多元线性回归的诊断

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

1.1 回归系数的假设检验

一元多重线性回归模型

对一元多重正态线性回归模型

$$Y_{n\times 1} = X_{n\times p}\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$$

其中 $Rank(X) = p$. 由于

1. β 的最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'Y \sim N_p(\beta, \sigma^2(X'X)^{-1})$, σ^2 的估计 $s^2 = \frac{1}{n-p}\hat{\epsilon}'\epsilon$ 且 $(n-p)s^2 \sim \sigma^2\chi^2_{n-p}$
2. s^2 和 $\hat{\beta}$ 相互独立.

从而

- β 的 $1-\alpha$ 置信域为

$$(\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \leq ps^2F_{p,n-p}(\alpha)$$

- 利用 Cauchy-Schwar 不等式可得 $c'\beta$ 的同时 $1-\alpha$ 置信区间 (Scheffe' confidence intervals) 为

$$c'\hat{\beta} \pm \sqrt{s^2 c'(X'X)^{-1}c}\sqrt{pF_{p,n-p}(\alpha)}$$

- 注意到 $\hat{\beta}_i \sim N_1(\beta_i, \sigma^2(X'X)^{-1}_{ii})$, 从而 $Var(\hat{\beta}_i)$ 的估计为 $\widehat{Var}(\hat{\beta}_i) = s^2(X'X)^{-1}_{ii}$, 因此

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{Var}(\hat{\beta}_i)}} \sim t_{n-p}$$

从而可得边际置信区间

$$\hat{\beta}_i \pm \sqrt{\widehat{Var}(\hat{\beta}_i)}t_{n-p}(\alpha/2)$$

对一元多重线性回归模型 $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$, 考虑假设 $H_0 : \beta_2 = 0 = L\beta$, 其中 β_2 为 $q \times 1$ 向量, $L_{n\times p} = (0, I_q)$. 在正态假

设下, 似然比检验可以使用. 也可以利用顺序平方 (sequential sum of squres):

$$\begin{aligned} SST &= SS_e + SS_{reg} \\ &= SS_e + SS_{reg}(X_1) + SS_{reg}(X_2|X_1) \end{aligned}$$

上式最后一项为

$$\begin{aligned} SS_{reg}(X_2|X_1) &= SS_{reg}(X_1, X_2) - SS_{reg}(X_1) \\ &= SS_e(X_1) - SS_e(X_1, X_2) \\ &= \|Y - X_1\hat{\beta}_1\|^2 - \|Y - X\hat{\beta}\|^2 \\ &= \hat{\beta}' L'[L'(X'X)^{-1}L]^{-1}L\hat{\beta} := SS_H \end{aligned}$$

从而拒绝域为

$$\frac{SS_H/(p-q)}{SS_e/(n-p)} > F_{p-q,n-p}(\alpha)$$

多元多重线性回归模型

对多元线性回归模型

$$\mathbf{Y}_{n\times m} = X_{n\times p}B_{p\times m} + \boldsymbol{\epsilon}_{n\times m}$$

若假定

$$\boldsymbol{\epsilon} \sim N_{n\times p}(0, I_n \otimes \Sigma), \quad \Sigma_{m\times m} > 0$$

则由前面 (上讲定理 4) 的讨论知

1. 当 X 满秩时候, B 和 Σ 的最大似然估计分别为

$$\hat{B} = (X'X)^{-1}X'\mathbf{Y}, \quad \hat{\Sigma}^* = \frac{1}{n}\mathbf{Y}P\mathbf{Y}$$

2. 而且, $\hat{B} \sim N_{p\times m}(B, (X'X)^{-1} \otimes \Sigma)$; $n\hat{\Sigma}^* \sim W_m(n-p, \Sigma)$.

此时, 记 $B = (B'_{(1)}, B'_{(2)})'$, 其中 $B_{(1)}$ 为 $q \times m$ 矩阵, 相应地, 记 $X = [X_1, X_2]$, 则

$$\mathbf{Y} = X_1B_{(1)} + X_2B_{(2)} + \boldsymbol{\epsilon}$$

考虑假设检验问题

$$H_0 : B_{(2)} = 0 \leftrightarrow H_1 : B_{(2)} \neq 0$$

令 $L_{s\times p} = (0, I_{p-q})$, 则上述假设等价于

$$H_0 : LB = 0 \leftrightarrow H_1 : LB \neq 0$$

由于假定了正态分布, 从而似然比检验是自然的.

1.1.1 似然比检验

在 H_0 下, $\mathbf{Y} = X_1 B_{(1)} + \boldsymbol{\epsilon}$, 因此

$$\max_{B_{(1)},\Sigma} L(B_{(1)}, \Sigma) = L(\hat{B}_{(1)}, \hat{\Sigma}_1)$$

其中

$$\hat{B}_{(1)} = (X_1'X_1)^{-1}X_1'\mathbf{Y},$$
$$\hat{\Sigma}_1 = \frac{1}{n}\mathbf{Y}(I - X_1(X_1'X_1)^{-1}X_1')\mathbf{Y}$$

从而似然比检验统计量为

$$\Lambda = \frac{\max_{B_{(1)},\Sigma} L(B_{(1)},\Sigma)}{\max_{B,\Sigma} L(B,\Sigma)} = \frac{L(\hat{B}_{(1)},\hat{\Sigma}_1)}{L(\hat{B},\hat{\Sigma}^*)} = \left(\frac{|\hat{\Sigma}^*|}{|\hat{\Sigma}_1|}\right)^{n/2}$$

因此当 Λ 过小时候拒绝零假设 H_0.

等价地, 当

$$-2log\Lambda = -nlog\left(\frac{|\hat{\Sigma}^*|}{|\hat{\Sigma}_1|}\right) = -nlog\frac{|n\hat{\Sigma}^*|}{|n\hat{\Sigma}^* + n(\hat{\Sigma}_1 - \hat{\Sigma}^*)|}$$

过大时候拒绝零假设 H_0.

- 由于 $df = dim(B, \Sigma) - dim(B_{(1)}, \Sigma) = (p-q)m$, 因此在零假设下,

$$-2log\Lambda \rightsquigarrow \chi^2_{(p-q)m}, n \to \infty$$

从而渐近 α 的似然比检验拒绝域为

$$-2log\Lambda > \chi^2_{(p-q)m}(\alpha)$$

- Bartlett(1954) 修正上述似然比检验统计量以得到更佳的卡方近似:

$$-(n-p-\frac{1}{2}(m-p+q+1))log\frac{|\hat{\Sigma}^*|}{|\hat{\Sigma}_1|} > \chi^2_{(p-q)m}(\alpha)$$

1.1.2 其它检验方法

对一元回归模型中回归系数的线性检验问题, 常常使用 sequential(extra) sum of squares (对回归平方和进行分解来体现依次引入解释变量所带来的贡献) 方法来进行检验.

对多元线性回归模型有如下平方和与交叉积分解

$$
\begin{aligned}
SSCP_T &= \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}' \\
&= (\mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}}) - n\bar{\mathbf{y}}\bar{\mathbf{y}}' \\
&= \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} + (\hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}') \\
&= SSCP_{res} + SSCP_{reg} \\
&= SSCP_{res} + SSCP_{reg}(X_1) + SSCP_{reg}(X_2|X_1)
\end{aligned}
$$

其中 $SSCP_{reg}(X_1)$ 表示解释变量仅有 X_1 时候的回归平方和与交叉积, $SSCP_{reg}(X_2|X_1)$ 表示在已有 X_1 时, 再增加 X_2 后回归平法和与交叉积的增量. 利用上讲中有约束条件的最小二乘估计结果, 有

$$
\begin{aligned}
SSCP_{reg}(X_2|X_1) &= SSCP_{reg} - SSCP_{reg}(X_1) \\
&= SSCP_{res}(X_1) - SSCP_{res} \\
&= \hat{B}'L'[L(X'X)^{-1}L']^{-1}L\hat{B}.
\end{aligned}
$$

记

$$SSCP_{reg}(X_2|X_1) = n(\hat{\Sigma}_1 - \hat{\Sigma}^*) := H$$
$$SSCP_{res} = n\hat{\Sigma}^* := E$$

流行的统计软件一般会报告四种多元检验统计量, 它们都是 HE^{-1} 的特征根的函数. 为此, 记 $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_s$ 为 HE^{-1} 的非零特征根, $s = min(m, p-q)$, 则

- Likelihood ratio:

$$-nlog\frac{|E|}{|E+H|} = nlog|I + HE^{-1}| = n\sum_{i=1}^{s} log(1+\eta_i)$$

- Wilk's Lambda:

$$\frac{|E|}{|E+H|} = \frac{1}{|I+HE^{-1}|} = \prod_{i=1}^{s}\frac{1}{1+\eta_i}$$

- Pillai's trace:

$$tr[H(H+E)^{-1}] = \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i}$$

- Hotelling-Lawley trace:

$$tr[HE^{-1}] = \sum_{i=1}^{s} \eta_i$$

- Roy's greatest root:

$$\frac{\eta_1}{1+\eta_1}$$

检验统计量的选择:

- Wilk's test 等价于似然比检验
- 当多个 η_i 较大时候, Roy's test 表现较差. 模拟研究表明当仅有一个 η_i 较大时候, Roy's test 的功效最好.

- 当样本量充分时候, Wilk's Lambda, Roy's test 和 Hotelling-Lawley trace test 渐近等价.
- 当四种检验方法报告的 p 值差异巨大时候, 我们使用 Wilk's test.

1.1.3 一般线性假设

对一般的线性假设

$$H_0 : A_{s\times p}B = C_{s\times m} \leftrightarrow H_1 : AB \neq C$$

其中 $Rank(A) = s < m$. 记 B_0 为其任一特解, 令 $\tilde{\mathbf{Y}} = \mathbf{Y} - XB_0$, $\tilde{B} = B - B_0$, 则原模型等价于

$$\tilde{\mathbf{Y}} = X\tilde{B} + \boldsymbol{\epsilon}$$

$$A\tilde{B} = 0$$

记 $\tilde{B}$ 的最小二乘估计为 $\tilde{B}_H$, 则 $\tilde{B}_H = (X'X)^{-1}X'\tilde{\mathbf{Y}} = \hat{B} - B_0$. 由之前的结论知

$$\begin{aligned} E &= (\tilde{\mathbf{Y}} - X\tilde{B}_H)'(\tilde{\mathbf{Y}} - X\tilde{B}_H) \\ &= (\mathbf{Y} - X\hat{B})'(\mathbf{Y} - X\hat{B}) \\ H &= \tilde{B}_H' A'[A(X'X)^{-1}A']^{-1}A\tilde{B}_H \\ &= (A\hat{B} - C)'[A(X'X)^{-1}A']^{-1}(A\hat{B} - C) \end{aligned}$$

从而可以使用之前的几种检验方法.

1.2 置信区间

回归系数矩阵及其线性函数

前面我们已经讨论了对回归系数矩阵或者其线性函数的假设检验问题. 利用假设检验和置信区间的等价性, 可以得出回归系数矩阵或者其线性函数的置信域. 例如

↑Example

对多元线性回归模型 $\mathbf{Y} = X_1 B_{(1)} + X_2 B_{(2)} + \boldsymbol{\epsilon}$, 求 $B_{(2)}$ 的 $1-\alpha$ 置信域.

↓Example

解 由于对假设 $H_0 : LB = B_{(2)0}$, 其中 $L = (0, I_{p-q})$, $B_{(2)0}$ 为固定的一个 $(p-q) \times m$ 矩阵. 其似然比检验统计量为

$$-nlog\frac{|E|}{|E+H|} \rightsquigarrow_{H_0} \chi^2_{(p-q)m}$$

其中

$$E = (\mathbf{Y} - X\hat{B})'(\mathbf{Y} - X\hat{B})$$

$$H = (L\hat{B} - B_{(2)0})'[L(X'X)^{-1}L']^{-1}(L\hat{B} - B_{(2)0})$$

从而由假设 H_0 的接受域

$$-nlog\frac{|E|}{|E+H|} \leq \chi^2_{(p-q)m}(\alpha)$$

所确定的 $B_{(2)0}$ 的取值区域即为所求的一个渐近 $1-\alpha$ 置信域.

也可以使用其他检验方法构造置信域. 由于假设检验统计量基于 HE^{-1} 的特征根, 因此上述置信域不直观而不常用. 对多元多重线性回归模型, 对其回归系数的假设进行检验是更感兴趣的.

预测场合下的置信域

在对回归模型拟合完成后, 当有了新的协变量 $\mathbf{x}_0$, 感兴趣的问题对回归模型

$$Y_0 = B'\mathbf{x}_0 + \epsilon_0$$

给出感兴趣的置信域.

预测平均响应的置信域

给定一组协变量值 $\mathbf{x}_0' = [1, x_{01}, \ldots, x_{0(p-1)}]$, 则平均响应 $B'\mathbf{x}_0$ 的预测值

$$\hat{Y}_0 = \hat{B}'\mathbf{x}_0 \sim N_m(B'\mathbf{x}_0, \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0\Sigma)$$

注意到 $(n-p)\hat{\Sigma}$ 服从 $W_m(n-p, \Sigma)$ 且与 $\hat{B}$ 相互独立, 于是

$$\begin{aligned} T^2 &= \Big(\frac{\hat{B}'\mathbf{x}_0 - B'\mathbf{x}_0}{\sqrt{\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0}}\Big)'\Big(\hat{\Sigma}^{-1}\Big)\Big(\frac{\hat{B}'\mathbf{x}_0 - B'\mathbf{x}_0}{\sqrt{\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0}}\Big) \\ &\sim \frac{m(n-p)}{n-p-m+1}F_{m,n-p-m+1} \end{aligned}$$

从而 $B'\mathbf{x}_0$ 的 $1-\alpha$ 置信域为

$$T^2 \le \frac{m(n-p)}{n-p+1-m} F_{m,n-p-m+1}(\alpha)$$

- 记 $\hat{\Sigma} = (\hat{\sigma}_{ij})$, 利用 Cauchy-Schwarz 不等式, 可得 $B'\mathbf{x}_0$ 的所有分量的 $1-\alpha$ 同时置信区间为

$$\mathbf{x}_0'\hat{B}_{(i)} \pm \sqrt{\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0\hat{\sigma}_{ii}} \sqrt{\frac{m(n-p)}{n-p-m+1} F_{m,n-p-m+1}(\alpha)},$$

$$i = 1, \ldots, m$$

- 上述同时置信区间一般较大, 应用上常常使用 Bonferroni 置信区间:

$$\mathbf{x}_0'\hat{B}_{(i)} \pm \sqrt{\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0\hat{\sigma}_{ii}} t_{n-p}(\alpha/m),$$

$$i = 1, \ldots, m$$

响应变量的置信域

对新的响应变量 $Y_0 = B'\mathbf{x}_0 + \epsilon_0$, 注意到

$$Y_0 - \hat{B}'\mathbf{x}_0 = (B - \hat{B})'\mathbf{x}_0 + \epsilon_0 \sim N_m(0, (1 + \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)\Sigma)$$

因此 Y_0 的 $1-\alpha$ 置信域为

$$\Big(\frac{Y_0 - \hat{B}'\mathbf{x}_0}{\sqrt{1 + \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0}}\Big)'\Big(\hat{\Sigma}^{-1}\Big)\Big(\frac{Y_0 - \hat{B}'\mathbf{x}_0}{\sqrt{1 + \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0}}\Big)$$
$$\leq \frac{m(n-p)}{n-p+1-m}F_{m,n-p+1-m}(\alpha)$$

- Y_0 的所有分量 Y_{0i} 的 $1-\alpha$ 同时置信区间为

$$\mathbf{x}_0'\hat{B}_{(i)} \pm \sqrt{(1 + \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)\hat{\sigma}_{ii}}\sqrt{\frac{m(n-p)}{n-p-m+1}F_{m,n-p-m+1}(\alpha)},$$
$$i = 1, \ldots, m$$

- Y_0 的所有分量 Y_{0i} 的 $1-\alpha$ Bonferroni 同时置信区间为

$$\mathbf{x}_0'\hat{B}_{(i)} \pm \sqrt{(1+\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)\hat{\sigma}_{ii}}t_{n-p}(\alpha/m),$$
$$i = 1, \ldots, m$$

1.3 模型诊断

对线性模型的诊断包括检查线性模型的假设条件是否满足, 以及对变量进行选择.

残差分析

对一元正态线性模型, 记 $\hat{\epsilon}$ 为残差, 则

- $E\hat{\epsilon} = 0, Var(\hat{\epsilon}) = \sigma^2(I - H)$
  - $\widehat{Var(\hat{\epsilon}_j)} = s^2(1 - h_{jj})$
  - 标准化残差 $\hat{\epsilon}_j^* = \frac{\hat{\epsilon}_j}{\sqrt{s^2(1-h_{jj})}} \rightsquigarrow N(0,1)$ 从而常常使用 Q-Q 图来检查这一点.
- 残差满足 $\hat{\epsilon}'\hat{Y} = 0$, $\hat{\epsilon}'X = 0$.
  - 残差 $\hat{\epsilon}_j$ 对 $\hat{Y}_j$ 作图, 若模型正确, 则图形应该以 0 为中心均匀分布, 没有明显的趋势.

– 残差 $\hat{\epsilon}_j$ 对任一解释变量 X_{kj} 作图, 若模型正确, 则图形应该以 0 为中心均匀分布, 没有明显的趋势.

- 误差独立性假设很重要, 但是不易检查, 常常使用残差对时间 (观测顺序) 作图. 比如 (**Durbin-Watson test**):

$$d = \frac{\sum_{j=2}^{n} (\hat{\epsilon}_j - \hat{\epsilon}_{j-1})^2}{\sum_{j=1}^{n} \hat{\epsilon}_j^2}$$

容易看出, $d \rightsquigarrow 2(1-r)$, 其中 r 为样本 (一步) 自相关系数. d 的值在 0 到 4 之间, $d = 2$ 表明不相关, d 显著小于 2 表明可能存在正相关, d 值显著大于 2 则可能存在负相关.

对多元多重线性回归模型, 其可以视为 m 个一元多重线性回归模型, 因此可以使用上面的残差诊断方法来检查多元多重线性回归模型的假设.

变量选择

多重线性模型变量选择有多种方法, 常用的包括调整的 R^2, C_p, AIC, BIC 等等. 对一个候选的线性回归模型 $\mathcal{M}$:

$$y_j = \beta_0 + \beta_1 x_{j1} + \cdots + \beta_{k-1} x_{j(k-1)} + \epsilon_j, j = 1, \ldots, n$$

- AIC 准则:

$$AIC = nlog(SS_e(\mathcal{M})) + 2 * k$$

- BIC 准则:

$$BIC = nlog(SS_e(\mathcal{M})) + k \cdot log(n)$$

对多元多重线性回归模型 $\mathcal{M}$):

$$Y_{n\times m} = X B_{k\times m} + \boldsymbol{\epsilon}$$

记 $|\hat{\Sigma}^*(\mathcal{M})| = |\frac{1}{n}SSPE(\mathcal{M})|$, 则模型选择准则 AIC 和 BIC 可以类似定义

$$AIC = nlog(|\hat{\Sigma}^*(\mathcal{M})|) + 2m * k$$

$$BIC = nlog(|\hat{\Sigma}^*(\mathcal{M})|) + k \cdot mlog(n)$$