2014.4.29

第二十一讲. 回归诊断(II): 影响分析

John Wilder Tukey

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.

John Wilder Tukey (June 16, 1915 – July 26, 2000) was an American mathematician best known for development of the **FFT algorithm** and **box plot**.

After the world war II, he returned to Princeton where he obtained his PhD in Math, dividing his time between the university and AT&T Bell Laboratories. Coined many terms, such as **bit (binary digit), software, data analysis**

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 and visualization. He is now regarded as a pioneer of big data analysis.

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

Summary of Tukey's principles for statistical practice

- The usefulness and limitation of mathematical statistics;
- The importance of having methods of statistical analysis that are robust to violations of the assumptions underlying their use;
- The need to amass experience of the behaviour of specific methods of analysis in order to provide guidance on their use;
- The importance of allowing the possibility of data's influencing the choice of method by which they are analysed;
- The need for statisticians to reject the role of 'guardian of proven truth', and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject;
- The iterative nature of data analysis;
- Implications of the increasing power, availability and cheapness of computing facilities;
- The training of statisticians.

Some Tukey words and phrase

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
de ciency
depth
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
polyef ciency
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 (rst in print)
stem-and-leaf
tapering
toolglass
trimming
twicing
vacuum cleaner
vague concept
window carpentry
winsorizing
Winsors principle
Zorns Lemma

影响分析

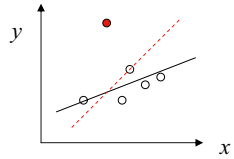
影响分析发现对回归分析影响过大的样本点：

- **异常点 (outlier)： 响应变量y异常**
- **高杠杆点：自变量x异常**
- **高影响样本点： (x,y) 异常**

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。

I. Outlier（Y异常：残差过大/小 ）



残差向量$\mathbf{e}=(\mathrm{I}_n-\mathrm{H})\mathbf{Y}, \mathrm{var}(\mathbf{e})=\sigma^2(\mathrm{I}_n-\mathrm{H})$

残差：$e_i=y_i-\hat{y}_i$， $i=1,2,...,n.$ $\mathrm{var}(e_i)=(1-h_{ii})\sigma^2$

$|e_i|$过大的y_i异常。也可使用标准化残差：

(1) 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，故$\hat{\sigma}$偏大。所以有时使用学 生化残差：

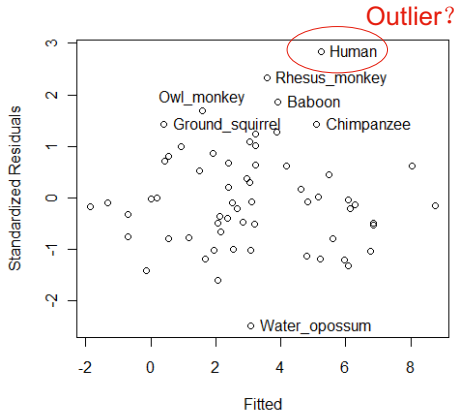
(2) 学生化残差：$r_i^*=\dfrac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}_{(-i)}}$, 其中$\hat{\sigma}_{(-i)}=\sqrt{\mathrm{e}_{(-i)}'\mathrm{e}_{(-i)}/(n-1-p)}$,

$\mathrm{e}_{(-i)}$为删除第i个观察后线性模型拟合的残差

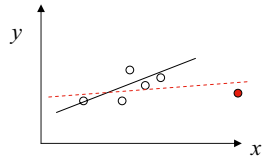
$|r_i|$或$|r_i^*|$较大时(比如≥ 2.576）， y_i被认为异常。

R： rstandard, rstudent

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



II. 高杠杆点（X异常，h过大）



记设计阵$\mathrm{X}=(\mathbf{1},\mathrm{Z})=\begin{pmatrix}1 & \tilde{\mathrm{x}}_1'\\ \dots & \dots\\ 1 & \tilde{\mathrm{x}}_1'\end{pmatrix}$, 令$\mathrm{Z}^{\perp}=\mathrm{Z}-\bar{\mathrm{x}}'\mathbf{1}=\begin{pmatrix}(\tilde{\mathrm{x}}_1-\bar{\mathrm{x}})'\\ \\ (\tilde{\mathrm{x}}_n-\bar{\mathrm{x}})'\end{pmatrix}$=自变量的中心化矩阵

因为$\mathrm{H}=\mathrm{P}_{\mathrm{X}}=\mathrm{P}_{\mathbf{1}}+\mathrm{P}_{\mathrm{Z}^{\perp}}=\dfrac{\mathbf{1}\mathbf{1}'}{n}+\mathrm{Z}^{\perp}(\mathrm{Z}^{\perp\prime}\mathrm{Z}^{\perp})^{-1}\mathrm{Z}^{\perp\prime}$,所以

$$h_{ii}=\frac{1}{n}+(\tilde{\mathrm{x}}_i-\bar{\mathrm{x}})'(\mathrm{Z}^{\perp\prime}\mathrm{Z}^{\perp})^{-1}(\tilde{\mathrm{x}}_i-\bar{\mathrm{x}})$$

$$=\frac{1}{n}+(\tilde{\mathrm{x}}_i-\bar{\mathrm{x}})'\left(\sum(\tilde{\mathrm{x}}_j-\bar{\mathrm{x}})(\tilde{\mathrm{x}}_j-\bar{\mathrm{x}})'\right)^{-1}(\tilde{\mathrm{x}}_i-\bar{\mathrm{x}}),$$

$H = X(X'X)^{-1}X'$的第i个对角元素h_{ii}称为杠杆或影响(*leverage*).
h_{ii}度量了第i个样本点的自变量与平 均值的距离.

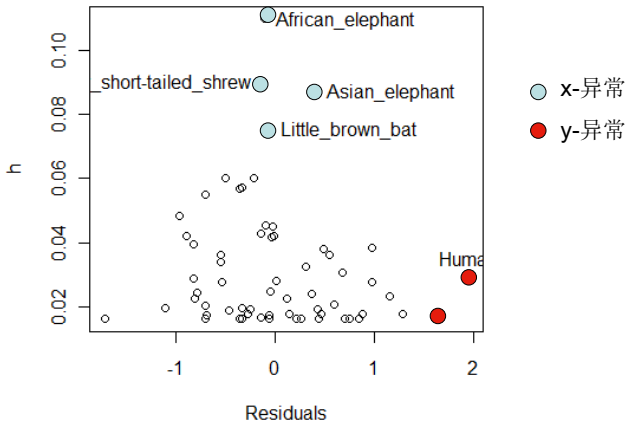
注：

(1) 因为$\sum_{i=1}^{n} h_{ii} = p$（参数个数），平均来看$h_{ii} \sim \frac{p}{n}$.

(2) 若$h_{ii} \approx 1$，$\mathrm{var}(e_i) = (1-h_{ii})\sigma^2 \approx 0$，则$e_i \approx 0$或$\hat{y}_i \approx y_i$.
相当于第i个样本点需要单独一个 参数进行拟合，
因此该点影响较大

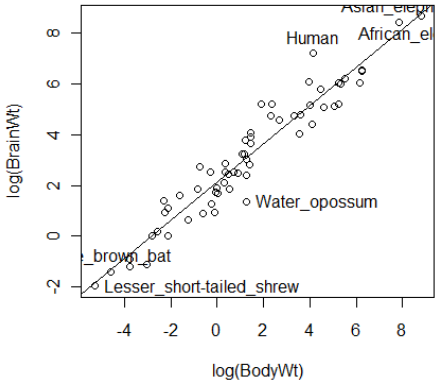
例1(续)：动物脑重量与体重的关系

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



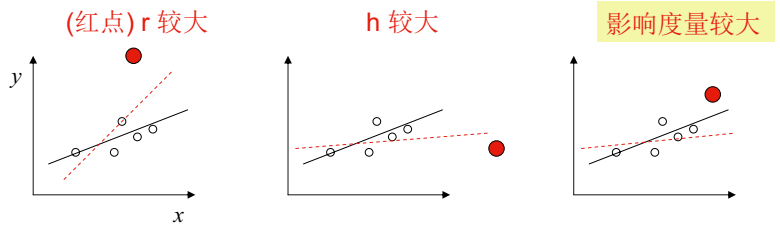
H = hatvalues (lm.out)

散点图上：y异常点(Human, Water_opossum) 远离回归线；高杠杆点(x-异常) 在两端。



III. 高影响样本点

第i个样本点(x_i, y_i)的x,y都可能不异常（r或h都不异常），
但它们综合起来可能对 回归分析影响较大(比如$r \times \sqrt{h}$较大)。



如何寻找高影响点？基本想法是: 如果删除一个样本点 (Jackknife) 会大幅度改变模型拟合效果，则它是高影响的。

基于所有数据的模型	删除第 i 个样本点后的模型
$Y = X\beta + \varepsilon$	$Y_{(-i)} = X_{(-i)}\beta + \varepsilon_{(-i)}$
$\Downarrow$	（删除第i行后记为$Y_{(-i)}$，$X_{(-i)}$）
$\hat{\beta} = (X'X)^{-1}X'Y,$	$\Downarrow$
$\hat{Y} = X\hat{\beta}$	$\hat{\beta}_{(-i)} = (X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)},$
	$\hat{Y}_{(-i)} = X_{(-i)}\hat{\beta}_{(-i)}$。

数据点i的影响体现在差异(DIFference)：$\hat{\beta} - \hat{\beta}_{(-i)}$，$\hat{Y} - \hat{Y}_{(-i)}$

常用的三种影响度量为: DFFITS，DFBETAS，Cook's D

它们都是基于比较删除样本点 i 后的分析结果与所有数据的分析结果

(i) DFBETAS

$\hat{\beta}_{k(-i)}$为删除第i行数据后β_k的LS估计，$\hat{\beta}_k$为基于所有数据的LS估计。

$\hat{\sigma}^2_{(-i)}$为删除第i行数据后σ^2的估计

$$DFBETAS_i = \frac{\hat{\beta}_k - \hat{\beta}_{k(-i)}}{\sqrt{c_{kk}\hat{\sigma}^2_{(-i)}}}, c_{kk} = \| \mathbf{x}_k^{\perp} \|^{-2} = (X'X)^{-1}\text{的}(k,k)\text{元}$$

回忆：$\text{var}(\hat{\beta}) = (X'X)^{-1}\sigma^2, \text{var}(\hat{\beta}_k) = c_{kk}\sigma^2$, 分母 $= \widehat{\text{var}(\hat{\beta}_k)} = c_{kk}\hat{\sigma}^2_{(-i)}$,

$\hat{\beta}_k$使用了样本i, 而$\hat{\beta}_{k(-i)}$没有, 前者可能有问题, 而后者可认为正常.

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

(ii). DFFITS

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{\sqrt{h_{ii}\hat{\sigma}^2_{(-i)}}}, \text{其中} \hat{Y}_{i(-i)} = \tilde{x}_i\hat{\beta}_{(-i)} = \text{删除第}i\text{行后对}Y_i\text{的预测}$$

$$\text{Recall: var}(\hat{Y}) = H\sigma^2, \text{var}(\hat{Y}_i) = h_{ii}\sigma^2. \text{ 视}\hat{Y}_{i(-i)}\text{为}\hat{Y}_i\text{"好"的预测,}$$

可以验证：$DFFITS_i = r_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}$, 其中 $r_i^* = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$

为学生化残差. 所以，为了求第i个观测值（$y_i, \tilde{x}_i$)的影响值$DFFITS_i$,

你不必删除该行数据重新回归。

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

(iii). Cook 距离 （最常用的影响度量）

Cook 距离：$D_i = \frac{\| \hat{Y} - \hat{Y}_{(-i)} \|^2}{p\hat{\sigma}^2} = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})'X'X(\hat{\beta} - \hat{\beta}_{(-i)})}{p\hat{\sigma}^2}$

为什么这样定义？D_i代表了去掉第 i 个样本点对所有拟合值 的影响

如果将$\hat{\beta}_{(-i)}$ 作为真值(不受样本点i影响), D具有F检验统计量的形式.

可以验证：$D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$，其中 $r_i = \frac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$ 为标准化残差。

D_i较大 $\Leftrightarrow$ h_{ii}较大（x_i高杠杆），或 $|r_i|$较大（y_i异常）。

R中：若 $D > 0.5$，影响偏大；若 $D > 1$，影响很大。

证明: $D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$，其中 $r_i = \frac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$。

证: 删除样本点i,拟合模型$Y_{(-i)} = X_{(-i)}\beta + \varepsilon_{(-i)}$,得$\hat{\beta}_{(-i)} = (X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)}$

记$X = \begin{pmatrix} \tilde{x}_1' \\ \dots \\ \tilde{x}_n' \end{pmatrix}$，则$X'X = \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i'$, $X'Y = \sum_{i=1}^{n} \tilde{x}_i y_i$

$$\Rightarrow X'X = X_{(-i)}'X_{(-i)} + \tilde{x}_i\tilde{x}_i', X'Y = \sum_{j=1}^{n} \tilde{x}_j y_j = X_{(-i)}'Y_{(-i)} + \tilde{x}_i y_i$$

利用事实：设$A_{p\times p}$对称, x，y为$p\times 1$向量，则$(A+xy')^{-1} = A^{-1} - \frac{A^{-1}xy'A^{-1}}{1+x'A^{-1}y}$,

$$\Rightarrow \hat{\beta}_{(-i)} = (X_{(-i)}'X_{(-i)})^{-1}X_{(-i)}'Y_{(-i)} = (X'X - \tilde{x}_i\tilde{x}_i')^{-1}(X'Y - \tilde{x}_i y_i)$$

$$= \left((X'X)^{-1} + \frac{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}}{1-\tilde{x}_i'(X'X)^{-1}\tilde{x}_i}\right)(X'Y - \tilde{x}_i y_i)$$

注意到$h_{ii} = \tilde{x}_i'(X'X)^{-1}\tilde{x}_i$

$$\hat{\beta}_{(-i)} = \left((X'X)^{-1} + \frac{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}}{1-\tilde{x}_i'(X'X)^{-1}\tilde{x}_i}\right)(X'Y - \tilde{x}_i y_i)$$

$$= \hat{\beta} - (X'X)^{-1}\tilde{x}_i y_i + \frac{1}{1-h_{ii}}\left\{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}X'Y - (X'X)^{-1}\tilde{x}_i\tilde{x}_i'(X'X)^{-1}\tilde{x}_i y_i\right\}$$

$$= \hat{\beta} - (X'X)^{-1}\tilde{x}_i y_i + \frac{1}{1-h_{ii}}\left\{(X'X)^{-1}\tilde{x}_i\tilde{x}_i'\hat{\beta} - h_{ii}(X'X)^{-1}\tilde{x}_i y_i\right\}$$

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

$$\Rightarrow X\hat{\beta} - X\hat{\beta}_{(-i)} = \frac{1}{1-h_{ii}}\left\{X(X'X)^{-1}\tilde{x}_i e_i\right\}$$

所以 $D_i = \frac{(\hat{\beta}-\hat{\beta}_{(-i)})'X'X(\hat{\beta}-\hat{\beta}_{(-i)})}{p\hat{\sigma}^2} = \frac{1}{p}\times\frac{h_{ii}}{1-h_{ii}}r_i^2$

注:由$\hat{Y}_i - \hat{Y}_{i(-i)} = \left(X\hat{\beta} - X\hat{\beta}_{(-i)}\right)_i = \frac{1}{1-h_{ii}}\left\{\tilde{x}_i'(X'X)^{-1}\tilde{x}_i e_i\right\} = \frac{h_{ii}e_i}{1-h_{ii}} \Rightarrow DFFITS_i = r_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}$

R: 影响度量

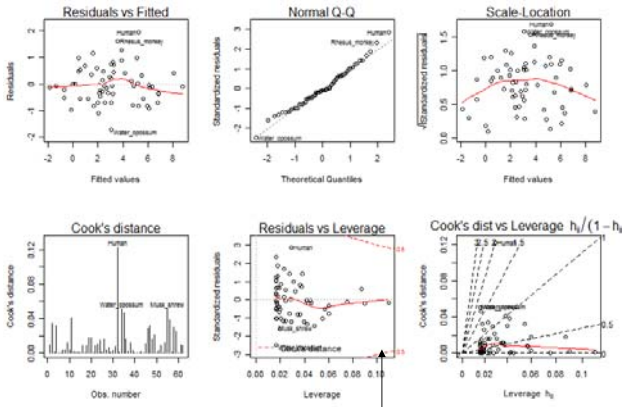
```
> dfbetas(lm.out)
> dffits(lm.out)
> cooks.distance(lm.out)
> hatvalues(lm.out)
> covratio(lm.out)   #参数估计方差的比 det(X_(-i)'X_(-i))^{-1} σ̂²_(-i) / det(X'X)^{-1} σ̂²
> influence.measures(lm.out)   #所有上面的影响度量值 。
```

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
```

	dfb.1_	dfb.l(BW	dffit	cov.r	cook.d	hat
Arctic_fox	0.12929	-0.00529	0.1386	1.01	9.6e-03	0.016
Owl_monkey	0.26065	-0.14732	0.2650	0.96	3.4e-02	0.023
Beaver	-0.05218	0.01663	-0.0524	1.05	1.4e-03	0.018
Cow	-0.04137	-0.21156	-0.2517	1.05	3.2e-02	0.055
…						

```
# R 回归诊断图（残差分析+影响分析)
> a =lm( BrainWt ~ BodyWt, data=log(brains)
> plot(a, 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

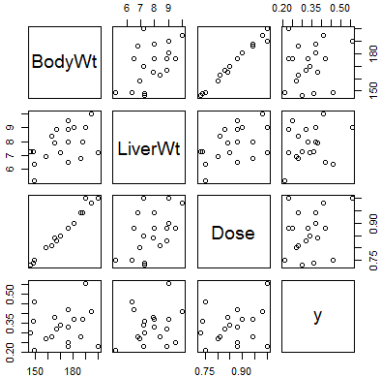
(6) D vs h: 影响分析

红色虚线为D-等高线.

例2. (Dennis Cook数据)k数据)。 一项试验希望研究动物肝脏对某种药物的吸收能力。为此随机选取19只小白鼠，口服该药物，剂量大小由体重决定 (大约为 40mg/kg 乘以体重)。一段时间后，测出肝中与所含药物重量，除以服用的剂量，得到肝吸收百分比 y。我们研究 y 我们研究 y 与体重BodyWt, 肝重LiverWt，剂量Dose的关系。

```
lm(  y ~ ., data = rat)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.266     0.195    1.4     0.19
BodyWt       -0.021     0.008   -2.7     0.02 *
LiverWt       0.014     0.017    0.8     0.42
Dose          4.178     1.523    2.7     0.02 *
---

Multiple R-squared:  0.36,
F-statistic: 2.9 on 3 and 15 DF,  p-value: 0.072
```



BowyWt和Dose都显著。

但理论上，这种决定药物剂量的方法决定了所测得到的 y 应该与三个变量无关。

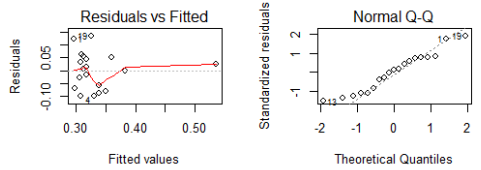
因为BodyWt, Dose几乎完全成正比，模型中去掉BodyWt 或Dose, 变量不再显著：

```
lm( y ~ . - Dose, data = rat)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17836    0.22778     0.8     0.4
BodyWt       0.00035    0.00151     0.2     0.8
LiverWt      0.01233    0.02041     0.6     0.6
```

```
lm( y ~ . - BodyWt, data = rat)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1174     0.2191      0.5     0.6
LiverWt      0.0087     0.0201      0.4     0.7
Dose         0.1736     0.2863      0.6     0.6
```

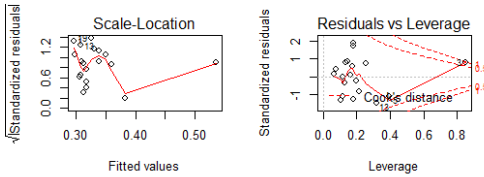
如何解释这种现象？

```
a = lm(formula = y ~ ., data = rat)
plot(a)
```



从第一个图可以发现第3个小白鼠的你拟合值过大，其杠杆值 h =0.85；

从第4个图发现其D =0.93, 是高影响点。



删除第三行数据重新分析：

```
lm(formula = y ~ ., data = rat[-3, ])

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.311    0.205     1.52     0.15
BodyWt       -0.008    0.019    -0.42     0.68
LiverWt       0.009    0.019     0.48     0.64
Dose          1.485    3.713     0.40     0.70


Multiple R-squared:  0.0211,
F-statistic:  0.1 on 3 and 14 DF,  p-value: 0.958
```

	BodyWt	LiverWt	Dose	y
1	176	6.5	0.88	0.42
2	176	9.5	0.88	0.25
3	190	9	1	0.56
4	176	8.9	0.88	0.23
5	200	7.2	1	0.23
6	167	8.9	0.83	0.32
7	188	8	0.94	0.37
8	195	10	0.98	0.41
9	176	8	0.88	0.33
10	165	7.9	0.84	0.38
11	158	6.9	0.8	0.27
12	148	7.3	0.74	0.36
13	149	5.2	0.75	0.21
14	163	8.4	0.81	0.28
15	170	7.2	0.85	0.34
16	186	6.8	0.94	0.28
17	146	7.3	0.73	0.3
18	181	9	0.9	0.37
19	149	6.4	0.75	0.46

或去除BodyWt, Dose之一：

```
lm(formula = y ~ . -Dose, data = rat[-3, ])
lm(formula = y ~ . - BodyWt, data = rat[-3, ])
```

得到的结果不再矛盾。

```
Influence measures of
        lm(formula = y ~ ., data = rat) :

   dfb.1_ dfb.BdyW dfb.LvrW dfb.Dose  dffit cov.r cook.d   hat
1  -0.038   0.315   -0.704   -0.244   0.892 0.631 0.169 0.178
2   0.143  -0.098   -0.482    0.126  -0.609 1.016 0.089 0.179
3  -0.231  -1.668    0.305    1.747   1.905 7.401 0.930 0.851
4   0.125  -0.127   -0.304    0.140  -0.494 0.860 0.057 0.108
5   0.522  -0.396    0.550    0.275  -0.909 1.524 0.203 0.392
6   0.002   0.014    0.029   -0.017   0.043 1.567 0.000 0.161
7  -0.184   0.150   -0.083   -0.118   0.310 1.289 0.025 0.137
8  -0.297   0.059    0.246   -0.040   0.426 1.520 0.047 0.254
9  -0.010   0.018    0.000   -0.017   0.043 1.402 0.000 0.067
10 -0.006   0.010   -0.003   -0.009  -0.014 1.496 0.000 0.120
11 -0.291   0.194    0.101   -0.173  -0.410 1.066 0.041 0.120
12  0.217  -0.025    0.052   -0.009   0.269 1.444 0.019 0.172
13 -0.772   0.144    0.766   -0.120  -1.099 0.972 0.273 0.316
14 -0.035  -0.046   -0.077    0.060  -0.142 1.461 0.005 0.131
15  0.019   0.041   -0.055   -0.038   0.119 1.359 0.004 0.076
16  0.123  -0.006    0.329   -0.049  -0.447 1.375 0.051 0.217
17 -0.104   0.015   -0.028    0.002  -0.126 1.607 0.004 0.195
18 -0.154   0.190    0.162   -0.189   0.352 1.270 0.032 0.149
19  0.856  -0.250   -0.295    0.171   0.995 0.517 0.200 0.178
```