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

第25讲 影响分析

2020.5.20

Jackknife 

附录：Jackknife (折刀,刀切法) 简介

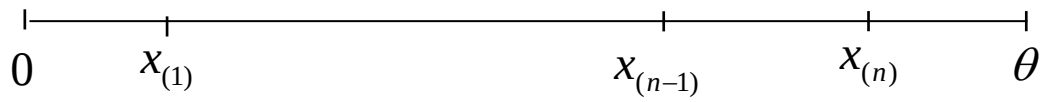
Tukey's Jackknife 方法除了在影响分析中有应用之外，更多地应用于预测的交叉验证和用于偏差校正和方差估计的刀切法中。The term Jackknife was coined by Tukey, "If you had exactly the right tool for the job, you'd use it. But if you don't, then you'd use a jackknife." Jackknife method is an all-purpose tool.

例1(估计上界)：样本$x_1,...,x_n$ $iid\sim$区间$[0,\theta]$，显然$x_{(n)}=\max(x_i)$低估θ，$x_{(n)}$需要上调。

如果假设均匀抽样，即 $x_1,...,x_n$ $iid\sim U(0,\theta)$, 则$[0,x_{(1)}],[x_{(1)},x_{(2)}],...,[x_{(n-1)},x_{(n)}],[x_{(n)},\theta]$

长度期望相同，前n个区间的平均长度$=\left(x_{(1)}+(x_{(2)}-x_{(1)})+..+(x_{(n)}-x_{(n-1)})\right)/n=x_{(n)}/n,$

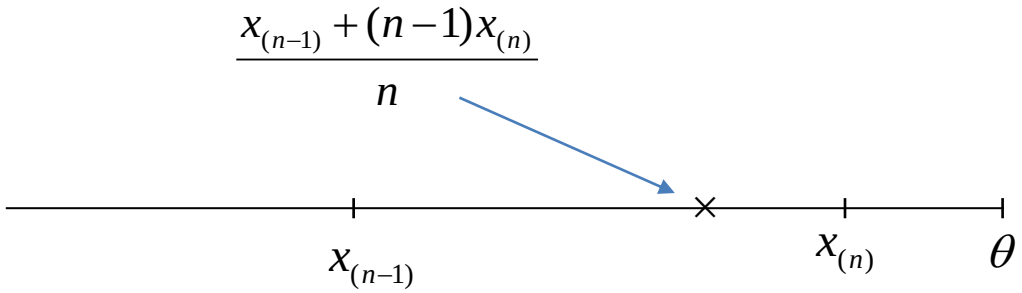
所以$x_{(n)}$需要上调至$x_{(n)}+x_{(n)}/n=\dfrac{n+1}{n}x_{(n)}.$



事实上，由$E(x_{(n)})=\dfrac{n}{n+1}\theta\Rightarrow E\left(\dfrac{n}{n+1}x_{(n)}\right)=\theta,$可很容易地得到这个无偏估计。

如果我们不知道总体分布，如何估计上界？我们使用删除一个数据点(delete-1)的方法：

假如$x_1,...,x_n$中某一个x_i没有采集到,概率是$1/n$(即等可能地删除一个数据点),那么最大值有可能是$x_{(n-1)}$,也可能是$x_{(n)}$,概率分别是$1/n$ 和$1-1/n$,平均来看，最大值为



它是$n-1$个样本点的最大值应该达到的值,与上界θ的距离$d_{n-1}=\theta-\frac{x_{(n-1)}+(n-1)x_{(n)}}{n}$,

而n个样本点的最大值与上界θ的距离$d_n=\theta-x_{(n)}$,因为两种情形下区间个数分别是$n,n+1$,令

$$nd_{n-1}=(n+1)d_n \Rightarrow \tilde{\theta}=2x_{(n)}-x_{(n-1)}$$

若令 $(n-1)d_{n-1}=nd_n \Rightarrow \tilde{\theta}=\frac{2n-1}{n}x_{(n)}-\frac{n-1}{n}x_{(n-1)}=2x_{(n)}-x_{(n-1)}-\frac{1}{n}\left(x_{(n)}-x_{(n-1)}\right)$

这是θ的Jackknife估计(见后页)。

Jackknife 校正偏差

问题及假设：给定基于样本$x_1,...,x_n$的θ的估计$\hat{\theta}=\hat{\theta}_n$, 假设

$$E(\hat{\theta})=\theta+\frac{c}{n},$$

我们希望估计$\hat{\theta}$的偏差$b=c/n.$

刀切法逐个删除数据点i，基于剩余的$n-1$个数据点得到θ的估计$\hat{\theta}^{(-i)}, i=1,...,n,$

注意到根据偏差假设，基于$n-1$个数据点的估计的偏差为$\frac{c}{n-1}$:

$$E(\hat{\theta}^{(-i)})=\theta+\frac{c}{n-1},$$

从而

$$E\left(\hat{\theta}^{(-i)}-\hat{\theta}\right)=\frac{c}{n-1}-\frac{c}{n}=\frac{c}{n(n-1)}$$

定义$\hat{\theta}^{(-i)}, i=1,...,n$的平均$\bar{\theta}=\sum_{i=1}^{n}\hat{\theta}^{(-i)}/n,$ 则

$$E(\bar{\theta}-\hat{\theta})=c/n(n-1).$$

即 $(n-1)E(\bar{\theta}-\hat{\theta})=c/n,$ 所以我们可以$(n-1)(\bar{\theta}-\hat{\theta})$估计$\hat{\theta}$的偏差$c/n.$

从而得到校正了偏差的 *Jackknife* 估计：

$$\tilde{\theta}_{Jackknife}=\hat{\theta}-(n-1)(\bar{\theta}-\hat{\theta})$$

注1：*Jacknife*的另一个主要用途在于计算统计量的方差。$\hat{\theta}$的方差的*Jackknife*估计：

$$\widehat{\mathrm{var}}(\hat{\theta})=\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^{(-i)}-\bar{\theta}\right)^2$$

注2: (Bootstrap自助法) 从样本$x_1,...,x_n$中有放回地抽取n个数据点，得θ的*Bootstrap* 估计$\hat{\theta}^*$,它被看作是$\hat{\theta}$的一个再抽样版本(*resampling*)。反复再抽样得到大量 $\hat{\theta}_1^*,\hat{\theta}_2^*,...,\hat{\theta}_B^*$,它们是$\hat{\theta}$的随机样本。

影响分析

远离数据中心的异常点会把回归直线“拉向”该点，对回归影响会比较大。因此影响分析首先判断每个样本点是否异常(远离中心)。影响分析试图发现对回归分析影响过大的样本点。

1. 异常点 (outlier)：响应变量y异常
2. 高杠杆点(high-leverage point)：自变量x异常
3. 高影响点(influential point)：(x, y)高影响，不一定异常

Jackknife

如果发现存在异常点或高影响点，如何处理？

1. 检查数据，如果是明显的记录错误，可删除或修改，但谨慎删除；
2. 采用稳健统计方法降低高影响点的影响。

模型：$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 I_n)$，$X$第一列为$\mathbf{1}$。

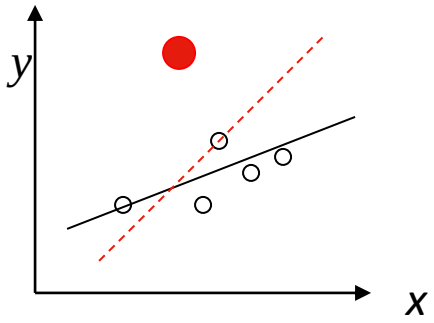
$\Leftrightarrow y_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i = \beta_0 + \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\gamma} + \varepsilon_i,\ \ \varepsilon_i \sim (0, \sigma^2), i = 1, \ldots, n$

其中

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \cdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1\ \tilde{\mathbf{x}}_1^{\mathsf{T}} \\ 1\ \tilde{\mathbf{x}}_2^{\mathsf{T}} \\ \cdots \\ 1\ \tilde{\mathbf{x}}_n^{\mathsf{T}} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p\text{-}1} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\gamma} \end{pmatrix}$$

LS估计$\hat{\boldsymbol{\beta}}$，拟合值$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = H\mathbf{y}$，残差$\mathbf{e} = (I - H)\mathbf{y}$

1. 异常点Outlier: y异常, 残差过大/小



残差 $e_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$,

如果标准化后的残差远离0，则第i个响应y_i可能是异常的。

因为残差向量$\mathbf{e} = (I_n - H)\mathbf{y}$, 所以 $\text{var}(\mathbf{e}) = \sigma^2(I_n - H) \Rightarrow \text{var}(e_i) = (1 - h_{ii})\sigma^2$

标准化残差： $r_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$， 其中 $\hat{\sigma} = \sqrt{\sum_{j=1}^{n} e_j^2 / (n-p)}$

注意上面定义中$\hat{\sigma}$估计用到了所有残差，如果y_i异常，对应的$|e_i|$偏大，$\hat{\sigma}$也偏大。所以有时使用如下的学生化残差：

学生化残差：

$$r_i^* = \frac{e_i}{\hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}}, \text{其中} \hat{\sigma}^{(-i)} = \sqrt{\| \mathbf{e}_{(-i)} \|^2 /(n-1-p)},$$

其中$\mathbf{e}_{(-i)}$为删除第i个观察后线性模型拟合的残差

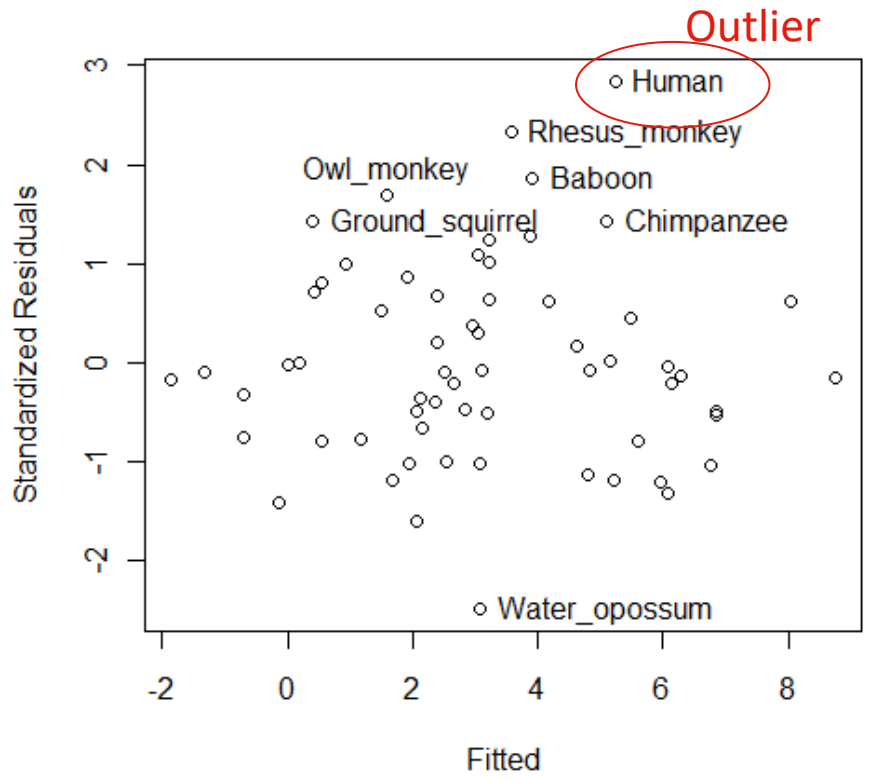
r_i，r_i^*近似地服从标准正态分布.

响应异常的判别标准：

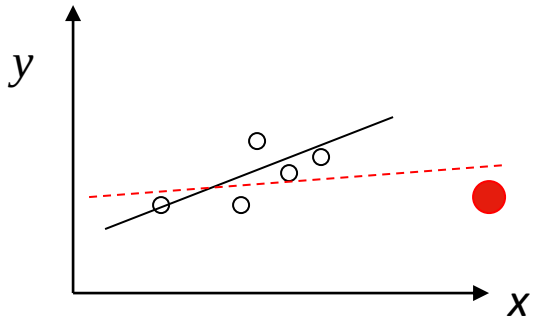
$|r_i|$或$|r_i^*|$较大时(比如≥ 3)，y_i可认为异常。

```
> rstandard(lm.out)
> rstudent(lm.out)
```

例1：动物脑重量与体重的关系：log(BrainWt)~log(BodyWt)



2. 高杠杆点：x异常，h过大



所有自变量为$\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_n$，$Z=(\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_n)^{\mathsf{T}}$，设计阵$X=(\mathbf{1},Z)=\begin{pmatrix}1 & \tilde{\mathbf{x}}_1^{\mathsf{T}}\\ ... & ...\\ 1 & \tilde{\mathbf{x}}_n^{\mathsf{T}}\end{pmatrix}$，

帽子矩阵/投影矩阵：$H=X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}=(h_{ij})$。

如果第i个自变量$\tilde{\mathbf{x}}_i$与中心$\bar{\mathbf{x}}/n$的马氏距离$d_S(\tilde{\mathbf{x}}_i,\ \bar{\mathbf{x}})$较大，则认为自变量$\tilde{\mathbf{x}}_i$异常。

$$d_S(\tilde{\mathbf{x}}_i,\ \bar{\mathbf{x}})=\sqrt{(\tilde{\mathbf{x}}_i-\bar{\mathbf{x}})^{\mathsf{T}}S^{-1}(\tilde{\mathbf{x}}_i-\bar{\mathbf{x}})}$$

自变量的样本均值：$\bar{\mathbf{x}}=(\tilde{\mathbf{x}}_1+...+\tilde{\mathbf{x}}_n)/n=Z^{\mathsf{T}}\mathbf{1}/n,$

样本方差：$S=\sum_{k=1}^{n}(\tilde{\mathbf{x}}_k-\bar{\mathbf{x}})(\tilde{\mathbf{x}}_k-\bar{\mathbf{x}})^{\mathsf{T}}/(n-1)=Z^{\perp\mathsf{T}}Z^{\perp}/(n-1)$