《线性回归》
— 线性模型的诊断

杨　　瑛

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

2019.05.07

Outline

1. 线性模型中的诊断方法和处理办法
   - 引言
   - 残差和帽子矩阵对角线

Outline

1. 线性模型中的诊断方法和处理办法
   - 引言
   - 残差和帽子矩阵对角线

主要内容：

- 主要考虑如何检验线性模型中假设偏离的方法，以及通过变量代换等克服这些可能的偏离.
- 最常用的诊断工具：各种类型的残差以及度量 （诸如帽子矩阵对角线）.
- 给定协变量$\mathbf{x} = (x_0, x_1, \cdots, x_{p-1})^T$ $(x_0 = 1)$. 我们欲建立模型

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

  但是如果真的模型是$\mathbf{E}[\mathbf{Y}|\mathbf{x}] = \mu(\mathbf{x})$, $\mu(\mathbf{x})$是$\mathbf{x}$的非线性函数，则上面模型中的β的LSE 将不在有任何意义.
- 我们需要可视化的方法【图形】了解μ的特性，决定线性形式$\mu(\mathbf{x}) = \mathbf{x}^T\beta$是否近似地满足。
- 如果不满足，我们将对$\mathbf{x}$进行适当的变换，以期达到较好拟合效果.

主要内容：

- 标准的线性回归模型假定方差函数Var[$\mathbf{Y}|\mathbf{x}$]不依赖于解释变量$\mathbf{x}$. 如果Var[$\mathbf{Y}|\mathbf{x}$] $= w(\mathbf{x})$, 其中$w(\mathbf{x})$不是常数，则LSE仍然是无偏估计，但不是有效估计.
- 如果$w(\mathbf{x})$已知，则可以实施加权最小二乘估计.
- 如果$w(\mathbf{x})$未知，必须先确定是否可以假定为常数.
- 如果不能确定$w(\mathbf{x})$是常数，则必须估计$w(\mathbf{x})$, 而且要把这个估计纳入到回归系数估计新的方法之中. 另外一种可行方法是对响应变量做适当的变换期望误差方差近似齐次.
- 即便方差函数是常数，误差ϵ也有可能是不独立的. 例如，数据按照一定时间顺序收集，则数据有可能是不独立. 所以还必须要判断误差ϵ是否是独立. Durbin-Watson检验是检验序列是否独立的有效方法之一.

主要内容：

- 如果随机误差不是正态的，但是只要协变量的联合分布是近似正态的，则估计不会出太大的问题.
- 如果不是这种情况，则需要对协变量，或者响应变量，或者二者同时进行变换，使得随机误差为正态分布.
- 异常值（outliers）会‘吸引’拟合直线或者拟合平面，将会导致对其余观测值拟合的较差. 如果协变量中有极端值，则影响可能是非常显著的. 这里我们采用两种观点：
  - ✓ 识别出异常值，在计算时降低异常值的权重或者将他们在拟合模型之前删除之
  - ✓ 使用稳健的回归方法来对抗异常值
- 检查回归设计矩阵列向量之间的共线性，然后给出处理共线性的方法(已讲)

主要内容（小结）：

- 在线性回归模型的应用中，需要对模型进行诊断和处理，涉及到如下的主要内容：
  - ✓ 异方差【检验】
  - ✓ 序列相关性【检验】
  - ✓ （随机误差的）正态性【检验】
  - ✓ 变换（协变量，响应变量，或者二者）（Box-Cox变换）
  - ✓ 异常值检验和判断（Cook's distance）
  - ✓ （列协变量之间的）共线性(方差膨胀因子)
- 这是线性回归诊断的几个重要方面，对于这些问题，既要有诊断的方法，还要有解决这些问题的方法.
- 在线性模型之下，这几个问题的研究相对独立.
- 各种残差和帽子矩阵是诊断的最基本的工具.
- 特别注意： 要从正反两个方面搞清楚每一个假设的作用是什么？如果这些假设不成立时，对估计和推断会有和影响。

- 考虑通常的模型

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

其中$\mathbf{X}$是秩为p的$n \times p$矩阵.

- 诊断模型的主要工具是残差. 设$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, 则残差定义为:

$$\begin{aligned} \mathbf{e} &= (\mathbf{I}_n - \mathbf{P})\,\mathbf{Y} \\ &= (\mathbf{I}_n - \mathbf{P})\,\epsilon \end{aligned}$$

后面一个等式成立是因为$(\mathbf{I}_n - \mathbf{P})\mathbf{X} = \mathbf{0}$.

- 拟合值是$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} = \mathbf{X}\widehat{\beta}$,
- 在回归诊断的文献中，学者们将投影矩阵$\mathbf{P}$称之为帽子矩阵(hat matrix), 因为将响应值$\mathbf{Y}_i$变换成立了拟合值$\widehat{\mathbf{Y}}_i$. 由此，投影矩阵P也会改写为用$\mathbf{H}$来表示. 后面用$\mathbf{H}$表示$\mathbf{P}$.

残差e的性质回顾：

这里简单回顾一下残差的性质：

- $\mathbf{H}^2 = \mathbf{H}.$
- $$E[\mathrm{e}] = (\mathbf{I}_n - \mathbf{H})\, E[\mathbf{Y}] = (\mathbf{I}_n - \mathbf{H})\, \mathbf{X}\beta = \mathbf{0}.$$
- $$\begin{aligned} \mathrm{Var}[\mathbf{e}] &= \mathrm{Var}\left[(\mathbf{I}_n - \mathbf{H})\, \mathbf{Y}\right] \\ &= (\mathbf{I}_n - \mathbf{H})\, \mathrm{Var}[\mathbf{Y}]\, (\mathbf{I}_n - \mathbf{H})^T \\ &= (\mathbf{I}_n - \mathbf{H})\, \sigma^2 \mathbf{I}_n\, (\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2\, (\mathbf{I}_n - \mathbf{H})\,. \end{aligned} \tag{1}$$

残差e的性质回顾(续)

- $E[\hat{\mathbf{Y}}] = \mathbf{H}E[\mathbf{Y}] = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$.
- $$\mathrm{Var}[\hat{\mathbf{Y}}] = \mathbf{H}\,\mathrm{var}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H}. \tag{2}$$
- $$\mathrm{Cov}[\mathbf{e}, \hat{\mathbf{Y}}] = \mathrm{Cov}\left[(\mathbf{I}_n - \mathbf{H})\,\mathbf{Y}, \mathbf{H}\mathbf{Y}\right] = \sigma^2\,(\mathbf{I}_n - \mathbf{H})\,\mathbf{H} = \mathbf{0}.$$

  在随机误差是正态性假定下，$\mathbf{e}$和$\hat{\mathbf{Y}}$是独立的.
- 设$\mathbf{H} = (h_{ij})$, 则$\mathbf{H}$的对角线的元素h_{ii}称为帽子矩阵对角线(hat matrix diagonals), 更简单地记为h_i.
- 由(1)知，
$$\mathrm{Var}[\epsilon_i] = \sigma^2(1 - h_i).$$

- 从上面的结果可以看出，如果模型是正确的，残差的方差依赖于帽子矩阵对角线. 由于这个原因，残差需要重整，使得其方差近似为单位方差. 考虑内部学生化残差(internally Studentized residual)：

$$r_i = \frac{e_i}{S(1-h_i)^{1/2}}, \tag{3}$$

其中$S^2 = \mathbf{e}^T\mathbf{e}/(n-p)$是$\sigma^2$的无偏估计.

- Cook and Weisberg (1982)证明了如下结论：

$$r_i \sim \text{Beta}(\frac{1}{2}, \frac{1}{2}(n-p-1)),$$

$r_1, \cdots, r_n$是同分布的（没有说是独立的！）【作业！】

- 由于残差可能受到异常值的影响，有些学者喜欢用外部学生化残差(externally Studentized residual)：

$$t_i = \frac{e_i}{S(i)\left(1-h_i\right)^{1/2}}, \tag{4}$$

其中$S(i)$的计算与S类似，但需要剔除第i个观测记录，这样以来有可去掉第i观测值可能是异常值的影响.

- 可以推导出残差外部化残差t_i的分布.

外部学生化残差分布

为研究外部学生化残差的分布，需要研究$S(i)$和S之间的关系.

Theorem

设$\widehat{\beta}$和$\widehat{\beta}(i)$分别表示β的利用全数据和剔除掉第i记录之后得到的*LSE*, 则

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i e_i}{1 - h_i}. \tag{5}$$

定理的证明：

用$\mathbf{X}(i)$表示从$\mathbf{X}$之中剔除掉第i行之后的矩阵. 因为：

$$\mathbf{X}(i)^T\mathbf{X}(i) = \mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T.$$

故我们从矩阵求逆的公式(见附录A.9.4)得到：

定理的证明(续)

$$\begin{aligned}
\left(\mathbf{X}(i)^T\mathbf{X}(i)\right)^{-1} &= \left(\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T\right)^{-1} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1 - \mathbf{x}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1 - h_i}. \quad (6)
\end{aligned}$$

定理的证明(续)

因此，

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(i) &= \left[\mathbf{X}(i)^T\mathbf{X}(i)\right]^{-1}\left(\mathbf{X}^T\mathbf{Y}-\mathbf{x}_iY_i\right)\\
&= \left[\mathbf{X}^T\mathbf{X}\right)^{-1}+\frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1-h_i}]\left(\mathbf{X}^T\mathbf{Y}-\mathbf{x}_iY_i\right)\\
&= \hat{\boldsymbol{\beta}}-\frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i}{1-h_i}\left[Y_i\left(1-h_i\right)-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+h_iY_i\right]\\
&= \hat{\boldsymbol{\beta}}-\frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_ie_i}{1-h_i}. \qquad (7)
\end{aligned}$$

利用上面的定理，我们有：

$$
\begin{aligned}
(n-p-1)S(i)^2 &= \sum_{l\neq i}\left[Y_l - \mathbf{x}_l^T\hat{\boldsymbol{\beta}}(i)\right]^2 \\
&= \sum_{l\neq i}\left[Y_l - \mathbf{x}_l^T\left(\hat{\boldsymbol{\beta}} - \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i e_i}{1-h_i}\right)\right]^2 \\
&= \sum_{l\neq i}\left(e_l + \frac{h_{li}e_i}{1-h_i}\right)^2 \\
&= \sum_{l=1}^{n}\left(e_l + \frac{h_{li}e_i}{1-h_i}\right)^2 - \frac{e_i^2}{(1-h_i)^2}. \qquad (8)
\end{aligned}
$$

因为$\mathbf{H}$是投影矩阵，故$\mathbf{He}=\mathbf{0}$, $\mathbf{H}^2=\mathbf{H}$, 由此可以得到：$\sum_l h_{li}e_l=0$, $\sum_l h_{il}^2=h_i$. 把这些关系应用到(8), 就得到：

$$(n-p-1)S(i)^2=(n-p)S^2-\frac{e_i^2}{1-h_i}. \tag{9}$$

从关系$e_i^2/(1-h_i)=r_i^2S^2$ 可以得到：

$$\begin{aligned} t_i^2 &= \frac{e_i^2(n-p-1)}{(n-p-1)S(i)^2(1-h_i)} \\ &= \frac{e_i^2(n-p-1)}{\left[(n-p)S^2-e_i^2/(1-h_i)\right](1-h_i)} \\ &= \frac{e_i^2}{S^2(1-h_i)}\frac{n-p-1}{n-p-r_i^2} \\ &= \frac{r_i^2(n-p-1)}{n-p-r_i^2}=\frac{B}{1-B}(n-p-1). \end{aligned} \tag{10}$$

其中$B = r_i^2(n-p)^{-1} \sim \text{Beta}(\frac{1}{2}, \frac{1}{2}(n-p-1))$.
Beta分布还有这样一个性质：如果$B \sim \text{Beta}(\alpha/2, \beta/2)$,
则$\beta B\{\alpha(1-B)\}^{-1} \sim F_{\alpha,\beta}$.
令$\alpha = 1, \beta = n-p-1$, 则有

$$t_i^2 \sim F_{1,(n-p-1)},$$

等价地有：

$$t_i \sim t_{n-p-1}.$$

Mahalanobis距离

帽子矩阵对角线其实可以解释为$p-1$维空间中距离的一种度量.

- 由Seber and Lee (2003)中的（3.52）式

$$h_i = n^{-1} + (n-1)^{-1}\mathrm{MD}_i, \tag{11}$$

其中MD_i是Mahalanobis距离：

$$\mathrm{MD}_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

表示$\mathbf{X}$的第i行（观测）与平均值$\bar{\mathbf{x}}$之间的距离. 因此，帽子矩阵对角线第i个点的异常程度.

- 上面的度量依赖于样本均值$\bar{\mathbf{X}}$和样本协方差矩阵，会受到异常值的影响. 后面还要研究和考察更加稳健的度量.

帽子矩阵对角线的进一步性质：

- 如果回归模型中含有截距项，则$\mathbf{X}$的第一列元素都是1.
  由Seber and Lee (2003) 中的(9.13)式得到

$$n^{-1} \leq h_i \leq 1. \tag{12}$$

此外，还有下面的结论：

$$\sum_i h_i = \mathrm{tr}(\mathbf{H}) = p. \tag{13}$$

故，帽子矩阵对角线之平均值为p/n.

例子

这里给出两个简单的例子.

Example

考虑过原点的回归模型

$$Y_i = \beta x_i + \varepsilon_i, \tag{14}$$

因为$\mathbf{X} = (x_1, x_2, \ldots, x_n)^T = \mathbf{x}$，故帽子矩阵为：

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{x}\mathbf{x}^T\|\mathbf{x}\|^{-2}$$

$h_i = x_i^2 / \sum_k x_k^2$.

例子

Example

考虑简单线性回归模型

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

帽子矩阵对角线为：

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}. \tag{15}$$

可以尝试解释这个结果.