

课程主页: <http://staff.ustc.edu.cn/~ynyang/2022>

第十六讲 回归诊断

2022.12.16

Tukey's Jackknife



基于线性回归分析结果，对线性模型的恰当性进行评判或诊断，并发现异常值等对回归结果影响过大的数据点。回归诊断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

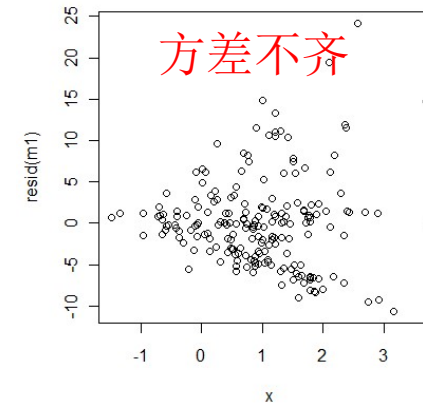
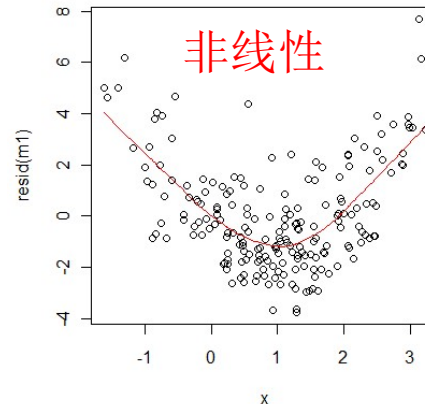
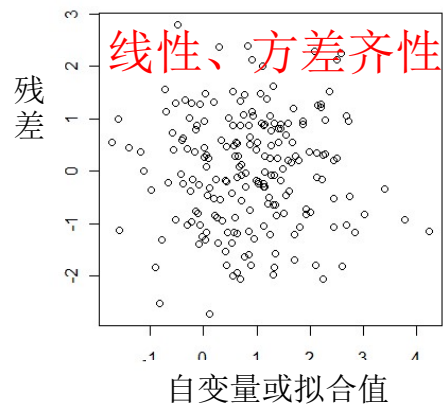
- R^2 或回归方程的显著性检验称为拟合优度量，是一种整体拟合程度的度量。 R^2 不足以发现细节上的模型的不合理性。
- 回归诊断提供了一些常用工具（残差图、标准化残差、杠杆值、Cook距离等），实际数据分析中不应局限于这些工具。

1. 残差分析

残差图

残差图：横轴为自变量，纵轴为残差的散点图。

线性模型假设误差与自变量无关，故自变量-残差散点图上应没有任何明显的趋势或相关关系，如下左图。当有多个自变量时，残差图的横轴可取为拟合值（拟合值是所有自变量的最优组合）。



应对策略

如果残差图表明线性模型假设不合理：

- 非线性（中图）：变量变换（Box-Cox变换或对数变换）、添加高阶项、非线性回归或广义线性回归模型(glm)
- 方差不齐（右图）：变量变换、GLS、glm

成语 Box and Cox: 轮流、互斥、小人物

George E. P. Box

G.E.P.Box (born 18 October 1919) is a statistician, who has made important contributions in the areas of quality control, time-series analysis, design of experiments.

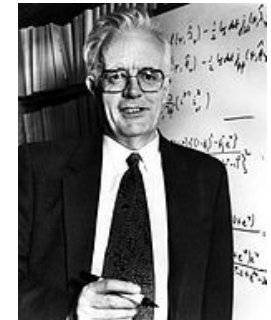
- Son-in-law of Fisher.
- Founder of Dept Stat, University of Wisconsin–Madison(1960).
- Contributions: Box–Cox transformations, Box-Jenkins models, Box–Behnken designs, robust statistics, etc.



Sir David Roxbee Cox

Sir D.R. Cox (born 1924, Birmingham, England) is a prominent British statistician. He worked in the Imperial College London, Oxford University.

- Proportional hazards model (Cox model), survival data.
- 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假设成立。

变换之后的变量的分布对称、均衡，接近正态。

基于此, Box - Cox变换的基本思想是：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。

Box-Cox 变换

Box-Cox变换：

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

选择 λ , 使得 $y^{(\lambda)}$ 的分布近似为 正态分布。

注1：具体实施的时候，可忽略分母 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2：最常用的是对数变换。

注3：BC变换是单调变换。

剖面似然法：
求解幂次 λ

似然函数
 $L = Pr(data)$

数据： $(y_i, x_i), i = 1, \dots, n$. 对某个 $\lambda \in R$, 响应变量做Box-Cox变换：

$$\mathbf{y} = (y_1, \dots, y_n)^\top \rightarrow \mathbf{y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^\top$$

求解 λ 变换, 使得变换后 $\mathbf{y}^{(\lambda)}$ 满足正态模型：

$$\mathbf{y}^{(\lambda)} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n) \Leftrightarrow \mathbf{y}^{(\lambda)} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n).$$

解：对于给定的 λ , $\mathbf{y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^\top$ 的联合密度函数为

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

所以似然函数, 即数据 $\mathbf{y} = (y_1, \dots, y_n)^\top$ 的联合密度为：

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

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

同时求解 $\lambda, \boldsymbol{\beta}, \sigma^2$ 比较困难。但 λ 给定时, $\boldsymbol{\beta}, \sigma^2$ 的极大似然估计容易求解：

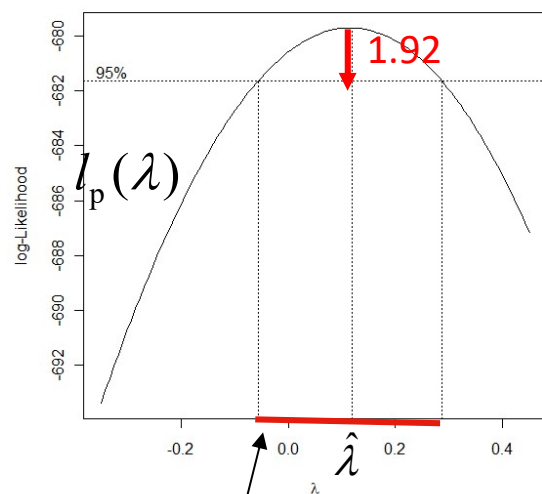
$$\hat{\boldsymbol{\beta}}(\lambda) = (X^\top X)^{-1} X^\top \mathbf{y}^{(\lambda)}, \quad \hat{\sigma}^2(\lambda) = \|\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}}(\lambda)\|^2 / n \triangleq RSS(\lambda) / n,$$

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

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

极大化无显式解。逐点搜索计算 $l_p(\lambda)$ ， $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} l_p(\lambda)$ 为最优 λ ，如下图。

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 $\lambda = 0.5$ 。



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。

$$1.92 = 3.84/2$$

$$3.84 = 1.96^2$$

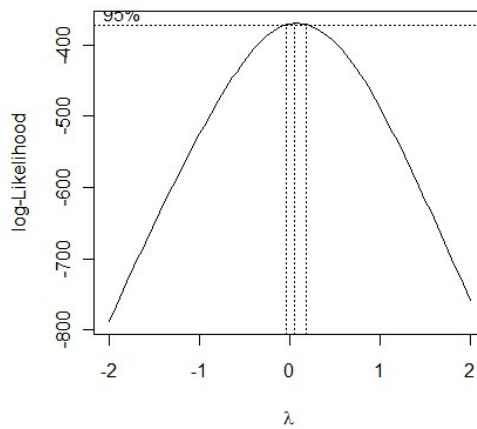
实际应用中，对自变量也可做Box-Cox变换

例1. 62种哺乳动物的脑重量与体重数据。

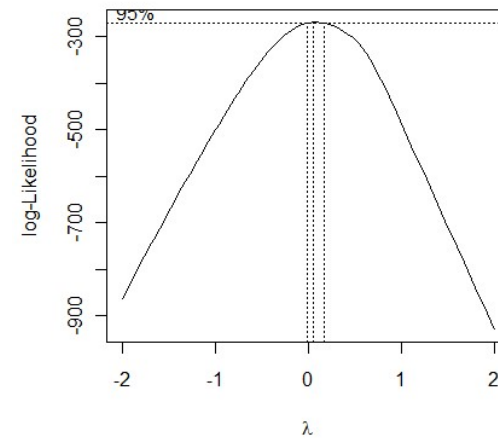
R: library(MASS)中的boxcox函数

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

其它变换

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）

附：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 $x_1, \dots, x_n \text{ iid} \sim B(1, p)$, 由CLT, $\frac{\sqrt{n}(\bar{x} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$,

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

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

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

Delta方法

引理(Delta方法). 若 $\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$, $n \rightarrow \infty$, 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 $\theta = E(y)$, $\sigma^2 = \sigma^2(\theta) = \text{var}(y)$, y 的方差稳定化变换定义为

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

解: 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

$$[g'(\theta)]^2 \sigma^2(\theta) = C \text{ (常数)}$$

$$\Rightarrow g(\theta) \propto \int \frac{1}{\sigma(\theta)} d\theta.$$

例2. 我们已知下述事实 (第2讲命题2):

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

方差稳定化变换 (称为Fisher's z-变换):

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

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

例3 比率数据(proportion data) 设 $x \sim B(n, p)$, $\hat{p} = x/n$, 则

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

方差稳定化变换:

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

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:

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

p 为概率, $\frac{p}{1-p}$ 称为 *odds*.

2. 影响分析

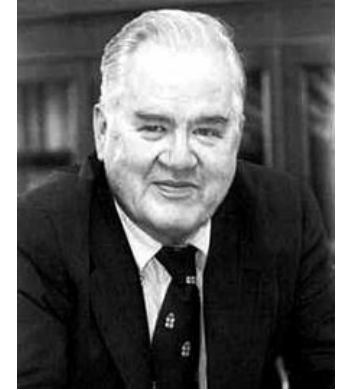
Tukey's quote: Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

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，除非是数据录入错误，否则一般不要删除，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) was an American mathematician best known for development of the **FFT algorithm** and **box plot**.

In 1970, he contributed significantly to what is today known as the **jackknife** estimation. In 1974, he developed, with Jerome H. Friedman, the concept of the **projection pursuit**. In 1977, he published a book, **Exploratory Data Analysis**, in which he emphasized the role of **data summary, visualization, robustness, influence analysis**. He is now regarded as a pioneer of data science.

Tukey's range test, Tukey's test of additivity, the Tukey lambda distribution and Tukey's lemma all bear his name.

He coined many terms (next page).

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

analysis

alias (in time series)

ANOVA

bandwidths

bagplot

batch

bispectrum

bit

biweight

blind distribution

borrowing strength

boxplot

cepstrum

coco

complex demodulation

confirmatory data analysis (CDA)

darius

data analysis

demodulation

efficiency

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, H

hinge

Huberizing

jackknife

linear programming

midmean

multihaver

Munkery

polyspectrum

polykay

polysampling

polyspectrum

prewhitening

quefrency

RadGaussianization

harmonic

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 $\mathbf{1}$ 。

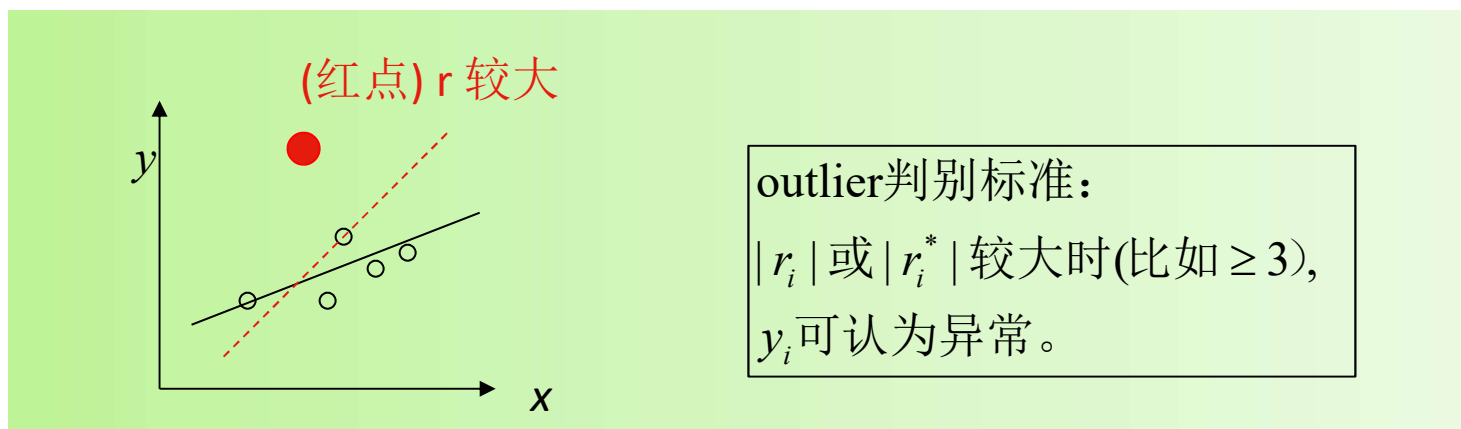
$\Leftrightarrow y_i = \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} + \varepsilon_i = \beta_0 + \mathbf{x}_i^\top \mathbf{b} + \varepsilon_i$, $\varepsilon_i \sim (0, \sigma^2), i = 1, \dots, n$

其中

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

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

(1) 异常点Outlier: 响应y (或残差)异常



$\text{var}(e_i) = (1 - h_{ii})\sigma^2$, 其中 $H = P_X = (h_{ij})_{n \times n}$ 为帽子矩阵 / 投影矩阵。

这是因为残差向量 $\mathbf{e} = (I_n - H)\mathbf{y}$, $\text{var}(\mathbf{e}) = \sigma^2(I_n - H)$ 。

标准化
残差

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

学生化
残差

如果 y_i 异常, $|e_i|$ 偏大, $\hat{\sigma}$ 也偏大, 所以估计 σ 时不用 e_i ,

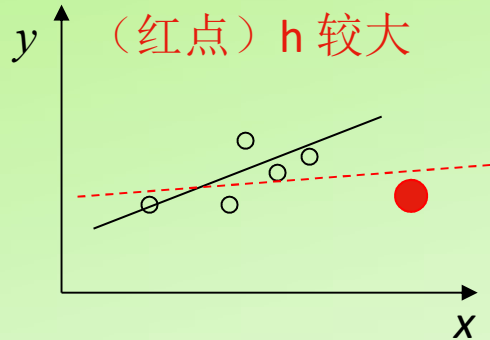
定义学生化残差: $r_i^* = \frac{e_i}{\hat{\sigma}^{(-i)} \sqrt{1 - h_{ii}}}$, 其中 $\hat{\sigma}^{(-i)} = \sqrt{\sum_{j \neq i} e_j^2 / (n - p - 1)}$,

R

> rstandard(lm.out)

> rstudent(lm.out)

(2) 高杠杆点: 自变量 \mathbf{x} 异常



自变量异常的判别标准:

若 $h_{ii} \gg \frac{p}{n}$, 第 i 个样本的自变量是高杠杆点或高影响的, 其中 h_{ii} 为 $H = P_X$ 的 (i, i) 元。

杠杆值

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

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

命题1. H 的第 i 个对角元 $h_{ii} = \frac{1}{n} + d_M^2(\mathbf{x}_i, \bar{\mathbf{x}})/(n-1)$, 且 $\frac{1}{n} \leq h_{ii} \leq 1$,
 其中 $d_M(\mathbf{x}_i, \bar{\mathbf{x}}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})^\top S^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})}$ 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: $X = \begin{pmatrix} \tilde{\mathbf{x}}_1^\top \\ \vdots \\ \tilde{\mathbf{x}}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^\top \end{pmatrix} = (\mathbf{1}, Z), P_X = X(X^\top X)^{-1}X^\top = \begin{pmatrix} \tilde{\mathbf{x}}_1^\top \\ \vdots \\ \tilde{\mathbf{x}}_n^\top \end{pmatrix} (X^\top X)^{-1} (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)$

的 (i, j) 元为 $h_{ij} = \tilde{\mathbf{x}}_i^\top (X^\top X)^{-1} \tilde{\mathbf{x}}_j^\top$ 。

因为 $X = (\mathbf{1}, Z), P_X = P_1 + P_{Z_c} = \frac{\mathbf{1}\mathbf{1}^\top}{n} + Z_c(Z_c^\top Z_c)^{-1}Z_c^\top$,

所以 $h_{ii} = 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^\top (Z_c^\top Z_c)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$
 $= 1/n + (\mathbf{x}_i - \bar{\mathbf{x}})^\top \left(\sum (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \right)^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ 。

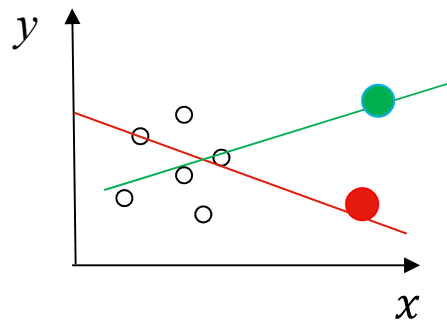
因为 $P_1 \leq H \leq I_n, 1/n \leq h_{ii} \leq 1$ 。

$$h_{ii} \approx 1$$

若 $h_{ii} \approx 1$, 则 \mathbf{x}_i 远离 $\bar{\mathbf{x}}$,

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

(\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) $\Leftrightarrow \mathbf{x}_i = \bar{\mathbf{x}}$,

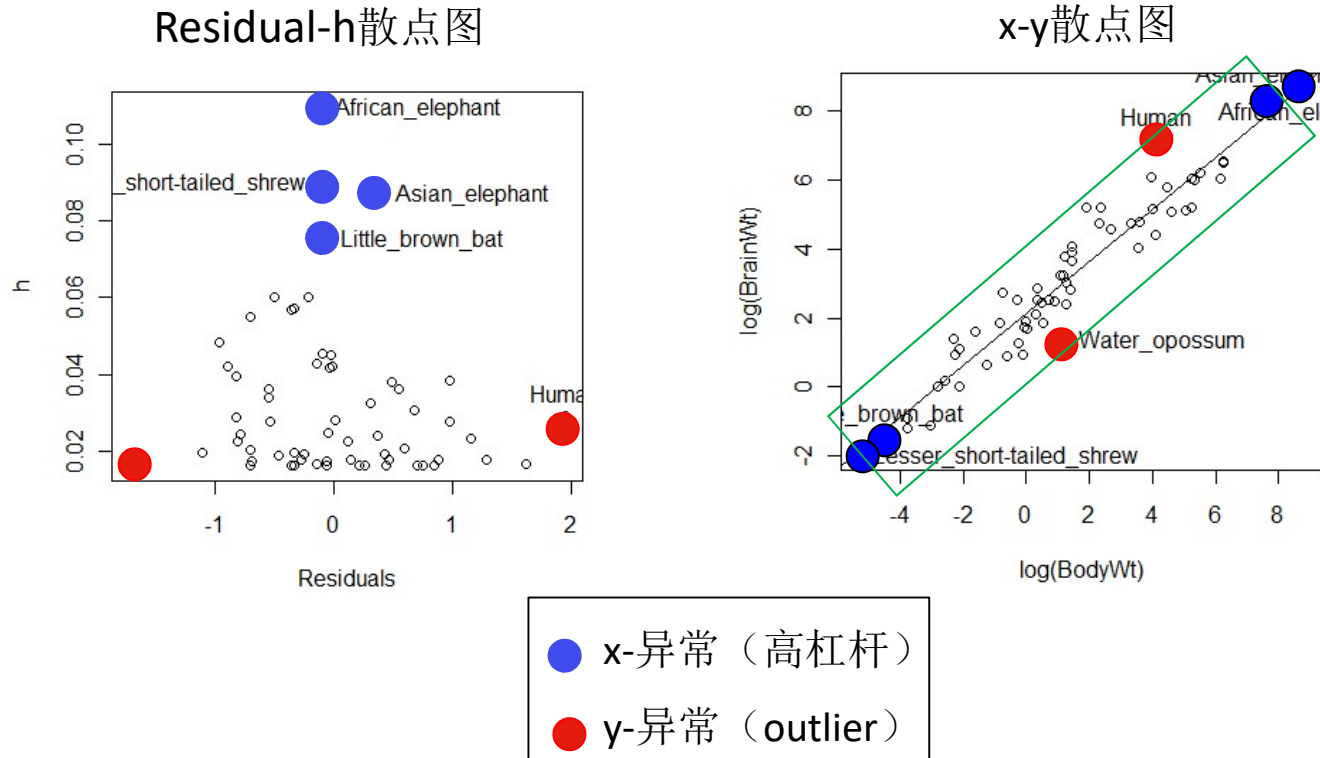
此时 $\hat{y}_i = \hat{\beta}_0 + \hat{\mathbf{b}}^T \mathbf{x}_i = (\bar{y} - \hat{\mathbf{b}}^T \bar{\mathbf{x}}) + \hat{\mathbf{b}}^T \bar{\mathbf{x}} = \bar{y}$.

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 $\sum_{i=1}^n \mathbf{x}_i (y_i - \bar{y} - (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{b}) = 0$

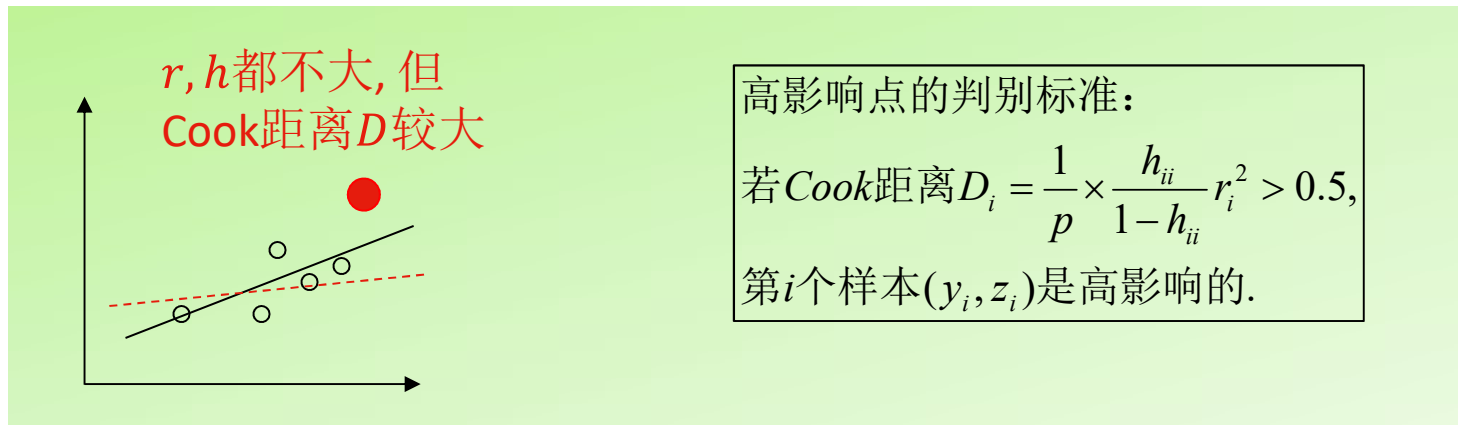
对任何 \mathbf{b} , 第 k 项为0, 不影响正则方程。

例4：动物脑重量与体重的关系 $lm.out=lm(\log(BrainWt)\sim\log(BodyWt), data=brains)$



- > `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: (\mathbf{x}, y) 异常



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, 这种影响反映在“原始模型删除一行数据会导致回归结果出现较大变化”。概括这种影响的度量包括

Cook距离 D , DFFITS、DFBETAS,...

考察删除一个样本点对模型拟合效果的影响的方法称为delete-1、leave-one-out, 也可称为Jackknife (但Jackknife通常专指用于偏差、方差估计)。

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 y_i, \mathbf{x}_i	差异 (DF)
数据	$\mathbf{y} = (y_1, \dots, y_n)^T$ $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$	$\mathbf{y}_{-i} = (\dots, y_{i-1}, y_{i+1} \dots)^T$ $X_{-i} = (\dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1} \dots)^T$	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	$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$	$\tilde{\boldsymbol{\beta}} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T \mathbf{y}_{-i}$	$\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$
y_i 的拟合/预测	$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$	$\tilde{y}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}$ (预测)	$\hat{y}_i - \tilde{y}_i$
所有拟合值	$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$	$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}$	$\ \hat{\mathbf{y}} - \tilde{\mathbf{y}}\ ^2$

注意： $\tilde{\boldsymbol{\beta}}$ 不依赖于删除的数据 (y_i, \mathbf{x}_i) ，所以 $\tilde{y}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}$ 不是拟合值

Cook距离基于所有拟合值之差 $\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2$ （表格右下角）。

基于 $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ ， $\hat{y}_i - \tilde{y}_i$ 构造的影响度量称为DFBETAS, DFFITS

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 $A - \mathbf{u}\mathbf{v}^T$ 可逆, 则有

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

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

附: Sherman-Morrison-Woodbury公式:

假设 U, V 是 $n \times k$ 矩阵, 若 $A, A - UV^T$ 可逆, 则

$$(A - UV^T)^{-1} = A^{-1} + A^{-1}U(I_k - V^T A^{-1}U)^{-1}V^T A^{-1}$$

特别地

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

左端 n 阶矩阵求逆转换为右端 k 阶求逆

注: $\det(A - UV^T) = \det(A) \det(I_k - V^T A^{-1}U)$

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) , $i = 1, 2, \dots$, 需要计算 $n \times n$ 矩阵的逆 $(A - U_i V_i^T)^{-1}$, SMW公式说明, 若 $k \ll n$, 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: 假设线性模型 $\mathbf{y}_{n \times 1} = X_{n \times p} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$, $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, LS估计为 $\hat{\boldsymbol{\beta}}$, 残差为 $e_i, i = 1, \dots, n$, 残差平方和为 $RSS = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^n e_i^2$. 对任何 $1 \leq i \leq n$, 删除数据点 i 后拟合模型 $\mathbf{y}_{-i} = X_{-i} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{-i}, \boldsymbol{\varepsilon}_{-i} \sim (0, \sigma^2 I_{n-1})$, 其LS估计 $\tilde{\boldsymbol{\beta}}$ 和残差平方和 RSS_{-i} 分别为:

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

$$(2) RSS_{-i} = RSS - e_i^2 / (1 - h_{ii}).$$

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

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

命题2说明:

- 删除一个样本点后的LS估计由完整数据模型下的LS估计和残差决定, 不必重新拟合
- 若删除样本点 i , 则残差平方和减小 $e_i^2 / (1 - h_{ii})$. 若该点的杠杆值 $h_{ii} \rightarrow 1$ (影响大), 则残差平方和会大幅度减小 (除非 $e_i = 0$).
- $\hat{y}_i - \tilde{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}} = e_i h_{ii} / (1 - h_{ii})$.

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

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

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

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

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

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

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

$\hat{\boldsymbol{\beta}}$

h_{ii}

(2) 由(1),基于删除样本点*i*拟合模型得到的残差平方和

$$\begin{aligned}
 RSS_{-i} &= \|\mathbf{y}_{-i} - X_{-i}\tilde{\boldsymbol{\beta}}\|^2 = \|\mathbf{y}_{-i} - X_{-i}\hat{\boldsymbol{\beta}} + X_{-i}(X^T X)^{-1}\mathbf{x}_i e / (1 - h_{ii})\|^2 \\
 &\stackrel{\Delta}{=} \|\mathbf{y}_{-i} - X_{-i}\hat{\boldsymbol{\beta}} + \mathbf{a}\|^2 = \|\mathbf{y}_{-i} - X_{-i}\hat{\boldsymbol{\beta}}\|^2 + \mathbf{a}^T \mathbf{a} + 2\mathbf{a}^T (\mathbf{y}_{-i} - X_{-i}\hat{\boldsymbol{\beta}})
 \end{aligned}$$

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

$$\begin{aligned}
 \text{而 } \mathbf{a}^T \mathbf{a} &= e_i^2 \mathbf{x}_i^T (X^T X)^{-1} X_{-i}^T X_{-i} (X^T X)^{-1} \mathbf{x}_i / (1 - h_{ii})^2 \\
 &= e_i^2 \mathbf{x}_i^T (X^T X)^{-1} \{X^T X - \mathbf{x}_i \mathbf{x}_i^T\} (X^T X)^{-1} \mathbf{x}_i / (1 - h_{ii})^2 = e_i^2 h_{ii} / (1 - h_{ii}),
 \end{aligned}$$

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

所以 $RSS_{-i} = RSS - e_i^2 / (1 - h_{ii}).$ 证毕。

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 $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X^T X)^{-1}$, 所以 $D_i = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T \hat{\sigma}^{-2}(X^T X)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) / p$
可理解为 $\hat{\boldsymbol{\beta}}$ 与 $\hat{\boldsymbol{\beta}}^{(-i)}$ 的平均马氏距离。

命题3: 对任何 $1 \leq i \leq n$, $D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$, 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 2(1), $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (X^T X)^{-1} \mathbf{x}_i e_i / (1-h_{ii})$,

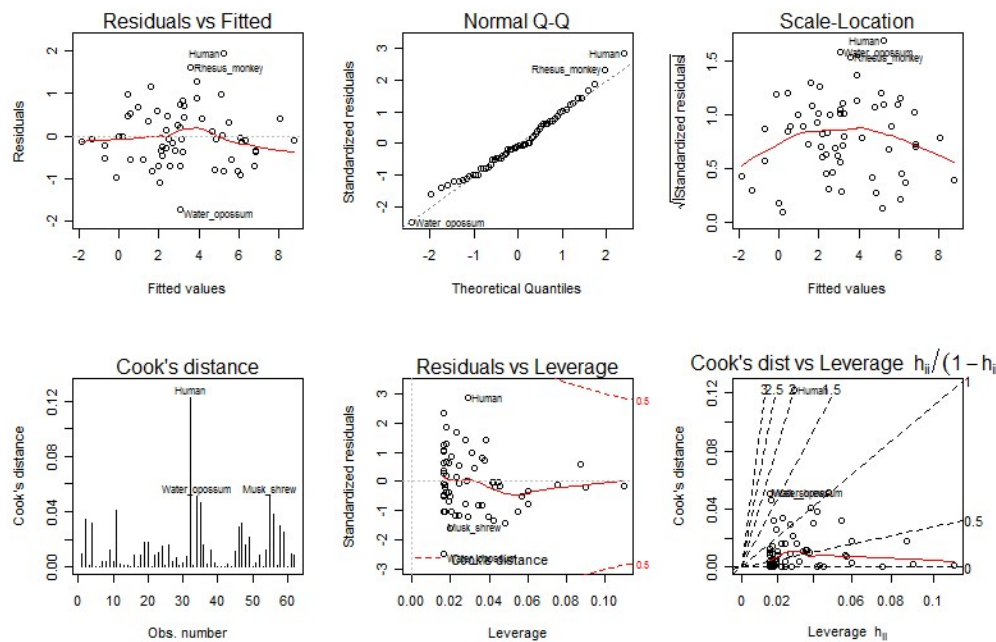
所以 $\hat{\mathbf{y}} - \tilde{\mathbf{y}} = X\hat{\boldsymbol{\beta}} - X\tilde{\boldsymbol{\beta}} = \frac{X(X^T X)^{-1} \mathbf{x}_i e_i}{1-h_{ii}} \Rightarrow D_i = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{p\hat{\sigma}^2} = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$

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

$DFBETAS_i(k) = \frac{\hat{\beta}_k - \tilde{\beta}_k}{\tilde{\sigma}\sqrt{(X^T X)^{-1}_{kk}}}$, $1 \leq k \leq p$. 若 $|DFBETAS_i(k)| \geq \frac{2}{\sqrt{n}}$, i 高影响

例4(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

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

即删除一个样本点后残差平方和减小。

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

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

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

思考：

1. 什么数据点影响较小？

2. 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？删除两个点是否也由类似结论？

附录1: Jackknife 方法简介

Jackknife
便携折刀



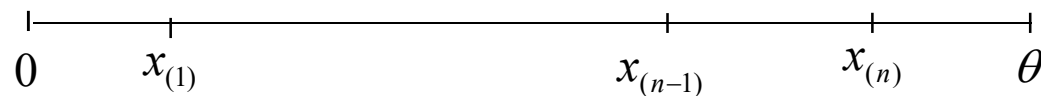
Invented by Quenouille(1949), named by Tukey (1958).

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): $x_1, \dots, x_n \text{ iid} \sim U(0, \theta)$, 记样本从小到大排列为

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

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



直观解法: 间隔 $[0, x_{(1)}], [x_{(1)}, x_{(2)}], \dots, [x_{(n)}, \theta]$ 期望长度相同, 前 n 个区间

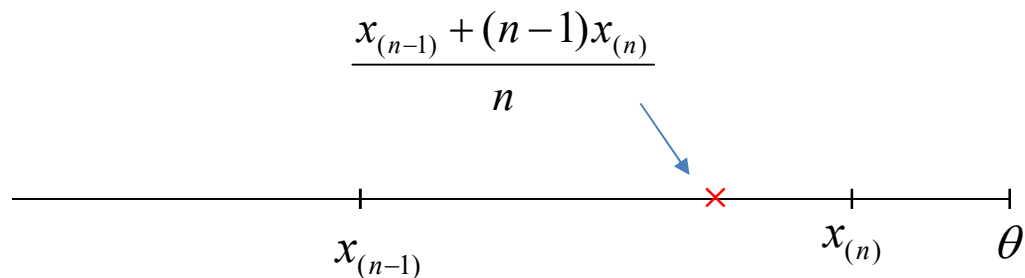
的平均长度: $\bar{d} = (x_{(1)} + (x_{(2)} - x_{(1)}) + \dots + (x_{(n)} - x_{(n-1)})) / n = x_{(n)} / n$, 最后

一个区间的长度 $\theta - x_{(n)}$ 应该约等于 \bar{d} : $\theta - x_{(n)} = \bar{d} = x_{(n)} / n \Rightarrow \hat{\theta}_2 = \frac{n+1}{n} x_{(n)}$.

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:

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

假如 x_1, \dots, 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

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

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。

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

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, \dots, 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 $\hat{\theta}^{(-i)}, i = 1, \dots, n, \bar{\theta} = \sum_{i=1}^n \hat{\theta}^{(-i)} / n$,

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 $\frac{c}{n-1}$ ：

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

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

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, $\hat{\theta}_1 = x_{(n)}$, $E(\hat{\theta}) = \theta + \frac{\theta}{n+1}$, 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

的最大值的预期值, 则 $\bar{\theta} - \hat{\theta} = (x_{(n-1)} - x_{(n)}) / n$.

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

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

Jackknife用于估计方差 (Tukey, 1958)

*Jackknife*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, \dots, x_n 的 θ 的一个估计 $\hat{\theta}$ 。目标：估计 $\text{var}(\hat{\theta})$ 。

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

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

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

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

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

我们下面验证*Jackknife*方法得到的也是该估计。

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

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

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

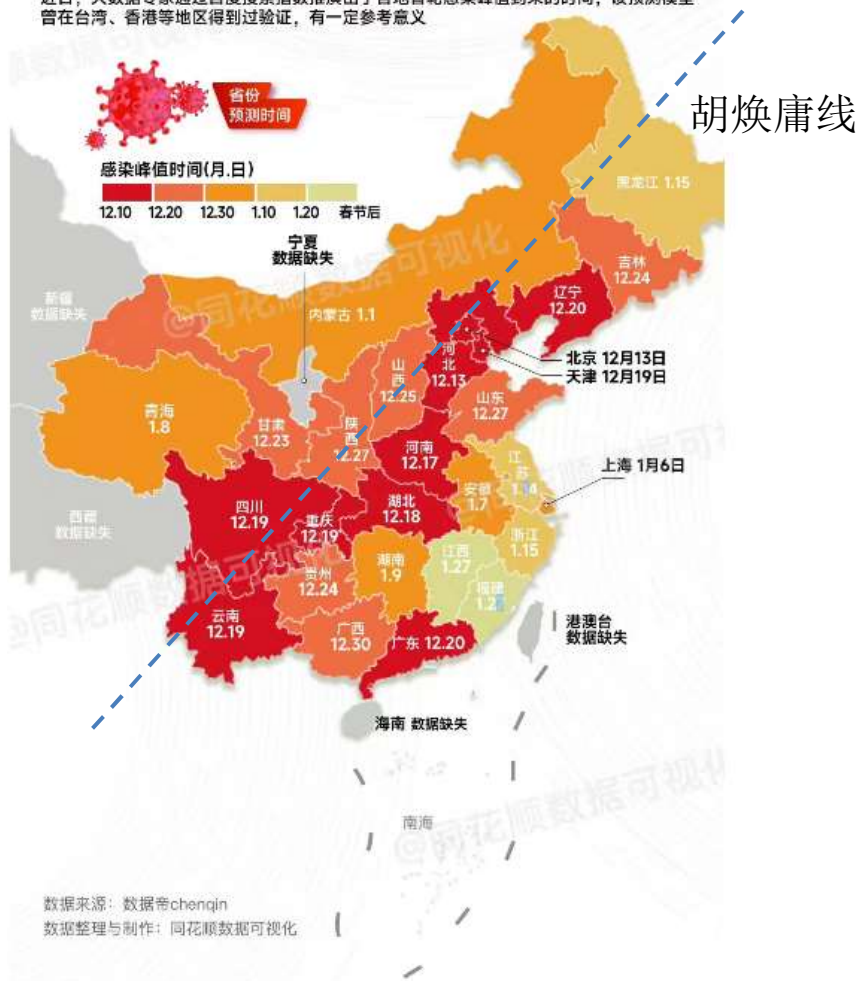
Jackknife与Bootstrap

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, \dots, 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, \dots, x_n ，参数 θ 的估计为 $\hat{\theta}$ 。从 x_1, \dots, 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 $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ ，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。

首轮疫情冲击来临 各省第一波感染高峰时间预测

近日，大数据专家通过百度搜索指数推演出了各地首轮感染峰值到来的时间，该预测模型曾在台湾、香港等地区得到过验证，有一定参考意义



Google Flu Trends was described as using the following method to gather information about flu trends.[4][5]

A linear model is used to compute the log-odds of Influenza-like illness (ILI) physician visit and the log-odds of ILI-related search query:

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \epsilon$$

P is the percentage of ILI physician visit and Q is the ILI-related query fraction computed in previous steps. β_0 is the intercept and β_1 is the coefficient, while ϵ is the error term.

上图基于12月10日数据帝的数据 <https://zhuanlan.zhihu.com/p/590989182>
数据帝chenqin每天都在更新《各城市首轮感染高峰期预测》