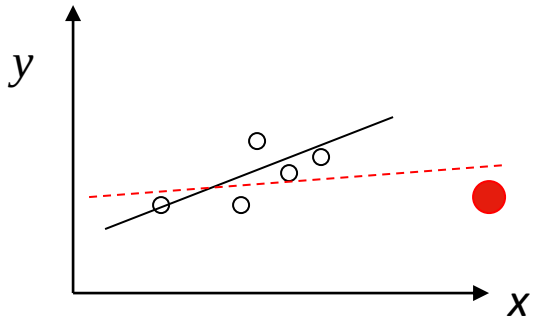
回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第26讲 影响分析（续）

2020.5.22

Jackknife 

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$X=\begin{pmatrix}\mathbf{x}_1^{\mathsf{T}}\\ \vdots \\ \mathbf{x}_n^{\mathsf{T}}\end{pmatrix}=(\mathbf{1},Z)=\begin{pmatrix}1 & \tilde{\mathbf{x}}_1^{\mathsf{T}}\\ \vdots & \vdots \\ 1 & \tilde{\mathbf{x}}_n^{\mathsf{T}}\end{pmatrix}$,

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

帽子矩阵/投影矩阵：$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} = (h_{ij}),\ h_{ij} = \mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_j = (1, \tilde{\mathbf{x}}_i^{\mathsf{T}})\begin{pmatrix} n & \mathbf{1}^{\mathsf{T}}Z \\ Z^{\mathsf{T}}\mathbf{1} & Z^{\mathsf{T}}Z \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ \tilde{\mathbf{x}}_j \end{pmatrix}$

命题1. H的第i个对角元 $h_{ii} = \frac{1}{n} + d_S^2(\tilde{\mathbf{x}}_i,\ \bar{\mathbf{x}})/(n-1)$, 且 $\frac{1}{n} \le h_{ii} \le 1$. (事实上, $h_{ii} < 1$).

证明：$X = (\mathbf{1}, Z)$, Z与$\mathbf{1}$正交化得Z的中心化矩阵：$Z^{\perp} = Z - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}} = \begin{pmatrix} (\tilde{\mathbf{x}}_1 - \bar{\mathbf{x}})^{\mathsf{T}} \\ \\ (\tilde{\mathbf{x}}_n - \bar{\mathbf{x}})^{\mathsf{T}} \end{pmatrix}$.

所以 $H = P_X = P_{\mathbf{1}} + P_{Z^{\perp}} = \frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{n} + Z^{\perp}(Z^{\perp\mathsf{T}}Z^{\perp})^{-1}Z^{\perp\mathsf{T}}$, 而$Z^{\perp\mathsf{T}}Z^{\perp} = (n-1)S$

所以H的第i个对角元素 : $h_{ii} = \frac{1}{n} + (\tilde{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathsf{T}}(Z^{\perp\mathsf{T}}Z^{\perp})^{-1}(\tilde{\mathbf{x}}_i - \bar{\mathbf{x}})$

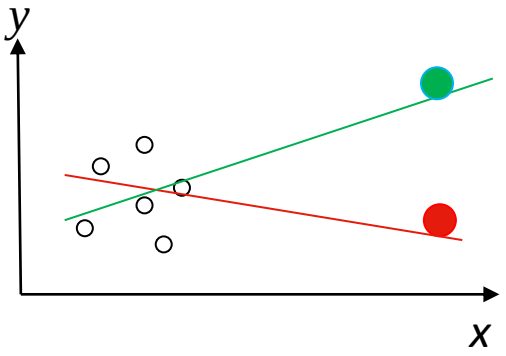
$= \frac{1}{n} + (\tilde{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathsf{T}}\big((n-1)S\big)^{-1}(\tilde{\mathbf{x}}_i - \bar{\mathbf{x}}) = \frac{1}{n} + d_S^2(\tilde{\mathbf{x}}_i,\ \bar{\mathbf{x}})/(n-1).$

显然$h_{ii} \ge \frac{1}{n}$, 等号成立当且仅当$\tilde{\mathbf{x}}_i = \bar{\mathbf{x}}$. 由$I_n - H \ge 0 \Rightarrow h_{ii} \le 1$ 。

定义杠杆(leverage)：$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} = (h_{ij})_{1\le i,j\le n}$的第$i$个对角元素$h_{ii}$称为杠杆(leverage)，它度量了第$i$个自变量远离中心的异常程度。

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

- 若$h_{ii} \approx 1$, $\tilde{\mathbf{x}}_i$远离$\bar{\mathbf{x}}$, $\mathrm{var}(e_i) = (1-h_{ii})\sigma^2 \approx 0 \Rightarrow e_i \approx 0$, $\hat{y}_i \approx y_i$, $(\tilde{\mathbf{x}}_i, y_i)$靠近回归直线而且主导回归直线的方向(右图）



- 若$h_{ii} \approx 1/n$，则$\tilde{\mathbf{x}}_i \approx \bar{\mathbf{x}}$，影响最小，此时

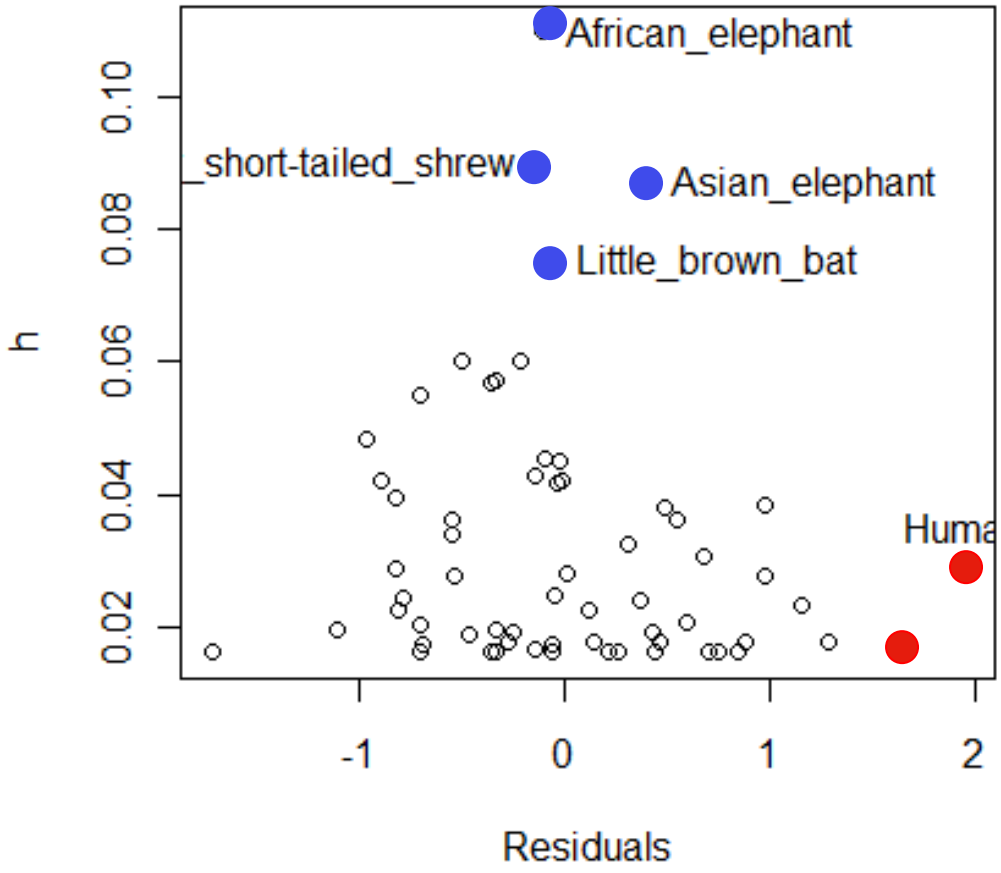
$$\hat{y}_i = \hat{\beta}_0 + \tilde{\mathbf{x}}_i^{\mathsf{T}}\hat{\boldsymbol{\gamma}} = (\bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}\hat{\boldsymbol{\gamma}}) + \tilde{\mathbf{x}}_i^{\mathsf{T}}\hat{\boldsymbol{\gamma}} \approx \bar{y},$$

事实上,若 $h_{ii} = 1/n, (\tilde{\mathbf{x}}_i, \hat{y}_i) = (\bar{\mathbf{x}}, \bar{y})$ 对回归结果没有任何影响。

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

例1(续)：动物脑重量与体重的关系

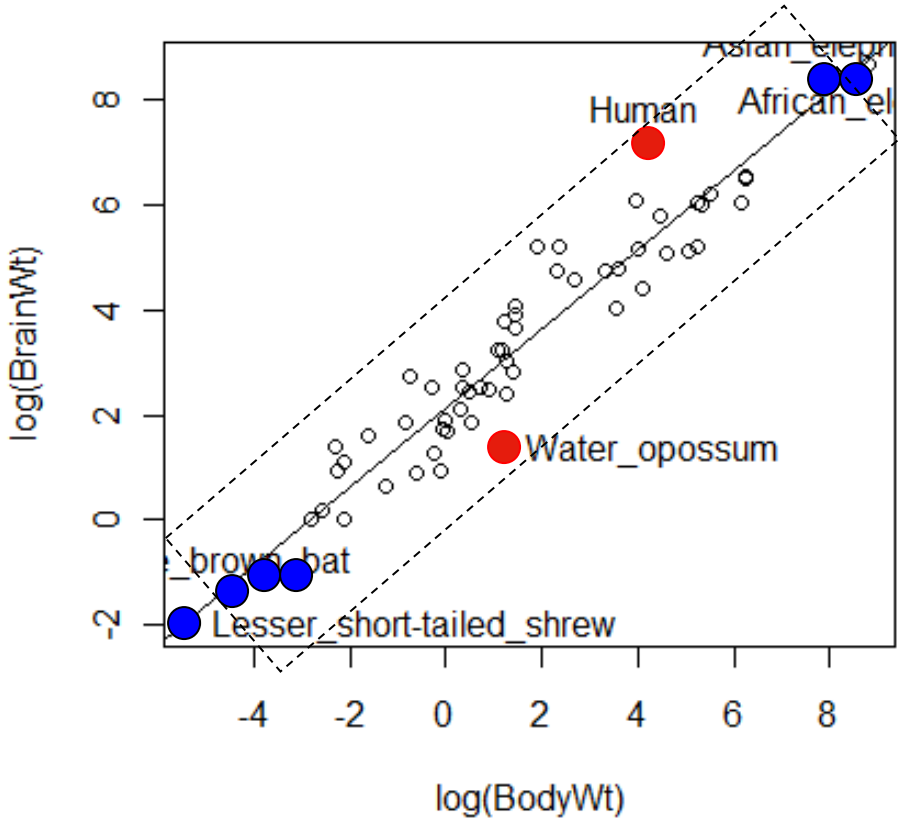
```
a=lm(log(BrainWt)~log(BodyWt), data=brains)
```



- x-异常（高杠杆）
- y-异常（outlier）

异常点outlier(红点Human, Water_opossum) 远离回归线;

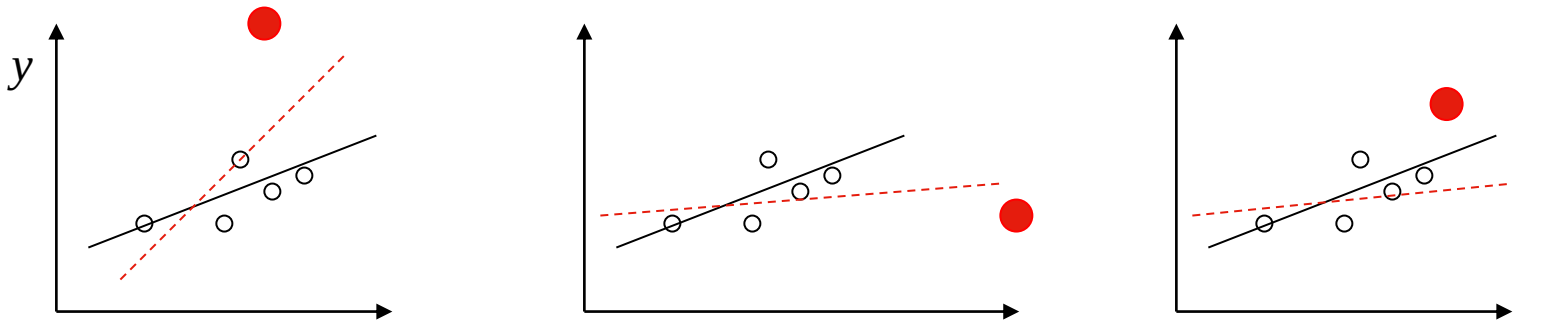
高杠杆点（蓝色）在两端。



对回归直线走向起决定作用的是外围（虚线）附近的那些点（高影响点）

3. 高影响点：(x,y) 异常

(红点) r 较大　　h 较大　　r,h都不大, 但D较大



有时，第i个样本点$(\mathbf{x}_i, y_i)$的自变量和响应都不异常，但它们合在一起(第三图)可能对回归分析的影响可能较大，我们以

Cook距离D、DFFITS、DFBETAS

度量这种影响,它们都是基于“在原始模型中删除一行数据所导致的拟合的变化”得到的。其中D应用最为广泛，其它影响度量通常作为补充。

Jackknife方法定义影响度量

John Tukey 用 delete-1 method （leave-one-out） 方法考察数据点的影响， 即删除一个样本点 导致模型拟合效果的变化程度。这可以称为Jackknife方法（但习惯上Jackknife专指用于求偏差、方差的deletel-1方法）

所有数据：$\mathbf{y}_{n\times 1}, X_{n\times p}$	$\mathbf{y}, X$删除第i行：$\mathbf{y}_{(-i)}, X_{(-i)}$ $(n-1)\times 1$ $(n-1)\times p$	差异(DF)
$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\mathbf{y}_{(-i)} = X_{(-i)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(-i)}$	
$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$	$\hat{\boldsymbol{\beta}}^{(-i)} = (X_{(-i)}{}^{\mathsf{T}}X_{(-i)})^{-1}X_{(-i)}{}^{\mathsf{T}}\mathbf{y}_{(-i)}$	$DFBETAS: \propto \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)}$
$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$	$\hat{\mathbf{y}}^{(-i)} = X\hat{\boldsymbol{\beta}}^{(-i)}$ 假设$X_{(-i)}{}^{\mathsf{T}}X_{(-i)}$可逆	$DFFITS: \propto (\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(-i)})_i = \mathbf{x}_i{}^{\mathsf{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_i{}^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(-i)}$ $Cook's\ D: \propto \|\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(-i)}\|^2 = \|X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}^{(-i)}\|^2$

下标$(-i)$表示删除数据第i行：$\mathbf{y}_{(-i)}$， $X_{(-i)}$.

上标$(-i)$表示基于$\mathbf{y}_{(-i)}$， $X_{(-i)}$得到的估计、拟合等：$\hat{\boldsymbol{\beta}}^{(-i)}, \hat{\mathbf{y}}^{(-i)}$.

例如，$\hat{\boldsymbol{\beta}}^{(-i)}: p\times 1$， $\mathbf{y}_{(-i)}: (n-1)\times 1$， $\hat{\mathbf{y}}^{(-i)}{}_{n\times 1} = X\hat{\boldsymbol{\beta}}^{(-i)} \neq X_{(-i)}\hat{\boldsymbol{\beta}}^{(-i)}$.

数据：$(\tilde{\mathbf{x}}_i, y_i),\ i = 1,...,n,$

模型：$y_i = \beta_0 + \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\gamma} + \varepsilon_i \hat{=} \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, i = 1,...,n$

所有数据$\mathbf{y}_{n\times 1} = (y_1,..., y_n)^{\mathsf{T}}, X_{n\times p} = (\mathbf{x}_1,..., \mathbf{x}_n)^{\mathsf{T}},$

模型：$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$

$\boldsymbol{\beta}$的LS估计：$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y},$

拟合值：$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}},$

σ^2的LS估计：$\hat{\sigma}^2 = \dfrac{RSS}{n-p} = \dfrac{1}{n-p}\| \mathbf{y} - X\hat{\boldsymbol{\beta}} \|^2$

删除第 i 个样本点$(y_i, \tilde{\mathbf{x}}_i)$,

数据：$\mathbf{y}_{(-i)} = (y_1,\ldots, y_{i-1}, y_{i-1},\ldots y_n)^{\mathsf{T}}, X_{(-i)} = (\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_n)^{\mathsf{T}}$,行数$(n-1)$

模型：$\mathbf{y}_{(-i)} = X_{(-i)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(-i)}$

$\boldsymbol{\beta}$的LS估计：$\hat{\boldsymbol{\beta}}^{(-i)} = \left(X_{(-i)}{}^{\mathsf{T}} X_{(-i)}\right)^{-1} X_{(-i)}{}^{\mathsf{T}} \mathbf{y}_{(-i)}$,

σ^2的LS估计：$\hat{\sigma}^{(-i)2} = \dfrac{1}{n-1-p} \| \mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}}^{(-i)} \|^2$

$\mathbf{y}$的拟合值：$\hat{\mathbf{y}}^{(-i)} = X\hat{\boldsymbol{\beta}}^{(-i)}$

注意 $\hat{\mathbf{y}}^{(-i)} = X\hat{\boldsymbol{\beta}}^{(-i)}$的第$i$行$\mathbf{x}_i{}^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(-i)}$是对$y_i$的预测，

$X_{(-i)}\hat{\boldsymbol{\beta}}^{(-i)}$是删除第$i$行数据后得到的$\mathbf{y}_{(-i)}$的拟合

定义（影响度量）：

$$\mathrm{DFBETAS}_i(k)=\frac{\hat{\beta}_k-\hat{\beta}_k^{(-i)}}{\hat{\sigma}^{(-i)}\sqrt{\left((X^{\mathsf{T}}X)^{-1}\right)_{kk}}},\quad 1\le k\le p,1\le i\le n.$$

$$\mathrm{DFFITS}_i=\frac{\left(\hat{\mathbf{y}}-\hat{\mathbf{y}}^{(-i)}\right)_i}{\hat{\sigma}^{(-i)}\sqrt{h_{ii}}}=\frac{\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}-\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(-i)}}{\hat{\sigma}^{(-i)}\sqrt{h_{ii}}},\quad 1\le i\le n.$$

$$\text{Cook 距离}\mathrm{D}_i=\frac{\|\hat{\mathbf{y}}-\hat{\mathbf{y}}^{(-i)}\|^2}{p\hat{\sigma}^2}=\frac{\|X\hat{\boldsymbol{\beta}}-X\hat{\boldsymbol{\beta}}^{(-i)}\|^2}{p\hat{\sigma}^2},\quad 1\le i\le n.$$

(1) 因为$\mathrm{var}(\hat{\beta}_k)=\sigma^2\left((X^{\mathsf{T}}X)^{-1}\right)_{kk}$, $\mathrm{DFBETAS}_i(k)$是$\hat{\beta}_k-\hat{\beta}_k^{(-i)}$的标准化，分母是$\hat{\beta}_k$的标准差,

(2) $\mathrm{var}(\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}})=\sigma^2\mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i=\sigma^2h_{ii}$, $DFFITS_i$是$\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}-\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(-i)}$的标准化;

(3) D_i类似于回归方程显著性的F检验。

需要注意各个度量不服从标准的t, F分布，通常如下判别影响点：

若 $|DFBETAS_i(k)| \geq \frac{2}{\sqrt{n}}$，第 i 个样本点对于估计 β_k 可视为是高影响的。

若 $|DFFITS_i| \geq 2\sqrt{\frac{p}{n}}$，则样本点 i 视为是高影响的。

若 $D_i > 0.5$, 第 i 个样本点影响较大；若 $D_i > 1$, 第 i 个样本点影响很大.

R: 影响度量

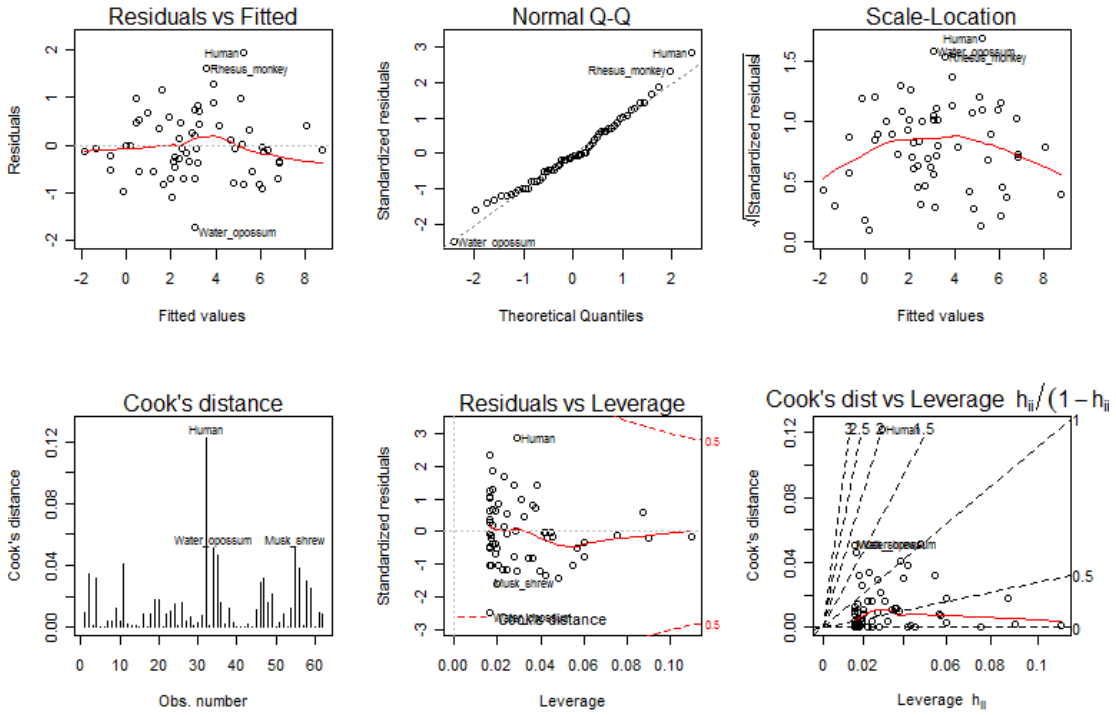
```
> dfbetas(lm.out)
> dffits(lm.out)
> cooks.distance(lm.out)
> hatvalues(lm.out)
> covratio(lm.out)   #参数估计方差的比 det[(X_(-i)^T X_(-i))^-1 σ̂^2_(-i)] / det[(X^T X)^-1 σ̂^2]
> influence.measures(lm.out)   #所有上面的影响度量值。
```

例1(续)：动物脑重量与体重的关系

```
a=lm(log(BrainWt)~log(BodyWt), data=brains)
b=influence.measures(a)$infmat
```

	DFBETA1	DFBETA2	DFFIT	Cov.ratio	D	hat
	dfb.1_	dfb.l(BW	dffit	cov.r	cook.d	hat
Arctic_fox	0.129	-0.005	0.139	1.011	0.01	0.016
Owl_monkey	0.261	-0.147	0.265	0.961	0.034	0.023
Beaver	-0.052	0.017	-0.052	1.048	0.001	0.018
Cow	-0.041	-0.212	-0.252	1.055	0.032	0.055
Gray_wolf	-0.006	-0.007	-0.012	1.06	0	0.025
Goat	0.014	0.014	0.025	1.057	0	0.023
Roe_deer	0.059	0.035	0.086	1.041	0.004	0.019
Guinea_pig	-0.092	0.036	-0.092	1.039	0.004	0.019
Vervet	0.144	0.005	0.159	0.999	0.012	0.016
Chinchilla	0.081	-0.048	0.083	1.05	0.004	0.024
Ground_squirrel	0.259	-0.219	0.288	1.003	0.041	0.038
Arctic_ground_squir	-0.068	0.028	-0.068	1.047	0.002	0.02
African_giant_rat	-0.05	0.02	-0.05	1.05	0.001	0.019
Human	0.22	0.353	0.527	0.797	0.122	0.029

```
# 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，红色虚线为D-等高线.