课程主页：http://staff.ustc.edu.cn/~ynyang/2023

第十三讲 回归诊断

2023.12.15

Tukey’s Jackknife 

回归诊断基于线性回归分析结果，对线性模型的合理性进行评判，并发现异常值等对回归结果影响恰当性和过大的数据点。主要包含两部分内容：

- **残差分析**：对模型假设（即线性性和方差齐性）的合理性进行诊断；
- **影响分析**：发现对回归分析结果影响较大的点。

注意

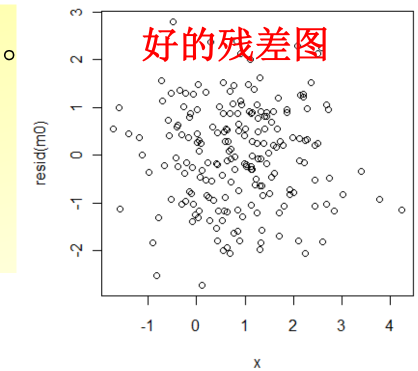
- R^2或回归方程的显著性检验$F = \frac{n-p}{p-1}\frac{R^2}{1-R^2}$称为拟合优度度量， 是一种整体拟合程度的度量, 不足以发现细节上的模型的不合理性。换言之，R^2或F值很大只能部分地说明模型合理或正确。
- 回归诊断提供了一些常用工具（残差图、标准化残差、杠杆值、Cook距离等），实际数据分析中不应局限于这些工具。

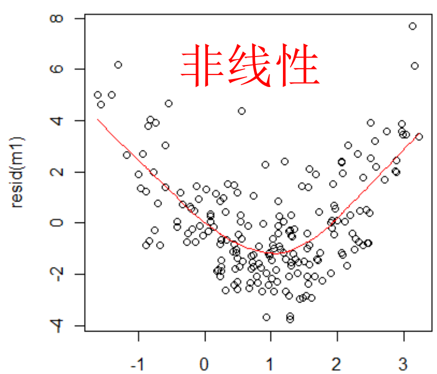
1. 残差分析

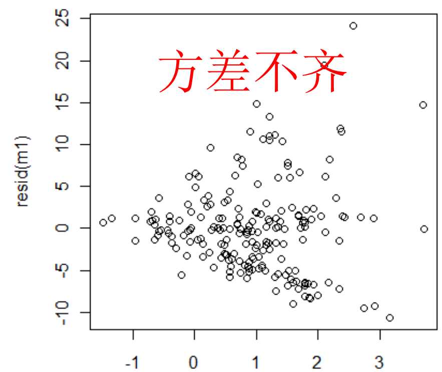
残差图

残差图：横轴为自变量或拟合值，纵轴为残差得散点图。

线性模型假设误差与自变量无关，故残差图上应没有任何明显的趋势，如右图所示。下图分别表明模型的线性假设和误差方差为常数的假设不成立。







R软件只提供拟合值-残差散点图，这是因为拟合值能最好地代表所有自变量。但有时也需要考察每个自变量的残差图。

解决方案

数据变换
非线性/多项式回归
GLM

数据变换
加权最小二乘
GLM

GLM：广义线性模型，针对指数族分布的回归模型。

成语 Box and Cox：轮流、互斥、小人物

George E. P. Box

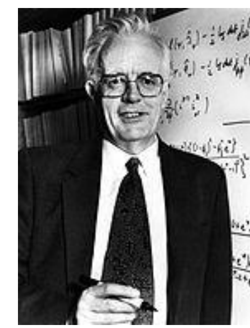
G.E.P.Box (1919-2013)，英国统计学家，Fisher的女婿，早期在北卡、普林斯顿工作，1960年创建威斯康星大学统计系，主要领域包括质量控制、时间序列和试验设计

Box–Cox transformations, Box-Jenkins models, Box–Behnken designs, robust statistics, etc.

David Roxbee Cox

Sir D.R. Cox (1924-2022)英国统计学家，任职于帝国理工、剑桥。

Proportional hazards model (Cox model), Cox process



Box-Cox变换： Box, George E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26 (2): 211–252.

Box-Cox变换

事实：联合正态分布情形下线性模型Gauss – Markov假设成立。

Box - Cox变换的基本思想：对于随机变量$y>0$，寻找某种幂次变换，使得变换之后的变量近似地服从正态分布（对称、均衡）。

Box-Cox变换

Box−Cox变换：

$$y>0 \rightarrow y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0 \end{cases}$$

选择λ, 使得 $y^{(\lambda)}$ 的分布服从正态分布。

注1：最常用的BC变换是对数变换。

注2：具体实施的时候，可忽略分母的 λ和分子上的−1，这里如此定义是为了将最常用的对数变换统一在幂次变换中。

注3：BC变换是单调变换，即不可能把U型变成线性。

注4：若y取值有正有负, 先平移再变换： $y>-a \rightarrow (y+a)^{\lambda}$

线性模型的BC变换

假设数据为 $(y_i, x_i), i=1,\dots,n$，假设存在幂次$\lambda \in R$, 使得响应变量的Box-Cox变换 $\mathbf{y}=(y_1,\dots,y_n)^{\mathsf{T}} \to \mathbf{y}^{(\lambda)}=(y_1^{(\lambda)},\dots,y_n^{(\lambda)})^{\mathsf{T}}$后

$\mathbf{y}^{(\lambda)}$满足正态模型：$\mathbf{y}^{(\lambda)} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0,\sigma^2 I_n) \Leftrightarrow \mathbf{y}^{(\lambda)} \sim N(X\boldsymbol{\beta},\sigma^2 I_n)$。

$\mathbf{y}^{(\lambda)}=(y_1^{(\lambda)},\dots,y_n^{(\lambda)})^{\mathsf{T}} \sim N(X\boldsymbol{\beta},\sigma^2 I_n)$的联合密度函数为

$$g(\mathbf{y}^{(\lambda)}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta}\|^2\right)$$

下面应用剖面极大似然方法（profile likelihood）求解最优的λ。

似然函数
$L = Pr(data)$

注意$\mathbf{y}^{(\lambda)}$含未知参数λ，其密度函数 $g(\mathbf{y}^{(\lambda)})$不是似然函数。

样本数据$\mathbf{y}=(y_1,\dots,y_n)^{\mathsf{T}}$的联合密度为似然函数：

$$L(\lambda,\beta,\sigma^2) = f(\mathbf{y}) = g(\mathbf{y}^{(\lambda)}) \times \left|\partial \mathbf{y}^{(\lambda)} / \partial \mathbf{y}\right|$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta}\|^2\right) \times \prod_{i=1}^{n} y_i^{\lambda-1}$$

对数似然函数：

$$l(\lambda,\beta,\sigma^2)=\log L(\lambda,\beta,\sigma^2)=C-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\|\mathbf{y}^{(\lambda)}-X\boldsymbol{\beta}\|^2+(\lambda-1)\sum\log(y_i),$$

极大似然法求解$\lambda,\boldsymbol{\beta},\sigma^2$使得$l(\lambda,\beta,\sigma^2)$最大，计算比较困难，但注意到

如果λ给定,那么$\boldsymbol{\beta},\sigma^2$的极大似然估计容易求解：

$$\hat{\boldsymbol{\beta}}(\lambda)=(X^\top X)^{-1}X^\top\mathbf{y}^{(\lambda)},\quad \hat{\sigma}^2(\lambda)=\|\mathbf{y}^{(\lambda)}-X\hat{\boldsymbol{\beta}}(\lambda)\|^2/n\triangleq RSS(\lambda)/n,$$

将$\hat{\boldsymbol{\beta}}(\lambda),\hat{\sigma}^2(\lambda)$代入对数似然函数，得仅含$\lambda$的对数剖面似然函数

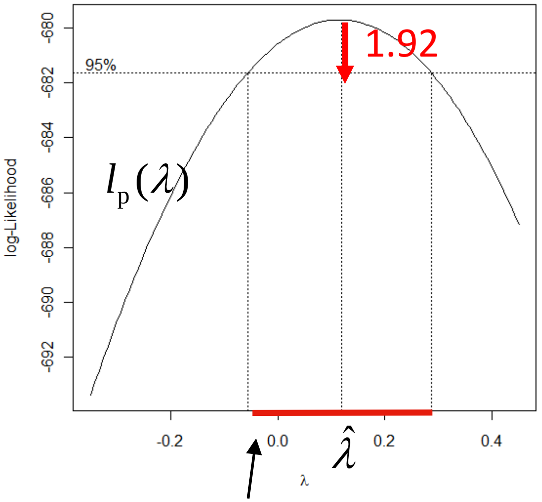
$$l_p(\lambda)=l(\lambda,\hat{\boldsymbol{\beta}}(\lambda),\hat{\sigma}^2(\lambda))=C-\frac{n}{2}\log RSS(\lambda)+(\lambda-1)\sum\log(y_i)$$

剖面似然方法极大化$l_p(\lambda)$，无显式解。

最简单的极大化求解方法：
逐点搜索，grid search

逐点搜索，对每一个网格点λ值，计算$l_p(\lambda)$,

$\hat{\lambda} = \underset{\lambda}{\mathrm{argmax}}\, l_p(\lambda)$为最优$\lambda$, 如下图。



λ的95%置信区间

似然比统计量$2(l_p(\hat{\lambda}) - l_p(\lambda)) \overset{\text{近似}}{\sim} \chi_1^2$

λ的95%置信区间:

$$\{\lambda : l_p(\lambda) \geq l_p(\hat{\lambda}) - 1.92\},$$

为从最高点下拉1.92处的水平线与$l_p(\lambda)$的两个交点之间的区间。

$1.92 = 3.84/2$

$3.84 = 1.96^2$

但实际应用中不一定取最优的$\hat{\lambda}$, 而是取其附近(95%置信区间内)的"容易解释"的值。比如,若$\hat{\lambda} = 0.61$，我们可以取$\lambda = 0.5$。

其它变换

方差稳定化变换适用范围较小，这里不做介绍（附录2）

离散变量连续化

因子变量合并水平，有次序的因子变量通过打分(*scoring*)转为连续变量。

(1) 因子变量相近的水平合并为一个水平；

(2) 有次序的因子变量通过打分转化为连续变量。例如，职称高、中、低分别打分为5,3,1。K水平因子需要$K-1$个参数，转化为连续变量后只需1个。（注意：谨慎打分！因子打分应该由专业人士提供）.

连续变量离散化

如果连续变量的数值含义不太具体明确的时候，或者连续的自变量与响应变量存在非线性关系时，可考虑将其离散化，转化为因子变量。

- 百分制转化为5分制（百分制成绩95和98都转化为5）；
- 血压值100和110没有本质差异，但它们和150有本质不同（150是高风险）。

例如$x=$血压值，$y=$某种疾病指标，模型$y=a+bx+\varepsilon$意味着不论血压多高，它对y的效应都是常数b,这显然不合理。以80,140为阈值将血压划分为高中低三个水平意味着上述模型转变成

$$y=a+b_1 1_{(x<80)}+b_2 1_{(x>140)}+\varepsilon$$

它表明低血压的效应为b_1，高血压的效应为b_2（相对于正常血压）

前述BC变换对响应变量y作BC变换：

$$boxcox(y\sim x)$$

在实际应用中，对自变量也可做Box-Cox变换:

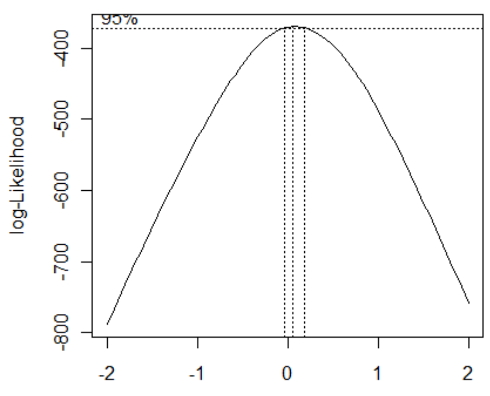
$$boxcox(x\sim y)$$

例1. 62种哺乳动物的脑重量与体重数据。

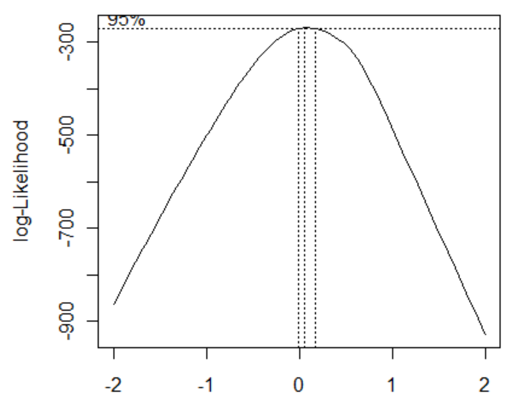
R: library(MASS)中的boxcox函数

左图：boxcox(BrainWt~BodyWt, data=brains) #响应BrainWt的BC变换

右图：boxcox(BodyWt~BrainWt, data=brains) #自变量BodyWt的BC变换



$\lambda \approx 0$, 对BrainWt做 log变换

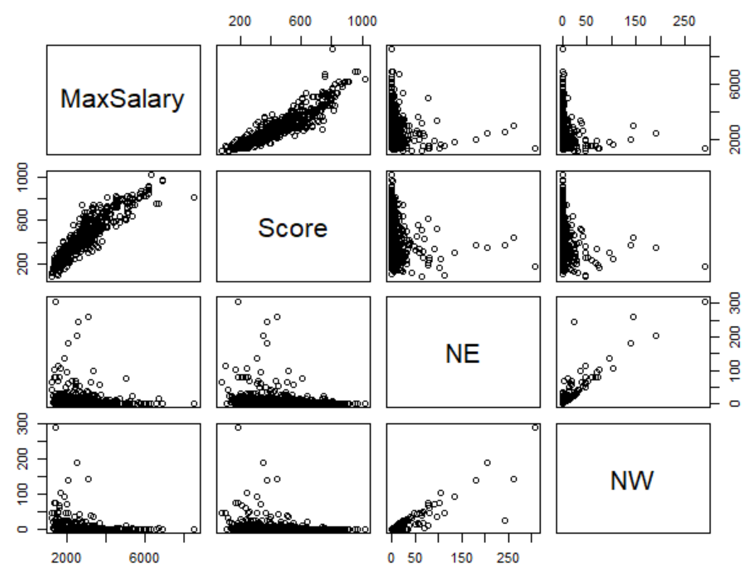


$\lambda \approx 0$, 对BodyWt做 log变换

例2. 数据集 salarygov (alr4) 汇总了美国政府某部门495种职位的信息，把包括每种职位的最高工资、每种职位的人数、女性人数和职位难度系数(Score)。目的是研究工资与职位难度的关系，特别地我们关心女性占主导的职位的工资情况是否偏低。变量具体描述如下：

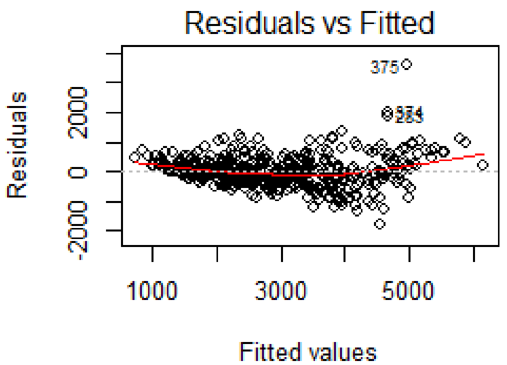
变量	描述
MaxSalary	职位最高工资
Score	职位难度系数（82-1017）
NE	该职位的雇员总数(number of employees)
NW	该职位的女性人数(number of women)

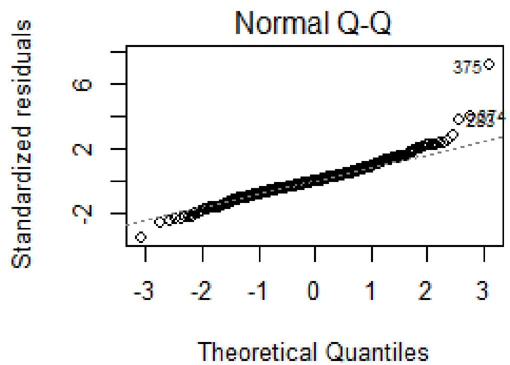
JobClass	NW	NE	Score	MaxSalary
Account_clerk	52	68	258	1549
Account_clerk_Intermediate	26	29	269	1712
Account_clerk_Principal	10	13	321	2182
Account_clerk_Senior	16	24	273	1982
Accountant	1	12	352	2555
Accountant_Chief	0	5	709	4060
……				



原始数据拟合线性模型

$$\text{MaxSalary}_i = \beta_0 + \beta_1 \times \text{Score}_i + \beta_2 \times \text{NW}_i + \beta_2 \times \text{NE}_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$$

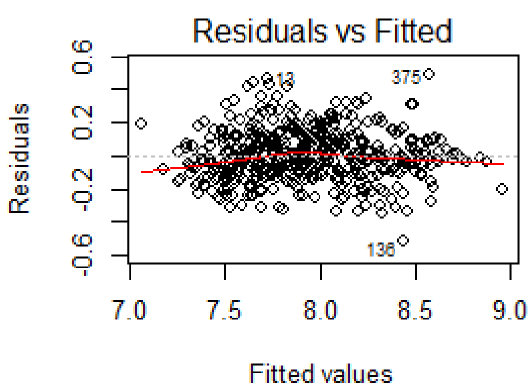


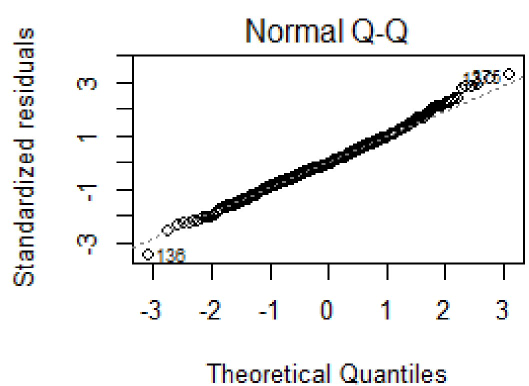


残差有非线性趋势和方差随拟合值增大的趋势。
第二个图qqnorm图检查残差是否符合正态分布

应用boxcox函数发现响应变量需要作对数变换，自变量无需变化，拟合

$$\log(\text{MaxSalary})_i = \beta_0 + \beta_1 \times \text{Score}_i + \beta_2 \times \text{NW}_i + \beta_2 \times \text{NE}_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$$

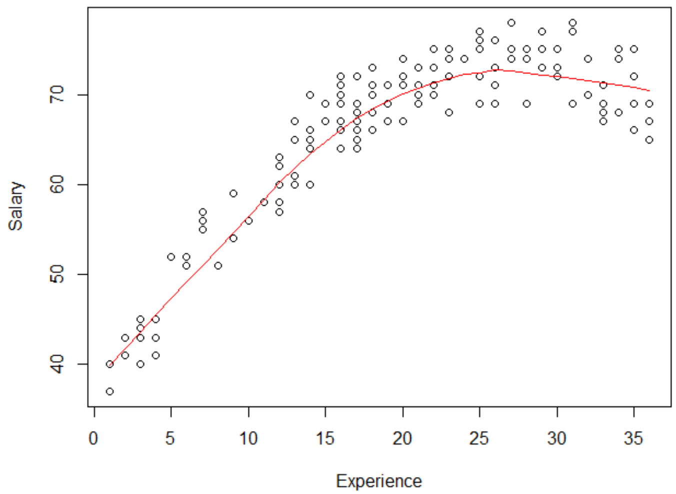




BC变换后，残差没有明显趋势。

例3. 数据集se (http://staff.ustc.edu.cn/~ynyang/2023/lab/se.xls) 是调查了134个职员(包括会计、工程师、系统管理员等）工资与工作经验数据。

变量	解释
Salary	年工资（1000$）
Experience	工龄（年）



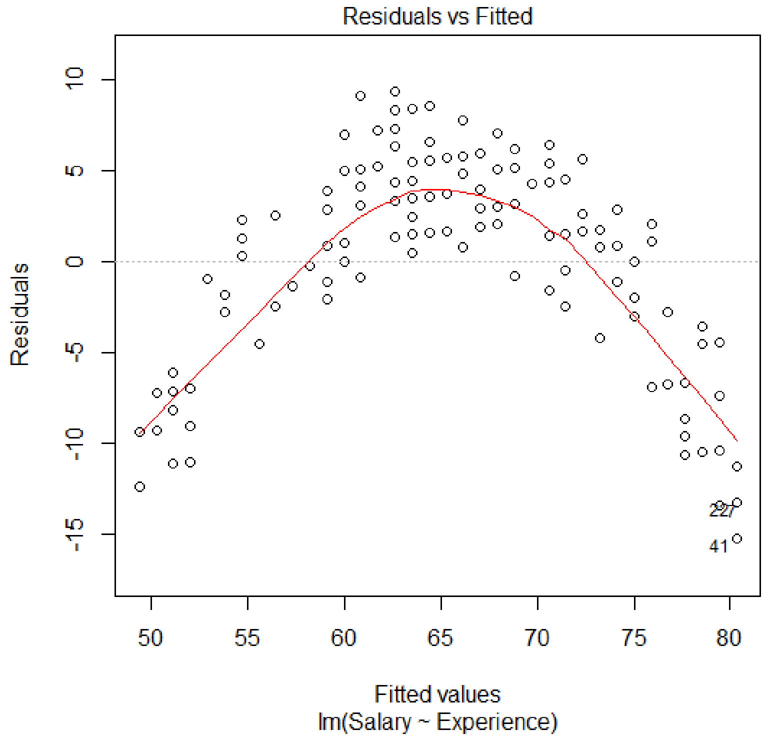
Salary	Experience
71	26
69	19
73	22
69	17
65	13
75	25
66	35
66	16
67	16
69	16
76	26
72	16
69	25
45	4
72	17
62	12
74	23

```
> plot( se )
> lines(lowess(se， f=2/3), col = 2) # lowess方法
```

拟合简单线性模型

```
> a=lm(Salary~Experience, data=se)
> plot(a )
```

$$\text{Salary} = \beta_0 + \beta_1 \times \text{Experience} + \varepsilon,$$

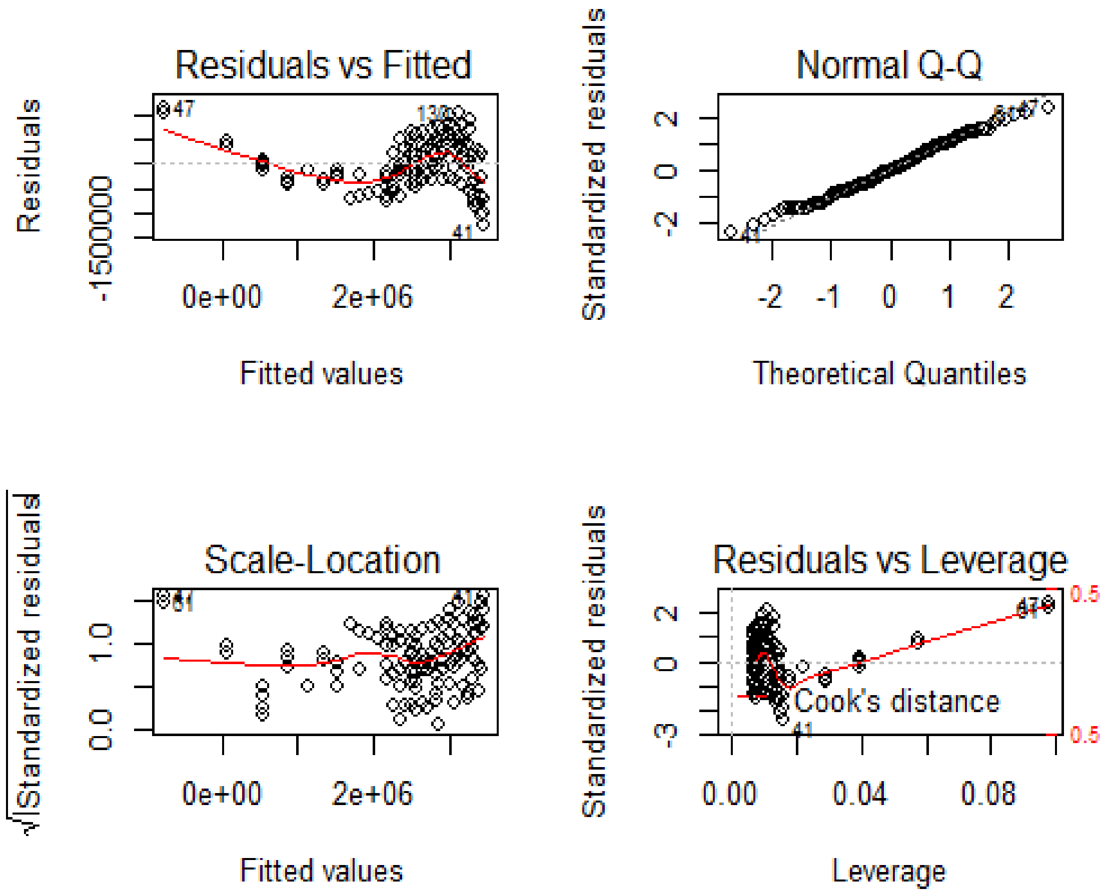


残差有非线性趋势

应用Box-Cox变换（Salary和Experience分别作立方和对数变换）后依旧有非线性现象。

$$\text{Salary}^3 = \beta_0 + \beta_1 \times \log(\text{Experience}) + \varepsilon,$$

Box-Cox 变换是单调变换。非单调的非线性无法用Box-Cox变换消除。



多项式拟合

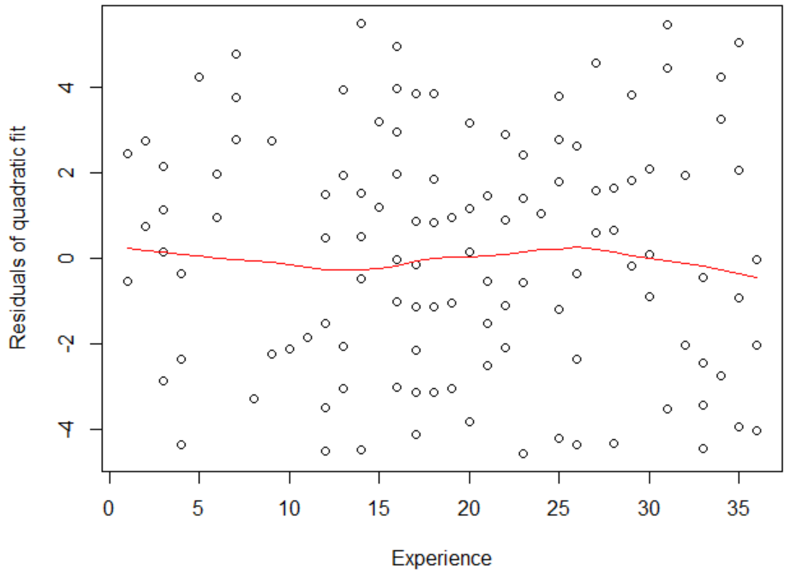
```
> lm(Salary~Experience+ I(Experience^2), data=se)
```

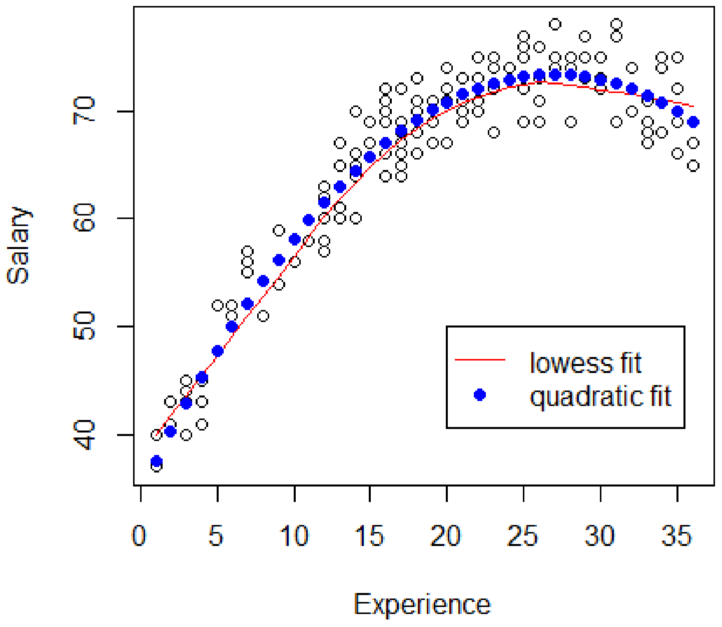
$$\text{Salary} = a + b \times \text{Experience} + c \times \text{Experience}^2 + \varepsilon$$

该模型是两个自变量的线性模型

$$\text{Salary} = a + b \times \text{Experience} + c \times \text{E2} + \varepsilon$$

（两个自变量分别是Experience, $E2 = \text{Experience}^2$）





2. 影响分析

远离数据中心的异常点对回归结果的影响较大，会把回归直线“拉向”该点。影响分析试图发现高影响点。

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

发现高影响点后寻找原因，一般不要轻易删除，而是采用稳健方法。倡导和推广影响分析和稳健统计的主要是John Tukey及其普林斯顿学派。

John Wilder Tukey

John Wilder Tukey (June 16, 1915 – July 26, 2000) 美国著名数学家，最著名的贡献是快速傅里叶变换FFT、Tukey’s lemma、盒型图boxplot 以及若干创造的新词。

1970年, Tukey提出了Jackknife估计； 1974年提出了投影追踪方法(projection pursuit); 1977年出版了《探索数据分析》一书（ Exploratory Data Analysis），强调数据汇总(summary)、可视化、稳健性、影响分析，统计软件的很多方法/函数都与Tukey的倡导有关（summary, boxplot,stem, qqplot, trimmed mean, …）。

Tukey‘s range test, Tukey’s test of additivity, the Tukey lambda distribution and Tukey's lemma.

Tukey‘s range test：单因素方差分析发现多组之间存在显著差异之后，通常需要事后分析(post hoc analysis after anova)，两两比较发现具体哪些组之间差异显著。Tukey的range test就是这样一个方法。

Terms and phrase coined by Tukey

alanysis
alias (in time series)
ANOVA
badmandments
bagplot
batch
bispectrum
bit
biweight
bland distribution
borrowing strength
boxplot
cepstrum
coco
complex demodulation
confimatory data analysis (CDA)
darius
data analysis
dedomulation
defficiency
depth (median of vectors)
dyadic ANOVA
exploratory data analysis (EDA)
faceless value
family of covers
fences
5-number summary
frogs
froots
finite character
Garden of Eden
hamming
(hanging) rootogram
hanning
hat matrix
hinge
Huberizing
jackknife
linear programming
midmean
multihaver
Munkery
polyspectrum
polykay
polysampling
polyspectrum
prewhitening
quefrency
RadGaussianization
rahmonic
regressogram
reroughing
rootogram
rough
running median
saphe cracking
schematic plots
slash distribution
smear-and-sweep
smelting
smoothing and decimation
software
stem-and-leaf
tapering
toolglass
trimming
twicing
vacuum cleaner
vague concept
window carpentry
winsorizing
Winsor's principle
Zorn's Lemma

记号

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

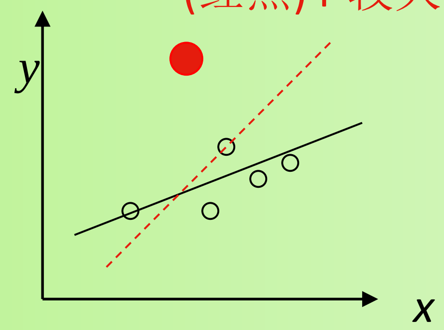
$$\Leftrightarrow y_i=\tilde{\mathbf{x}}_i{}^\top\boldsymbol{\beta}+\varepsilon_i=\beta_0+\mathbf{x}_i{}^\top\mathbf{b}+\varepsilon_i,\quad \varepsilon_i\sim(0,\sigma^2), i=1,\ldots,n$$

$$\text{其中}\mathbf{y}=\begin{pmatrix}y_1\\y_2\\\cdots\\y_n\end{pmatrix},\quad X=\begin{pmatrix}\tilde{\mathbf{x}}_1{}^\top\\\tilde{\mathbf{x}}_2{}^\top\\\cdots\\\tilde{\mathbf{x}}_n{}^\top\end{pmatrix}=\begin{pmatrix}1 & \mathbf{x}_1{}^\top\\1 & \mathbf{x}_2{}^\top\\ & \cdots\\1 & \mathbf{x}_n{}^\top\end{pmatrix}=(\mathbf{1},Z),\ Z=\begin{pmatrix}\mathbf{x}_1{}^\top\\\mathbf{x}_2{}^\top\\\cdots\\\mathbf{x}_n{}^\top\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\\\cdots\\\beta_{p\text{-}1}\end{pmatrix}=\begin{pmatrix}\beta_0\\\mathbf{b}\end{pmatrix}$$

LS估计$\hat{\boldsymbol{\beta}}$，拟合值$\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}}=H\mathbf{y}$，残差$\mathbf{e}=(I-H)\mathbf{y}$，其中$H=P_X$为投影阵，也称为hat matrix（Tukey）:

$$H\mathbf{y}=\widehat{\mathbf{y}}$$

(1) 异常点Outlier：响应y (或残差)异常



outlier判别标准：

$|r_i|$或$|r_i^*|$较大时(比如≥ 3)，

y_i可认为异常。

R

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

因为残差向量$\mathbf{e}=(I_n-H)\mathbf{y}$，$\mathrm{var}(\mathbf{e})=\sigma^2(I_n-H)$，所以

$\mathrm{var}(e_i)=(1-h_{ii})\sigma^2$，其中$H=(h_{ij})_{n\times n}$，$h_{ii}$为$H$的$(i,i)$元。

标准化残差

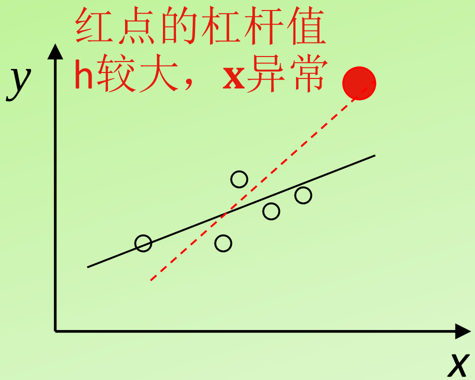
对任何$1\leq i\leq n$，$r_i=\dfrac{e_i}{\hat\sigma\sqrt{1-h_{ii}}}$，其中$\hat\sigma=\sqrt{\sum_{j=1}^n e_j^2/(n-p)}$.

如果y_i异常，$|e_i|$偏大，$\hat\sigma$也偏大，所以估计σ时应不用e_i，

学生化残差

学生化残差：$r_i^*=\dfrac{e_i}{\hat\sigma^{(-i)}\sqrt{1-h_{ii}}}$，其中$\hat\sigma^{(-i)}=\sqrt{\sum_{j\neq i}e_j^2/(n-p-1)}$。

(2) 高杠杆点：自变量 x 异常



自变量异常的判别标准：

若杠杆值$h_{ii} >> \dfrac{p}{n}$,第i个样本的自变量是高杠杆点或高影响的,其中h_{ii}为H的(i,i)元。

杠杆值

投影矩阵/帽子矩阵$H = P_X = X(X^{\top}X)^{-1}X^{\top}$。

帽子矩阵$H = P_X$的(i,i)元h_{ii}称为自变量$\mathbf{x}_i$的杠杆值(leverage)，它表示自变量$\mathbf{x}_i$与$\bar{\mathbf{x}}$之间的马氏距离（命题1）。

高杠杆

因为$H = P_X$对称幂等，所以$tr(H) = rank(H) = rank(X) = p$(假设$X$列满秩)，

$\sum_{i=1}^{n} h_{ii} = tr(H) = p$,平均来看$h_{ii} \sim \dfrac{p}{n}$，远高于该值即认为$\mathbf{x}_i$是高影响的。

由$X=\begin{pmatrix}\tilde{\mathbf{x}}_1^{\top}\\ \vdots \\ \tilde{\mathbf{x}}_n^{\top}\end{pmatrix}=\begin{pmatrix}1 & \mathbf{x}_1^{\top}\\ \vdots & \vdots \\ 1 & \mathbf{x}_n^{\top}\end{pmatrix}=(\mathbf{1},Z)$，知$H=X(X^{\top}X)^{-1}X^{\top}=\begin{pmatrix}\tilde{\mathbf{x}}_1^{\top}\\ \vdots \\ \tilde{\mathbf{x}}_n^{\top}\end{pmatrix}(X^{\top}X)^{-1}\left(\tilde{\mathbf{x}}_1,\dots,\tilde{\mathbf{x}}_n\right)$

所以$h_{ij}=\tilde{\mathbf{x}}_i^{\top}(X^{\top}X)^{-1}\tilde{\mathbf{x}}_j^{\top}$。利用$X=(\mathbf{1},Z)$,我们可得到杠杆值$h_{ii}$更精细的刻画：

命题1. H的第i个对角元 $h_{ii}=\dfrac{1}{n}+d_M^2(\mathbf{x}_i,\bar{\mathbf{x}})/(n-1)$，且$\dfrac{1}{n}\le h_{ii}\le 1$.,

其中$d_M(\mathbf{x}_i,\bar{\mathbf{x}})=\sqrt{(\mathbf{x}_i-\bar{\mathbf{x}})^{\top}S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})}$（称为$\mathbf{x}_i$与$\bar{\mathbf{x}}$的马氏距离），

$S=\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})^{\top}/(n-1)$为样本协方差矩阵。

证明：因为$X=(1,Z)$，$H=P_X=P_{\mathbf{1}}+P_{Z_c}=\dfrac{\mathbf{1}\mathbf{1}^{\top}}{n}+Z_c(Z_c{}^{\top}Z_c)^{-1}Z_c{}^{\top}$,所以

$$\begin{aligned}h_{ii}&=1/n+(\mathbf{x}_i-\bar{\mathbf{x}})^{\top}(Z_c{}^{\top}Z_c)^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})=1/n+(\mathbf{x}_i-\bar{\mathbf{x}})^{\top}\left(\sum(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\top}\right)^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\\&=1/n+(\mathbf{x}_i-\bar{\mathbf{x}})^{\top}[(n-1)S]^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\end{aligned}$$

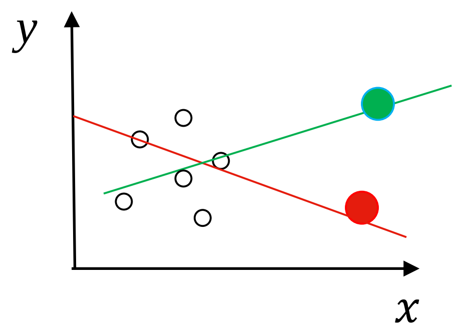
另外,$P_{\mathbf{1}}\le H\le I_n\Rightarrow 1/n\le h_{ii}\le 1$。

$h_{ii} \approx 1$

若$h_{ii} \approx 1$，由命题1知$d_M^2(\mathbf{x}_i, \bar{\mathbf{x}})$较大，$\mathbf{x}_i$远离$\bar{\mathbf{x}}$，

另外，$\mathrm{var}(e_i) = (1-h_{ii})\sigma^2 \approx 0 \Rightarrow e_i \approx 0,\ \hat{y}_i \approx y_i$，

$(\mathbf{x}_i, y_i)$靠近回归直线而且主导回归直线的方向(下图).



$h_{ii} = 1/n$

由命题1知：杠杆值$h_{ii} = 1/n$（最小）$\Leftrightarrow \mathbf{x}_i = \bar{\mathbf{x}}$，

此时 $\hat{y}_i = \hat{\beta}_0 + \hat{\mathbf{b}}^\top \mathbf{x}_i = (\bar{y} - \hat{\mathbf{b}}^\top \bar{\mathbf{x}}) + \hat{\mathbf{b}}^\top \bar{\mathbf{x}} = \bar{y}$.

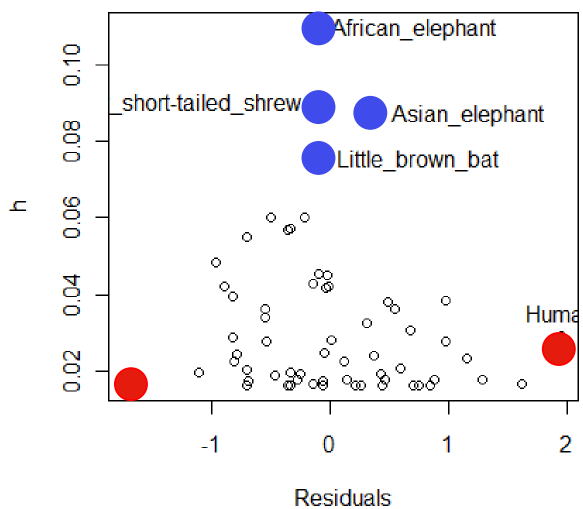
练习：若$(\mathbf{x}_k, y_k) = (\bar{\mathbf{x}}, \bar{y})$，则删除第$k$个样本不改变LS估计。

这是因为关于$\mathbf{b}$的正则方程为$\sum_{i=1}^{n} \mathbf{x}_i (y_i - \bar{y} - (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{b}) = 0$

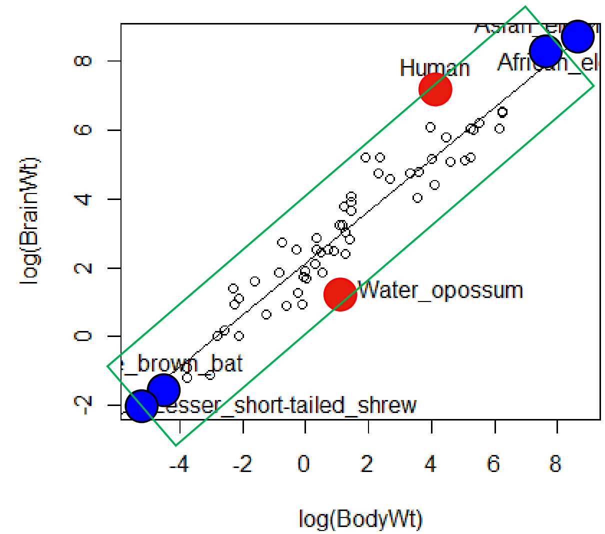
对任何$\mathbf{b}$，第k项为0，不影响正则方程。

例1(续）：动物脑重量与体重的关系 lm.out=lm(log(BrainWt)~log(BodyWt), data=brains)

Residual-h散点图



x-y散点图

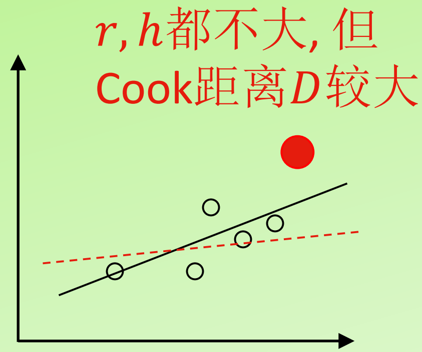


x-异常（高杠杆）

y-异常（outlier）

```
> hatvalues (lm.out)   #杠杆值
> influence.measures(lm.out)  #杠杆值及其它影响度量
```

(3) 高影响点：(x, y) 异常



高影响点的判别标准：

若Cook距离$D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2 > 0.5$,

第i个样本(y_i, z_i)是高影响的.

第i个样本点$(\mathbf{x}_i, y_i)$的自变量和响应都不异常，但它们综合在一起可能是高影响的。如何评估和发现这种高影响？

delete-1、leave-one-out方法评估删除一个样本点后模型拟合效果的变化，如果变换很大，则该样本点（包含自变量和响应）是高影响的。这种“删除一个样本点”的方法也可称为Jackknife （但Jackknife通常专指用于偏差、方差估计）。由此得到的影响度量包括:

Cook’s 距离D，DFBETAS，DFFITS.

对于线性模型$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，我们考察delete-1的影响。为了记号简单，我们这里记X的每一行为 $\mathbf{x}_i$，而不再用记号 $\tilde{\mathbf{x}}_i$

线性模型的delete-1

	完整数据	删除第i行数据	差异 (DF)	影响度量
数据	$\mathbf{y} = (y_1, \dots, y_n)^\top$ $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$	$\mathbf{y}_{-i} = (\dots, y_{i-1}, y_{i+1} \dots)^\top$ $X_{-i} = (\dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1} \dots)^\top$	$y_i, \mathbf{x}_i$	
模型	$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\mathbf{y}_{-i} = X_{-i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{-i}$		
LS估计	$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$	$\widetilde{\boldsymbol{\beta}} = (X_{-i}^\top X_{-i})^{-1} X_{-i}^\top \mathbf{y}_{-i}$	$\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$	DFBETAS
y_i的拟合	$\hat{y}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$	$\tilde{y}_i = \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}$ (预测)	$\hat{y}_i - \tilde{y}_i$	DFFITs
所有拟合	$\hat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$	$\tilde{\mathbf{y}} = X\widetilde{\boldsymbol{\beta}}$,	$\|\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|\|^2$	Cook's D

注意：$\widetilde{\boldsymbol{\beta}}$ 不依赖于删除的数据$(y_i, \mathbf{x}_i)$，所以$\tilde{y}_i = \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}$ 不是通常的拟合值，而应称为预测。

我们主要讨论Cook's D

引理1.（Sherman - Morrison - Woorbury公式的特殊情况）

假设$A_{n\times n}$可逆，对任何$n\times 1$向量$\mathbf{u},\mathbf{v}$, 若$A-\mathbf{u}\mathbf{v}^\top$可逆，则有

$$(A-\mathbf{u}\mathbf{v}^\top)^{-1}=A^{-1}+\frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1-\mathbf{v}^\top A^{-1}\mathbf{u}},$$

特别地$(I_n-\mathbf{u}\mathbf{v}^\top)^{-1}=I_n+\mathbf{u}\mathbf{v}^\top/(1-\mathbf{v}^\top\mathbf{u}).$

附：Sherman-Morrison-Woodbury公式:

假设U,V是$n\times k$矩阵，若$A_{n\times n}A, A-UV^\top$可逆，则

$$(A_{n\times n}-UV^\top)^{-1}=A^{-1}+A^{-1}U(I_k-V^\top A^{-1}U)^{-1}V^\top A^{-1}$$

特别地

$$(I_n-UV^\top)^{-1}=I_n+U(I_k-V^\top U)^{-1}V^\top$$

左端n阶矩阵求逆转换为右端k阶求逆

注：行列式类似的结果 $\det(A-UV^\top)=\det(A)\det(I_k-V^\top A^{-1}U)$

应用：当n很大时，n阶矩阵的求逆非常耗时。如果对不同的(U_i,V_i)，$i=1,2,\ldots$，需要计算$n\times n$矩阵的逆$\left(A-U_iV_i^T\right)^{-1}$，SMW公式说明，若$k<<n$, 我们只需要计算一次大矩阵的逆$A^{-1}$以及若干$k$阶小矩阵的逆.

命题2. 假设模型$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 I_n), X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, LS估计为$\widehat{\boldsymbol{\beta}}$，残差为$e_i, i = 1, \dots, n$, 残差平方和为$RSS = \|\mathbf{y} - X\widehat{\boldsymbol{\beta}}\|^2 = \sum e_i^2$。删除第 i 个数据点$y_i, \mathbf{x}_i$ 后拟合模型$\mathbf{y}_{-i} = X_{-i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{-i}, \boldsymbol{\varepsilon}_{-i} \sim (\mathbf{0}, \sigma^2 I_{n-1})$，其LS估计和残差平方和分别记作 $\widetilde{\boldsymbol{\beta}} = (X_{-i}^\top X_{-i})^{-1} X_{-i}^\top \mathbf{y}_{-i}$和$\widetilde{RSS} = \|\mathbf{y}_{-i} - X_{-i}\widetilde{\boldsymbol{\beta}}\|^2$, 则

(1) $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - (X^\top X)^{-1}\mathbf{x}_i e_i/(1 - h_{ii})$

(2) $\widetilde{RSS} = RSS - e_i^2/(1 - h_{ii})$

$$\mathbf{y}_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)^\top,$$
$$X_{-i} = (\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)^\top$$

命题2说明:

- 删除一个样本点后的LS估计由完整数据的LS估计和残差决定，不必重新拟合。
- $\widetilde{RSS} \le RSS$，若删除样本点 i ，则残差平方和减小。若该点的杠杆值$h_{ii} \to 1$（影响大），则残差平方和通常会大幅度减小（除非$e_i = 0$）。
- 若$e_i = 0$，删除数据点 i 对最小二乘没有任何影响。这也可以从正则方程看出：$0 = \sum_{k=1}^n \mathbf{x}_k(\mathbf{y}_k - \mathbf{x}_k^\top\widehat{\boldsymbol{\beta}}) = \sum_{k\neq i} \mathbf{x}_k(\mathbf{y}_k - \mathbf{x}_k^\top\widehat{\boldsymbol{\beta}})$

$$\underline{\text{命题2的证明:}}\quad (1)\quad X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\top},\qquad X^{\top}X=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\top},\ X^{\top}\mathbf{y}=\sum_{i=1}^{n}\mathbf{x}_i y_i$$

$$\Rightarrow X^{\top}X=X_{-i}^{\top}X_{-i}+\mathbf{x}_i\mathbf{x}_i^{\top},\quad X^{\top}\mathbf{y}=\sum_{j=1}^{n}\mathbf{x}_j y_j=X_{-i}^{\top}\mathbf{y}_{-i}+\mathbf{x}_i y_i$$

$$\text{注意}h_{ii}=\mathbf{x}_i^{\top}(X^{\top}X)^{-1}\mathbf{x}_i$$

$$\text{由引理1,}\ (X_{-i}^{\top}X_{-i})^{-1}=(X^{\top}X-\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}=(X^{\top}X)^{-1}+\frac{(X^{\top}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\top}(X^{\top}X)^{-1}}{1-\mathbf{x}_i^{\top}(X^{\top}X)^{-1}\mathbf{x}_i},$$

$$\Rightarrow\tilde{\boldsymbol{\beta}}=\left(X_{-i}^{\top}X_{-i}\right)^{-1}X_{-i}^{\top}\mathbf{y}_{-i}=\left((X^{\top}X)^{-1}+\frac{(X^{\top}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\top}(X^{\top}X)^{-1}}{1-\mathbf{x}_i^{\top}(X^{\top}X)^{-1}\mathbf{x}_i}\right)\left(X^{\top}\mathbf{y}-\mathbf{x}_i y_i\right)$$

$$=\hat{\boldsymbol{\beta}}-(X^{\top}X)^{-1}\mathbf{x}_i y_i+\frac{(X^{\top}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\top}(X^{\top}X)^{-1}X^{\top}\mathbf{y}-(X^{\top}X)^{-1}\mathbf{x}_i\underbrace{\mathbf{x}_i^{\top}(X^{\top}X)^{-1}\mathbf{x}_i}_{h_{ii}}y_i}{1-h_{ii}}$$

$$=\hat{\boldsymbol{\beta}}-\frac{(X^{\top}X)^{-1}\mathbf{x}_i\left(y_i-\mathbf{x}_i^{\top}\hat{\boldsymbol{\beta}}\right)}{1-h_{ii}}=\hat{\boldsymbol{\beta}}-\frac{(X^{\top}X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$$

(2) 由(1)，

$$\widetilde{RSS} = \left\|\mathbf{y}_{-i} - X_{-i}\widetilde{\boldsymbol{\beta}}\right\|^2 = \left\|\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}} + X_{-i}(X^\top X)^{-1}\mathbf{x}_i e_i/(1-h_{ii})\right\|^2$$

$$\triangleq \left\|\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}} + \mathbf{a}\right\|^2 = \left\|\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}}\right\|^2 + \mathbf{a}^\top\mathbf{a} + 2\mathbf{a}^\top(\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}})$$

显然，第一项 $\left\|\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}}\right\|^2 = \sum_{k\neq i} e_k^2 = RSS - e_i^2$，

第二项

$$\mathbf{a}^\top\mathbf{a} = e_i^2\mathbf{x}_i^\top(X^\top X)^{-1}\left[X_{-i}^\top X_{-i}\right](X^\top X)^{-1}\mathbf{x}_i/(1-h_{ii})^2$$

$$= e_i^2\mathbf{x}_i^\top(X^\top X)^{-1}\left[X^\top X - \mathbf{x}_i\mathbf{x}_i^\top\right](X^\top X)^{-1}\mathbf{x}_i/(1-h_{ii})^2$$

$$= e_i^2 h_{ii}/(1-h_{ii})$$

类似地，$2\mathbf{a}^\top\left(\mathbf{y}_{-i} - X_{-i}\widehat{\boldsymbol{\beta}}\right) = -2e_i^2 h_{ii}/(1-h_{ii})$

所以 $\widetilde{RSS} = RSS - e_i^2 - e_i^2 h_{ii}/(1-h_{ii}) \;\; = RSS - e_i^2/(1-h_{ii})$.

(3) 由(1), 预测值$\tilde{y}_i = \mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}} = \mathbf{x}_i^\top(\widehat{\boldsymbol{\beta}} - (X^\top X)^{-1}\mathbf{x}_i e_i/(1-h_{ii})$
$= \hat{y}_i - e_i h_{ii}\,/(1-h_{ii})$

Cook距离

定义：Cook距离 $D_i = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{p\hat{\sigma}^2} = \frac{\|X\hat{\boldsymbol{\beta}} - X\tilde{\boldsymbol{\beta}}\|^2}{p\hat{\sigma}^2}$，

若$D_i > 0.5$，样本点i有较大影响；若 $D_i > 0.5$，样本点i有很大影响。

注：因为$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1}$，所以$\mathrm{D}_i = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top \hat{\sigma}^{-2} (X^\top X)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) / p$

可理解为$\hat{\boldsymbol{\beta}}$与$\tilde{\boldsymbol{\beta}}$的平均马氏距离。

命题3：对任何$1 \le i \le n$，$D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$，其中$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$为标准化残差（该表达表明cook距离综合考虑了$h$和$r$）。

证明：由命题2⑴，$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (X^\top X)^{-1} \mathbf{x}_i e_i / (1-h_{ii})$，

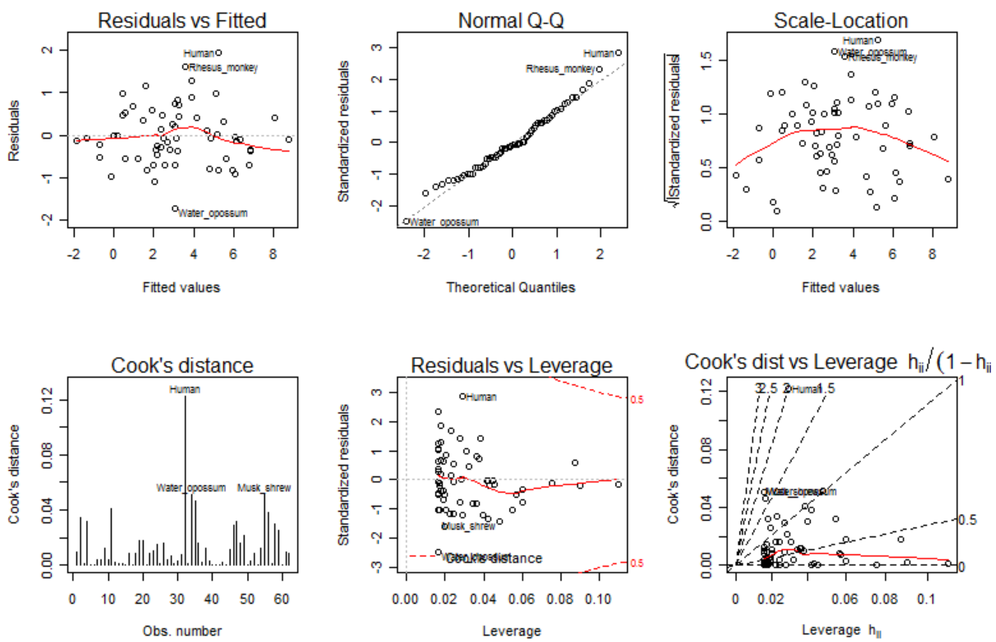
所以 $\hat{\mathbf{y}} - \tilde{\mathbf{y}} = X\hat{\boldsymbol{\beta}} - X\tilde{\boldsymbol{\beta}} = \frac{X(X^\top X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}} \Rightarrow D_i = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{p\hat{\sigma}^2} = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$

其它影响度量：$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \tilde{y}_i}{\tilde{\sigma}\sqrt{h_{ii}}}$，若$|DFFITS_i| \ge 2\sqrt{\frac{p}{n}}$，样本点$i$高影响

$\mathrm{DFBETAS}_i(k) = \frac{\hat{\beta}_k - \tilde{\beta}_k}{\tilde{\sigma}\sqrt{\left((X^\top X)^{-1}\right)_{kk}}}$，$1 \le k \le p$. 若$|DFBETAS_i(k)| \ge \frac{2}{\sqrt{n}}$，$i$高影响

例1(续)：动物脑重量与体重的关系， R 回归诊断图（残差分析+影响分析)

```
> myfit =lm( BrainWt ~ BodyWt, data=log(brains)
> plot(myfit, which=1:6)      #default: which=c(1,2,3,5)
```



6个图分别为

(1) 残差图：线性? 等方差?

(2) qqnorm: 误差正态?

(3) 刻度-位置图(残差图的补充)： 线性？等方差？

(4) Cook‘s D: 影响分析

(5) 残差-杠杆图：影响分析

(6) D vs h: 影响分析

红色实线为非线性拟合（lowess ），红色虚线为cook距离D-等高线(D=0.5, D=1).

例5. 判断正误：
(1) 若删除一个样本点，则残差平方和一定减小或不变（命题2）。
(2) 若回归模型中删除一个变量，则残差平方和一定增加或不变。
(3) 若回归模型中增加一个变量，则决定系数一定增加（或不变）。

(1) 实际上更简单地，从最小二乘法目标函数容易看出

$$RSS_n = min\sum_{i=1}^{n}(y_i - \tilde{\mathbf{x}}_i^{\mathrm{T}}\boldsymbol{\beta})^2 \geq min\sum_{i=1}^{n-1}(y_i - \tilde{\mathbf{x}}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = RSS_{n-1}$$

即删除一个样本点后残差平方和减小。

(2) $X = (X_1, X_2), \boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}$

$$\begin{aligned}\mathrm{RSS}(X_1, X_2) &= min_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} ||\mathbf{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2||^2\\ &\leq min_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2=\mathbf{0}} ||\mathbf{y} - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2||^2\\ &= min_{\boldsymbol{\beta}_1} ||\mathbf{y} - X_1\boldsymbol{\beta}_1||^2 = \mathrm{RSS}(X_1)\end{aligned}$$

(3) 因为$R^2 = 1 - \mathrm{RSS}/SS_{总}$， 故添加自变量个数时，RSS减少， 决定系数增加。

思考：

为什么删除一个数据点后的拟合可由完整模型的拟合决定？删除两个点是否也有类似结论？

附录1：Jackknife 方法简介

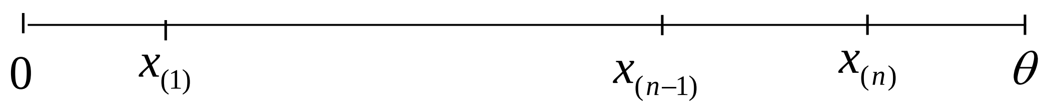
Jackknife
便携折刀

Invented by Quenouille(1949), named by Tukey (1958).
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(万能) tool.

例A1(估计上界)：$x_1,\dots,x_n\ iid \sim U(0,\theta)$，记样本从小到大排列为

$$x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$$

显然极大似然估计$\hat{\theta}_1 = x_{(n)} = \max(x_i)$低估了$\theta$。$E(\hat{\theta}_1) = \dfrac{n}{n+1}\theta = \theta - \dfrac{\theta}{n+1}$

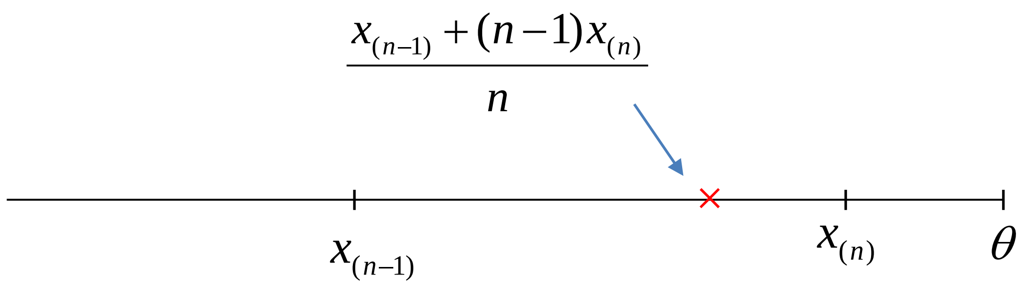


<u>直观解法</u>：间隔$[0,x_{(1)}],[x_{(1)},x_{(2)}],\dots,[x_{(n)},\theta]$期望长度相同，前$n$个区间的平均长度：$\bar{d} = \left(x_{(1)} + (x_{(2)} - x_{(1)}) + \dots + (x_{(n)} - x_{(n-1)})\right)/n = x_{(n)}/n$，最后一个区间的长度$\theta - x_{(n)}$应该约等于$\bar{d}$：$\theta - x_{(n)} = \bar{d} \Rightarrow \hat{\theta}_2 = \dfrac{n+1}{n} x_{(n)}$.

例A2. 从例A1直观解法知为了估计θ，估计最后一个间隔可能是关键：

$$d_n = \theta - x_{(n)} \tag{1}$$

假如$x_1,\ldots,x_n$中某一个点没有采集到,那么$n-1$个x的最大值有可能是$x_{(n-1)}$，也可能是$x_{(n)}$,概率分别是$1/n$ 和$1-1/n$. 平均看，$n-1$个数据点的最大值为



它与上界θ的距离

$$d_{n-1} = \theta - \frac{x_{(n-1)}+(n-1)x_{(n)}}{n}, \tag{2}$$

上述d_n和d_{n-1}分别是样本数目为n和$n-1$的时候，最后一个区间的长度。

令$(n-1)d_{n-1} = nd_n \Rightarrow \hat{\theta}_3 = 2x_{(n)} - x_{(n-1)} - \frac{1}{n}\left(x_{(n)} - x_{(n-1)}\right)$,

称为Jackknife估计，其偏差小于$\hat{\theta}_1$，$\hat{\theta}_2$的偏差。

Jackknife用于校正偏差 (Quenouille,1949)

问题及假设：给定基于样本$x_1,\ldots,x_n$的θ的估计$\hat{\theta}$, 假设其偏差$b_n=c/n$，即

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

我们希望估计偏差b_n,并校正$\hat{\theta}$.

Jackknife估计

记基于删除数据点i得到θ的估计为$\hat{\theta}^{(-i)}, i=1,\ldots,n, \bar{\theta}=\sum_{i=1}^{n}\hat{\theta}^{(-i)}/n,$

则偏差b的估计为$(n-1)(\bar{\theta}-\hat{\theta})$，校正偏差后的Jackknife估计为

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

证：根据偏差假设，基于$n-1$个数据点的估计$\hat{\theta}^{(-i)}$的偏差为$\dfrac{c}{n-1}$:

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

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

即 $(n-1)E(\bar{\theta}-\hat{\theta})=b_n$， 所以$(n-1)(\bar{\theta}-\hat{\theta})$是偏差$b_n$的无偏估计.

从而 $\hat{\theta}_{\text{Jackknife}}=\hat{\theta}-(n-1)(\bar{\theta}-\hat{\theta})$是的无偏估计。

例A2中，$\hat{\theta}_1 = x_{(n)}, E(\hat{\theta}) = \theta + \dfrac{\theta}{n+1}$, 我们希望用*Jackknife*校正 $\hat{\theta}_1$的偏差.

$$\hat{\theta}^{(-i)} = \text{删除}x_i\text{后的最大值} = \max_{j \neq i} x_j = \begin{cases} x_{(n)} & \text{若} x_i \neq x_{(n)} \\ x_{(n-1)} & x_i = x_{(n)} \end{cases}$$

令$\bar{\theta} = \sum_{i=1}^{n} \hat{\theta}^{(-i)} / n = ((n-1)x_{(n)} + x_{(n-1)}) / n$, 此即例2中对$n-1$个样本点的最大值的预期值，则$\bar{\theta} - \hat{\theta} = (x_{(n-1)} - x_{(n)}) / n.$

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

$= 2x_{(n)} - x_{(n-1)} - \dfrac{1}{n}\left(x_{(n)} - x_{(n-1)}\right)$,与例A2得到的结果相同。

Jackknife用于估计方差 (Tukey, 1958)

*Jacknife*的另一个主要用途是计算复杂统计量（比如最大统计量$\max(x_i)$）的方差。

未知参数θ，假设$\hat{\theta}$是基于简单样本$x_1,\ldots,x_n$的θ的一个估计$\hat{\theta}$。目标：估计$\mathrm{var}(\hat{\theta})$。

记$\hat{\theta}^{(-i)}$为删除x_i得到的θ的估计，记$\hat{\theta}^{(-i)}$，$i=1,\ldots,n$,的样本均值和样本方差

$$\bar{\theta}_{Jack}=\sum_{i=1}^{n}\hat{\theta}^{(-i)}/n,\quad s_{Jack}^2=\sum_{i=1}^{n}\left(\hat{\theta}^{(-i)}-\bar{\theta}\right)^2/(n-1),$$

则$\hat{\theta}$的方差的*Jackknife*估计：$\widehat{\mathrm{var}}(\hat{\theta})=\dfrac{(n-1)^2}{n}s_{Jack}^2=\dfrac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^{(-i)}-\bar{\theta}\right)^2$

例A3. 若θ是总体均值，σ^2是总体方差，即$E(x_i)=\theta,\mathrm{var}(x_i)=\sigma^2$.

假设通常的估计：$\hat{\theta}=\bar{x}$，我们已知$\mathrm{var}(\hat{\theta})=\sigma^2/n$，其估计$\mathrm{var}(\hat{\theta})=\dfrac{s_x^2}{n}$。

我们下面验证Jackknife方法得到的也是该估计。

$$\hat{\theta}=\bar{x},\ 则\ \hat{\theta}^{(-i)}=(x_1+\ldots+x_{i-1}+x_{i+1}+\ldots x_n)/(n-1)=\frac{n\bar{x}-x_i}{n-1},$$

则容易验证$\hat{\theta}^{(-i)}$，$i=1,\ldots,n$的样本均值$\bar{\theta}_{\mathrm{Jack}}=\bar{x}$，样本方差$s_{\mathrm{Jack}}^2=\dfrac{1}{(n-1)^3}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$

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

Jackknife与Bootstrap

例A2中随机删除一个样本点，即随机抽取$n-1$个点会影响或改变最大统计量，利用这种改变，我们可以评估最大统计量的偏差。类似地，如果从n个数据点$x_1,\ldots,x_n$随机有放回地抽取n个点，同样会改变最大统计量，进而为估计最大统计量的偏差或其他性质提供证据。这称为Bootstrap自助法。

假设样本$x_1,\ldots,x_n$,参数θ的估计为$\hat{\theta}$。从$x_1,\ldots,x_n$中有放回地抽取n个数据点，得θ的*Bootstrap*估计$\hat{\theta}^*$，反复抽样B次，得到$\hat{\theta}_1^*,\hat{\theta}_2^*,\ldots,\hat{\theta}_B^*$,它们是$\hat{\theta}$的再抽样版本, 可用来估计$\hat{\theta}$的分布或相关的量。

附录2：方差稳定化变换

正态分布$N(\mu,\sigma^2)$的方差与均值是两个自由参数，但基于中心极限定理的渐近正态分布的方差和均值可能有关，比如

假设伯努利变量$x_1,\dots,x_n \ iid \sim B(1,p)$, 由CLT，$\dfrac{\sqrt{n}(\bar{x}-p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$,

当$np>5$时,近似地 $\bar{x} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$

假设$X \sim Pois(\lambda)$泊松分布, 则$\dfrac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1),\ \ \lambda \to \infty$

当$\lambda>5$时，近似地$X \sim N(\lambda,\lambda)$.

Delta方法

引理(Delta方法). 若$\sqrt{n}(X_n-\theta) \xrightarrow{d} N(0,\sigma^2(\theta)),\ \ n\to\infty$, 假设$g'(\theta)$存在且非0, 则$\sqrt{n}(g(X_n)-g(\theta)) \xrightarrow{d} N(0,[g'(\theta)]^2\sigma^2(\theta))$.
(多元情形类似)

证明：泰勒展开，近似地：$\sqrt{n}(g(X_n)-g(\theta)) \approx \sqrt{n}g'(\theta)(X_n-\theta)$。

方差稳定化变换

设$\theta = E(y)$，$\sigma^2 = \sigma^2(\theta) = \mathrm{var}(y)$，$y$的方差稳定化变换定义为

$$g(y) = \int_0^y \frac{1}{\sigma(\theta)} d\theta.$$

由上页引理，我们寻求方差稳定化变换g，使得变换后$g(X_n)$的方差与θ无关，即

$$[g'(\theta)]^2\sigma^2(\theta) = C\text{（常数）} \Rightarrow g(\theta) \propto \int \frac{1}{\sigma(\theta)} d\theta\text{。}$$

例A4. 我们已知下述事实：

假设$(x_1, y_1),\ldots,(x_n, y_n)$ iid ~ 二元正态，设ρ为总体相关系数，r_n为样本相关系数，则 $\sqrt{n}\,(r_n - \rho) \xrightarrow{d} N\left(0, (1-\rho^2)^2\right)$，当$n \to \infty$.

方差稳定化变换（称为Fisher's z-变换）：

$$\mathrm{g}(r_n) \propto \frac{1}{2}\int_0^{r_n} \frac{1}{(1-\rho^2)} d\rho = \frac{1}{2}\log\left(\frac{1+r_n}{1-r_n}\right) = \mathrm{atanh}(r_n).$$

$$\sqrt{n}\left[\mathrm{atanh}(r_n) - \mathrm{atanh}(\rho)\right] \xrightarrow{d} N(0,1),$$

例A5 比率数据(proportion data) 设$x \sim \mathrm{B}(n, p)$, $\hat{p} = x/n$,则

$$\mathrm{E}(\hat{p}) = p，\sigma^2 = \mathrm{var}(\hat{p}) = p(1-p)/n.$$

方差稳定化变换：

$$\mathrm{g}(x) \propto \int_0^x \frac{1}{\sigma(p)} dp = \int_0^x \frac{1}{\sqrt{p(1-p)}} dp = \arcsin\sqrt{p}$$

略去常数2，可取$\hat{p}$的方差稳定化变换为$\arcsin\sqrt{\hat{p}}$。

但更为常用的是logit变换：

$$p \to \mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

p为概率，$\dfrac{p}{1-p}$称为*odds.*