应用回归分析

上海财经大学　统计与管理学院

第九章异常值与强影响值

章节概括：

- 异常值
- 异常值检验
- 强影响值

异常值

- 异常值分为两种情况：

  一种是关于因变量y异常；

  另一种是关于自变量x异常

- 在残差分析中，认为超过$\pm 3\hat{\sigma}$的残差为异常值。

- 当数据中存在关于 y 的异常观察值时，异常值把回归线拉向自己，使异常值本身的残差减少，而其余观察值的残差增大，这时回归标准差$\hat{\sigma}$也会增大，因而用“3σ”准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。

异常值

- 线性模型

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathrm{E}(\mathbf{e}) = \mathbf{0} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

- 正常值

$$\mathrm{E}(Y|X = \mathbf{x}_j) = \mathbf{x}_j'\boldsymbol{\beta}$$

- 异常值

$$\mathrm{E}(Y|X = \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \delta$$

- 异常值检验

$$\delta = 0$$

异常值检验

- 检验I
- 若i个观测可能为异常值，对应定义虚拟变量 U
- 用X和U回归Y
- T-test检验 $\delta = 0$
- 自由度为 $n - p' - 1$

异常值检验

- 检验II
- 1.删除第i个观测，$\hat{\beta}_{(i)}$ $\hat{\sigma}^2_{(i)}$
- 2.依据保留的(n-1)个观测，估计系数和方差
- 3.对删除第i个观测，计算 $\hat{y}_{i(i)} = \mathbf{x}_i' \hat{\beta}_{(i)}$

  注 y_i and $\hat{y}_{i(i)}$ 独立，且

$$\mathrm{Var}(y_i - \hat{y}_{i(i)}) = \sigma^2 + \sigma^2 \mathbf{x}_i' (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1} \mathbf{x}_i$$

- 4.注 $\mathrm{E}(y_i - \hat{y}_{i(i)}) = \delta$,假设误差为正态分布，则

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + \mathbf{x}_i' (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1} \mathbf{x}_i}}$$

在零假设成立时服从t分布，自由度为 $n - p' - 1$

标准化残差

- 标准化残差 (standardized residual)

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

期望为 0，方差为 1

- 学生化残差 (studentized residual, W.S. Gosset)

$$t_i = r_i \left(\frac{n-p'-1}{n-p'-r_i^2} \right)^{1/2} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

不用再回归

异方差

- 异方差

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{W}^{-1}$$

- 残差

$$\hat{e}_i = \sqrt{w_i}(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)$$

- 其余相同

多重检验

- 单个假设检验

$$n = 65,\ p' = 4$$

$$P(t(60) > 2.0) = 0.05$$

- 多重假设检验

65个独立的假设检验

$$P(t(60) > 2.0) = 0.964$$

Bonferroni不等式

- 若每个假设检验水平为a,
  则n个假设检验水平不超过 na
- 十分保守，提供的一个概率上界

- 若n个假设检验水平为 α
  单个检验的水平定为 $(\alpha/n) \times 100\%$

$$.05/65 = .00077$$

$$65(.00077) = .05$$

Forbe's数据

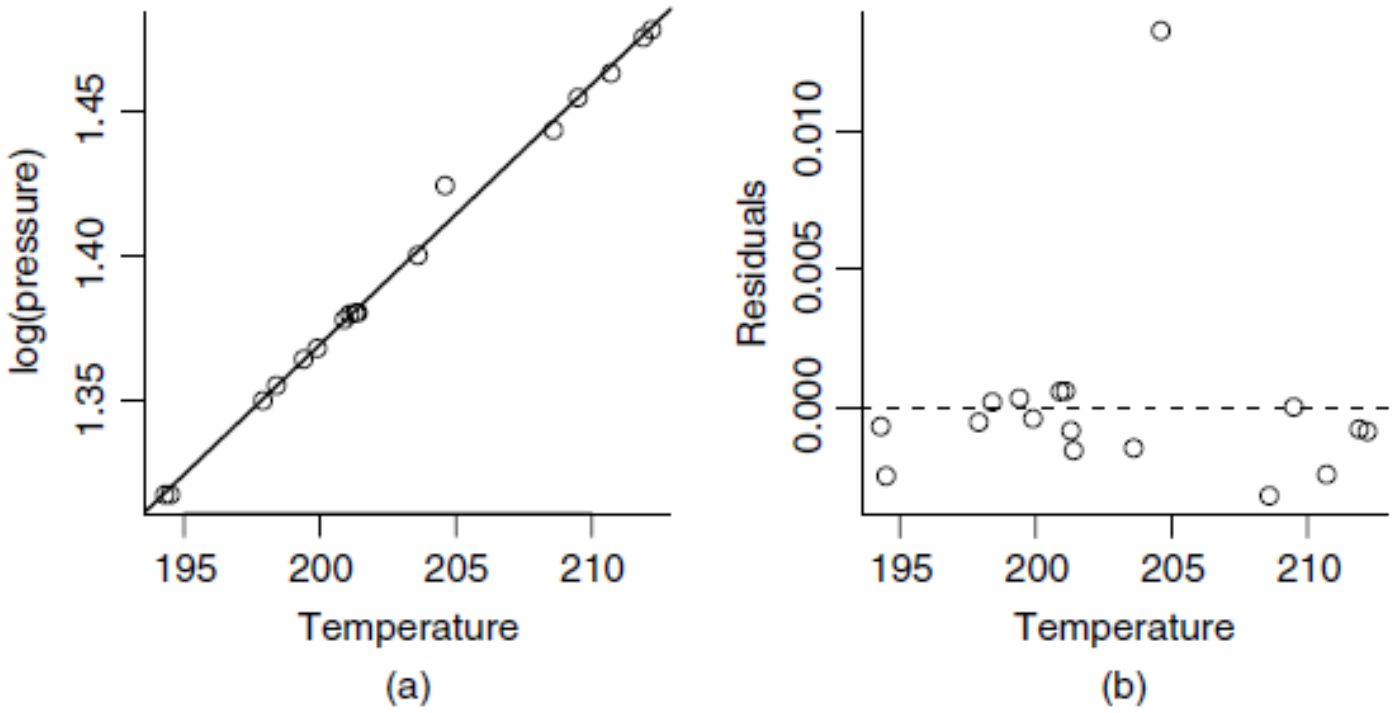

FIG. 1.4 (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of log(*Pressure*) on *Temp*. (b) Residuals versus *Temp*.

Forbe's数据

- 第12个观测可能为异常值
- 计算标准残差

$$\hat{e}_i = 1.36,\ \hat{\sigma} = 0.379 \qquad h_{12,12} = 0.0639$$

$$r_{12} = \frac{1.3592}{0.379\sqrt{1 - .0639}} = 3.7078$$

- T检验

$$t_i = 3.7078\left(\frac{17 - 2 - 1}{17 - 2 - 3.7078^2}\right)^{1/2} = 12.40$$

$$P(|t(14)| > 12.40) = 6.13 \times 10^{-9}$$

- Bonferroni p值 $17 \times 6.13 \times 10^{-9} = 1.04 \times 10^{-7}$

强影响值

$$\mathrm{Var}(\hat{e}_i) = \hat{\sigma}^2(1 - h_{ii})$$

- h_{ii}是帽子矩阵中主对角线的第i个元素，它是调节e_i方差大小的杠杆，因而称h_{ii}为第i个观察值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值h_{ii}也是表示自变量的第i次观测值与自变量平均值之间距离的远近。较大的杠杆值的残差偏小，这是因为大杠杆值的观测点远离样本中心，能够把回归方程拉向自己，因而把杠杆值大的样本点称为强影响点。

强影响值

- 包含异常值

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- 删除异常值

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}$$

- 对比系数估计

Cook's距离

- 库克距离

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\sigma}^2}$$

- 若 $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ $\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)}$

$$D_i = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{p'\hat{\sigma}^2}$$

- D_i 越大影响越大

Cook's距离

- A.12

$$D_i = \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}}$$

- r_i 大，在第i个观测点拟合的不好
- h_{ii} 大，$\mathbf{x}_i$ 距离样本均值 $\bar{\mathbf{x}}$. 远
- 对于库克距离，判断其大小的方法比较复杂，一个粗略的标准是

  当D_i<0.5时，认为不是异常值点，

  当D_i>1时，　认为是异常值点。

Rat数据

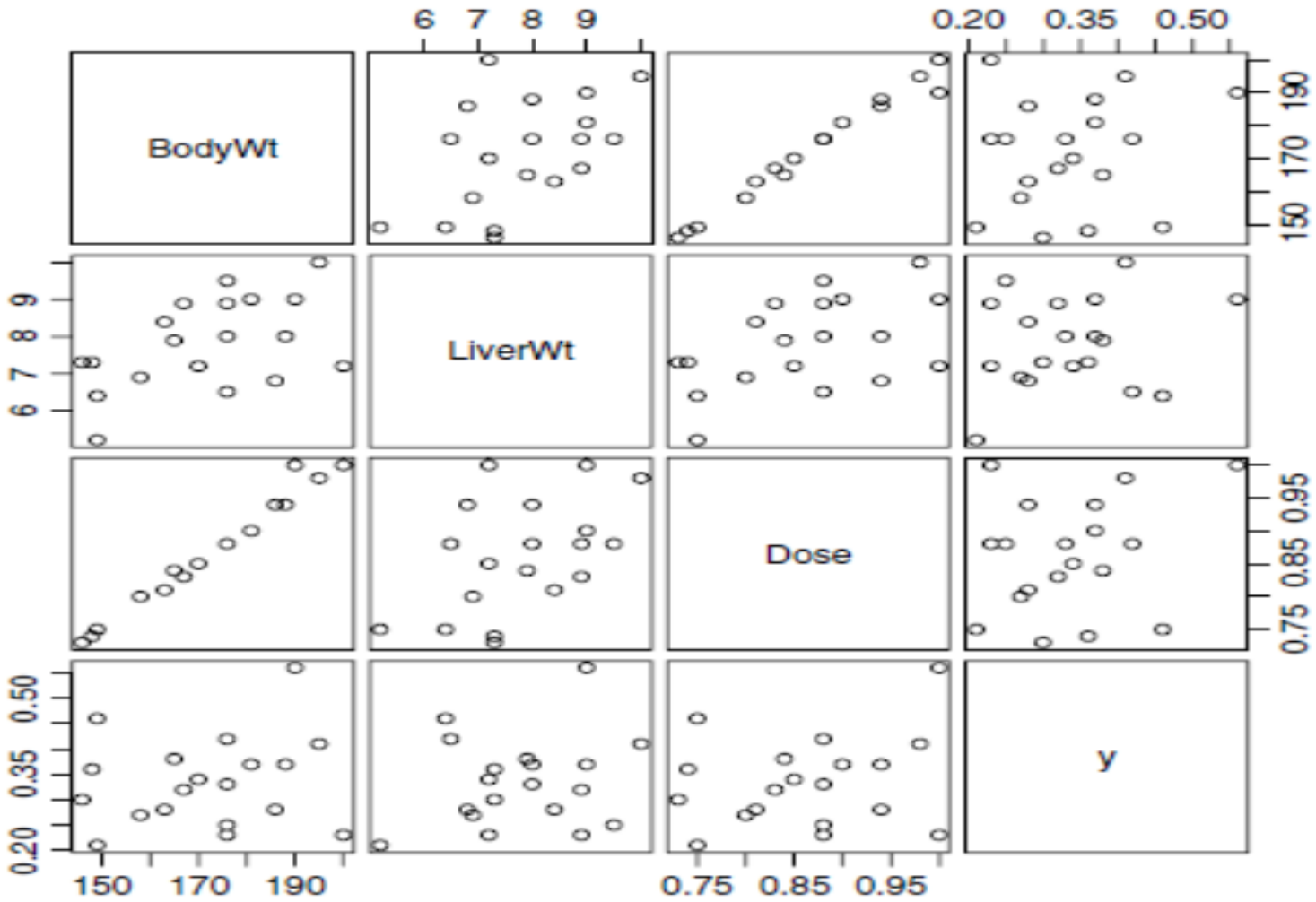

FIG. 9.2 Scatterplot matrix for the rat data.

Rat数据

TABLE 9.1 Regression Summary for the Rat Data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWt      -0.021246   0.007974  -2.664   0.0177
LiverWt      0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-Squared: 0.3639
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

Rat数据

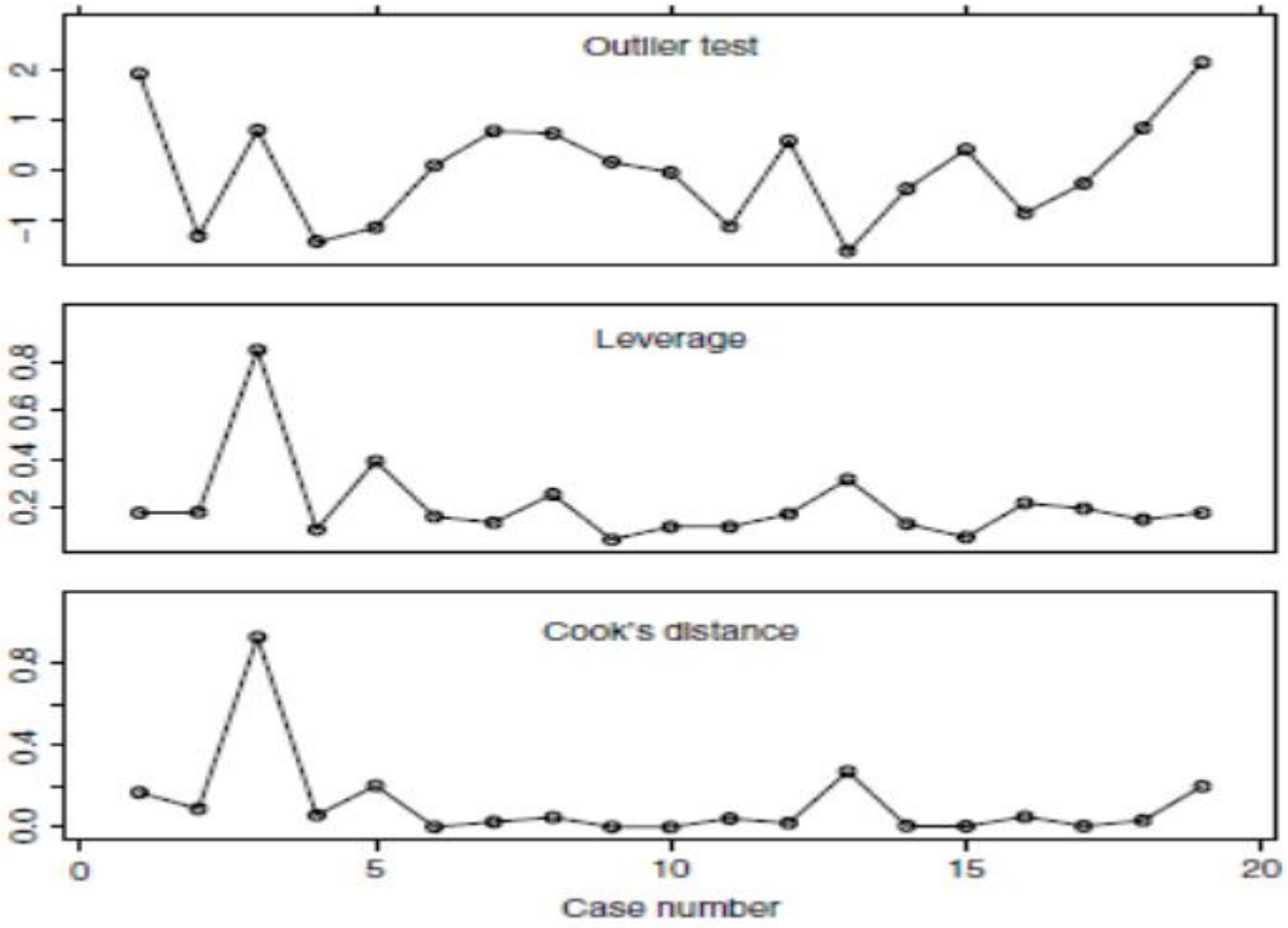

FIG. 9.3 Diagnostic statistics for the rat data.

Rat数据

TABLE 9.2 Regression Summary for the Rat Data with Case 3 Deleted

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWt      -0.007783   0.018717  -0.416    0.684
LiverWt      0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-Squared: 0.02106
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585
```

正态QQ-Plot

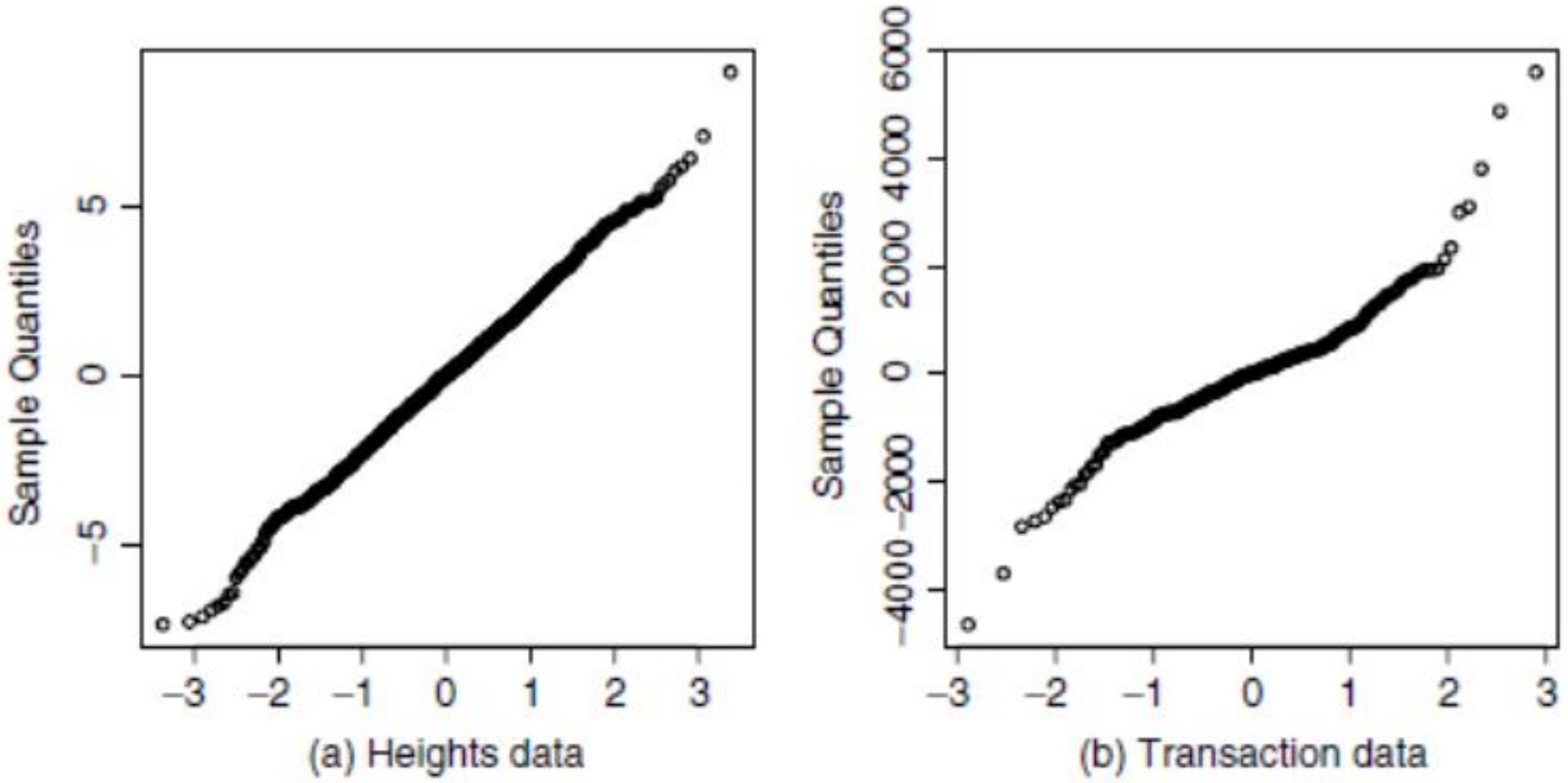

FIG. 9.5 Normal probability plots of residuals for (a) the heights data and (b) the transactions data.

处理方法

异常值原因	异常值消除方法
1. 数据登记误差，存在抄写或录入的错误	重新核实数据
2. 数据测量误差	重新测量数据
3. 数据随机误差	删除或重新观测异常值数据
4. 缺少重要自变量	增加必要的自变量
5. 缺少观测数据	增加观测数据，适当扩大自变量取值范围
6. 存在异方差	采用加权线性回归
7. 模型选用错误，线性模型不适用	改用非线性回归模型

Thank You !