\end{aligned}$$

$$\text{Model H: } \mathbf{Y}_i = \theta_0 + \epsilon_i, i = 1, \cdots, n, \epsilon_i \sim \ iid\ N(0, \sigma^2).$$

- 如何度量Model F和Model H的差别呢？一种办法是以前讲过的似然比检验方法。

模型显著性检验

- 另外一种想法是考察两种模型之下的残差平方和. 用RSS和RSS_H分别表示Model F 和Model H之下的残差平方和，即

$$\begin{aligned}\mathrm{RSS} &= \|\mathbf{Y}-\mathbf{X}\hat{\theta}\|^2 \quad (\text{自由度}: n-(p+1))\\ \mathrm{RSS}_H &= \|\mathbf{Y}-\overline{Y}\|^2 \quad (\text{自由度}: n-1).\end{aligned}$$

- 对于这两个残差平方和总有：

$$\mathrm{RSS} \le \mathrm{RSS}_H.$$

- 直观的想法就是，如果$\mathrm{RSS}_H-\mathrm{RSS}$取值较大，则拒绝$H_0$.

- 根据模型的假设以及残差RSS和RSS_H的性质，我们构造如下的检验统计量：

$$\begin{aligned} F &= \frac{(\mathrm{RSS}_H - \mathrm{RSS})/([n-1]-[n-(p+1)])}{\mathrm{RSS}/(n-p-1)} \\ &= \frac{(\mathrm{RSS}_H - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}. \end{aligned}$$

- 可以证明：

$$F \sim F_{p,n-p-1}.$$

- 如果$F \geq F^{\alpha}_{p,n-p-1}$ ($F^{\alpha}_{p,n-p-1}$是$F_{p,n-p-1}$分布的上α分位点)，则拒绝零假设H_0. 拒绝零假设的话，我们可以说回归时显著的，x_{ij}的作用不能被完全忽视。拒绝零假设并不意味着拟合方程$\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$必然是合理的。

R：lm中各个量的计算及其原理

- 至此，R中summary(lm(y∼x1+...+xp))中所有的量都可以计算了！
- R 文件F-test-LM.r 给出了每一个量的计算过程。【电脑演示！】
- 要求掌握每个量的计算原理。
- anova()常和lm()一起使用。

作业：

(1) 阅读Seber and Lee (2003)的第四章之后，完成p.113中的Ex 4c: 1,2,3.

(2) 试将第14页中的结果推广高维多元正态分布的情形。

【第十三讲】