多元线性回归：回归诊断

Multiple Linear Regression

王树佳 | 深圳大学经济学院

sjwang123@163.com

学习目标

检查多元回归模型

- 模型假设是否符合实际
- 数据是否存在问题

第七章 多变量关系分析

为什么要进行回归诊断（也是敏感性分析）？

模型的误差项ε假设：

（1）误差项ε的均值为0；

（2）对不同的观察值，误差项ε的方差 σ^2 都相同；

（3）对不同的观察值，误差项ε相互独立；

（4）误差项ε服从正态分布。

如果这些假设的一项或多项不成立，那么回归关系的显著性检验、预测及区间估计的结果就可能不成立

本节将对这些假设是否成立、数据是否正确等问题进行检查和诊断。

回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

回归诊断的最重要的工具是残差(Residual)，因此我们首先介绍残差与残差图，以及与其相关的杠杆值(Leverage)等概念。

残差(Residual)：

第 i 个观察值的残差就是第 i 个因变量的实际观察值 y_i 与预测值 $\hat{y}_i$ 的差值，即

$$e_i = y_i - \hat{y}_i, i = 1, 2, \cdots, n$$

可以证明：在模型的假设下，残差的平均值和标准差为

$$E(e_i)=0 \quad \sigma_{e_i}=\sigma\sqrt{1-h_i}$$

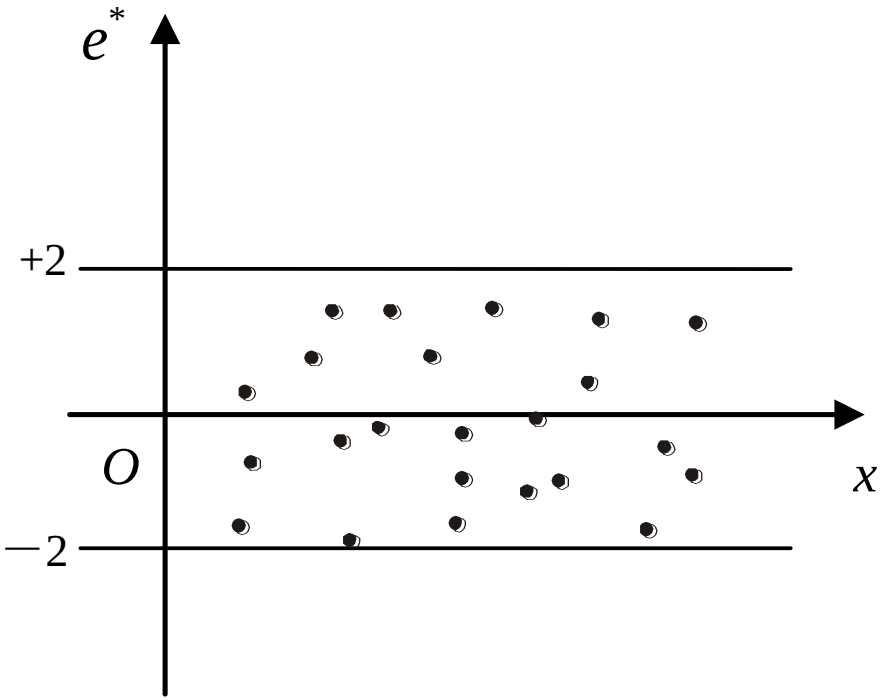
其中h_i为所谓“帽子矩阵”的对角线元素，称为**杠杆值**(Leverage)

标准化残差(Standardized residual)：

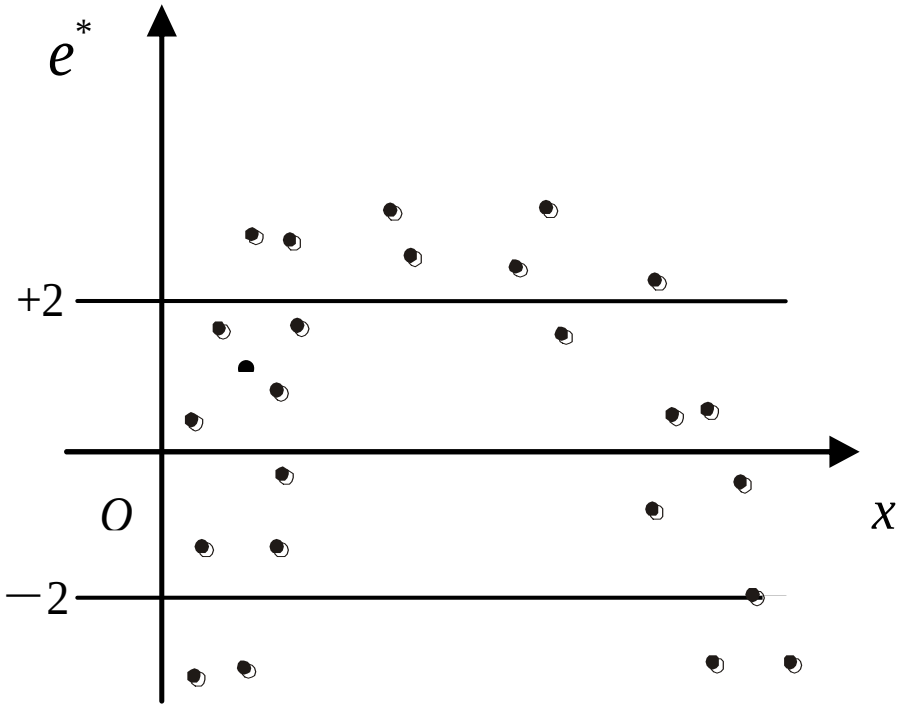
$$e_i^*=\frac{y_i-\hat{y}_i}{s\sqrt{1-h_i}}$$

其中s为误差项ε的标准差σ的估计量，它等于均方误差(MSE)的平方根，称为标准误 (Standard error)。

残差图：残差序列的散点图



模型假设合理



回归方程为曲线
或残差有自相关

残差图：残差序列的散点图

好模型的残差图：

在标准化残差对自变量（或因变量y的拟合值）的残差图中，如果绝大多数点都落在（—2，+2）的水平带状区间之内，且不带有任何系统趋势、完全随机地分布在该带状之中，则说明所采用的回归方程对样本数据的拟合是良好的，随机误差的假设是正确的。

如果残差图出现一些特定的形状，则反映了有些模型假设可能不成立。

第二节 回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

方差齐性：随机误差ε 的方差是相同的
即 $\mathrm{var}(\varepsilon_i)=\sigma^2, i=1,2,\cdots,n$ ，称为**方差齐性**(Homoscedasticity)

否则，则称随机误差ε 的方差具有**异方差性**(Heteroscedasticity)。

模型存在异方差性的可能后果：

1、回归系数估计量的方差会变大，从而使得回归方程不可靠；

2、参数的显著性检验（t–检验）失去意义；

3、预测精度降低。

出现异方差性的可能原因：

1、使用的是处理过的平均数据。

2、因变量的方差与其平均数有关。

3、方差与自变量有关。因变量的方差是自变量的一个函数，在经济模型中经常会碰到。例如，我们要研究住房支出与家庭收入及其他因素的关系，因变量的方差就很可能是家庭收入的函数。

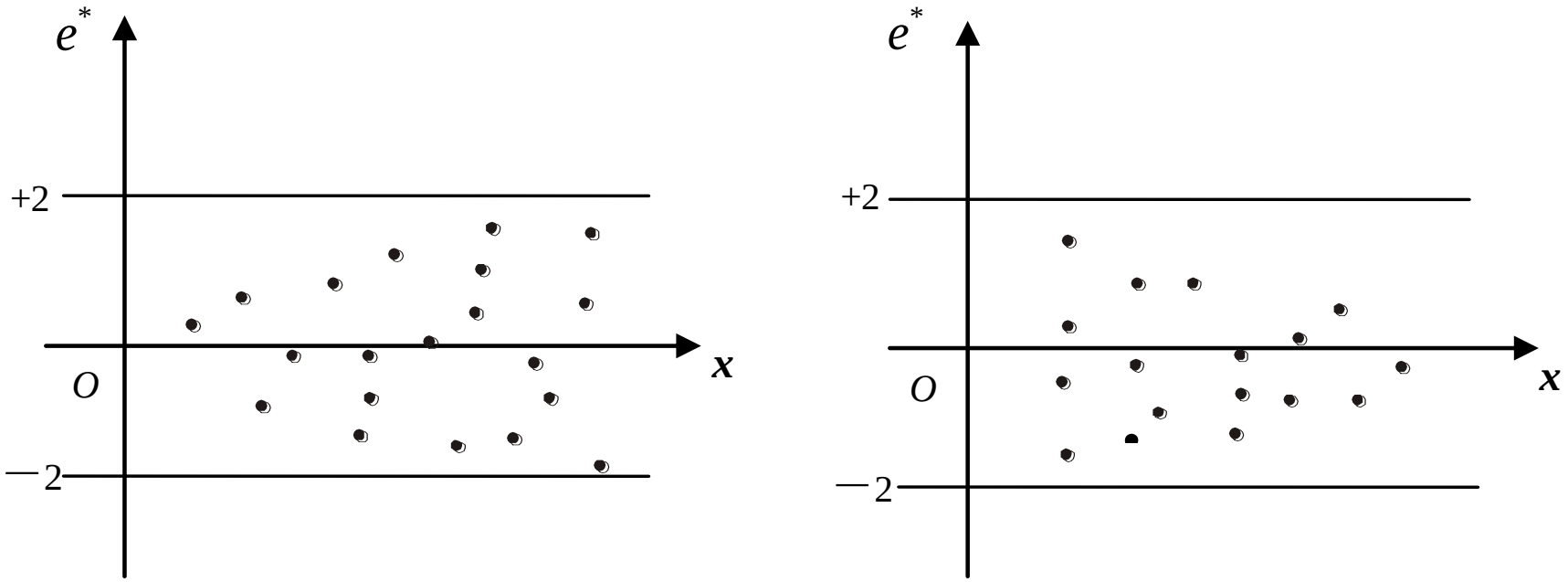
4、数据来源于不同观察者，不同地方或其他不同条件，则其方差很可能不同。

比如我国个人收入的方差，与时间有关系。

异方差性的检验方法：

1、残差图法；

2、检验法（Goldfeld-Quandt检验法、Glesjer检验法和等级相关系数检验法）。



随机误差项具有异方差性的标准化残差图

第二节 回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

自相关：对不同的观察值，误差项ε 是假设相互独立的。如果误差项ε的逐次值之间存在相关性，则称为**自相关**(Autocorrelation)。

$$\varepsilon_t = \rho\varepsilon_{t-1} + a_t, t = 1,2,\cdots$$

其中ρ是绝对值不大于1的参数，a_t是一个均值为0，方差为σ^2 的正态随机变量。

- ✓ 如果ρ不等于0，则表示存在自相关；
- ➢ 如果ρ >0，则表示存在正的自相关，
- ➢ 如果ρ <0，则表示存在负的自相关。

出现自相关的可能原因：

1、因变量本身存在自相关。许多经济变量往往存在自相关，特别是时间数列数据。

2、略去了存在自相关的自变量。在建立回归模型时，不重要的自变量往往被剔除出去。如果被剔除的这些自变量是自相关的，则往往会导致随机误差项的自相关性。

3、回归方程不正确。比如，某商品销售量受季节影响，如果采用线性函数作回归方程，季节波动就被并入了随机误差项，从而导致随机误差在时间上的相关性。

4、重要因素的影响。某些重要因素或事件的影响会持续相当一段时间，如战争、金融危机、自然灾害等，在影响期内，模型的误差项就会呈现自相关。

出现自相关的可能后果：

1、回归方程的最小二乘估计的方差会增大。
2、回归系数的t-检验的标准误不正确，t-检验失效。
3、回归方程不可用于预测。

自相关的检验方法：

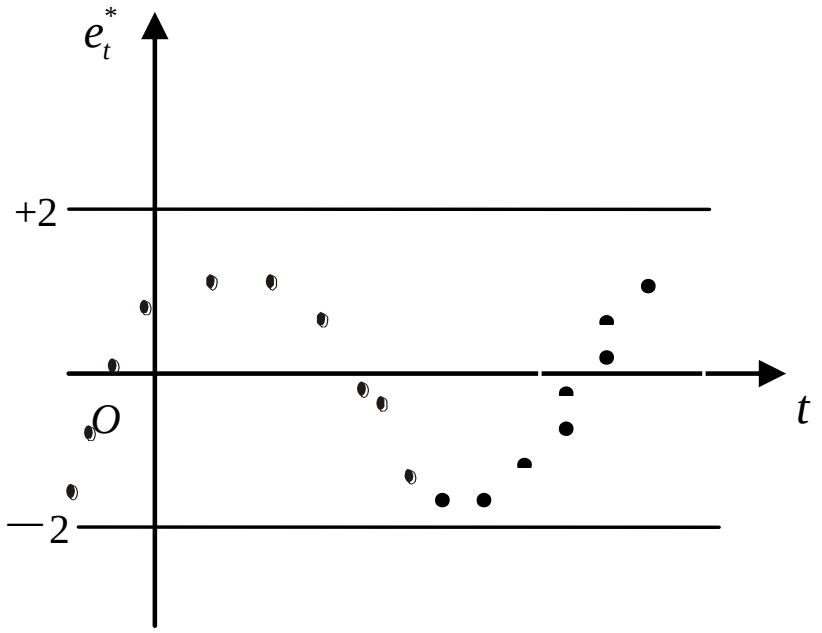
1、残差图法：

把残差（标准化残差）作为纵坐标，把时间（观察值顺序）作为横坐标做残差图。如果按时间出现一定的形状，则表示存在自相关。

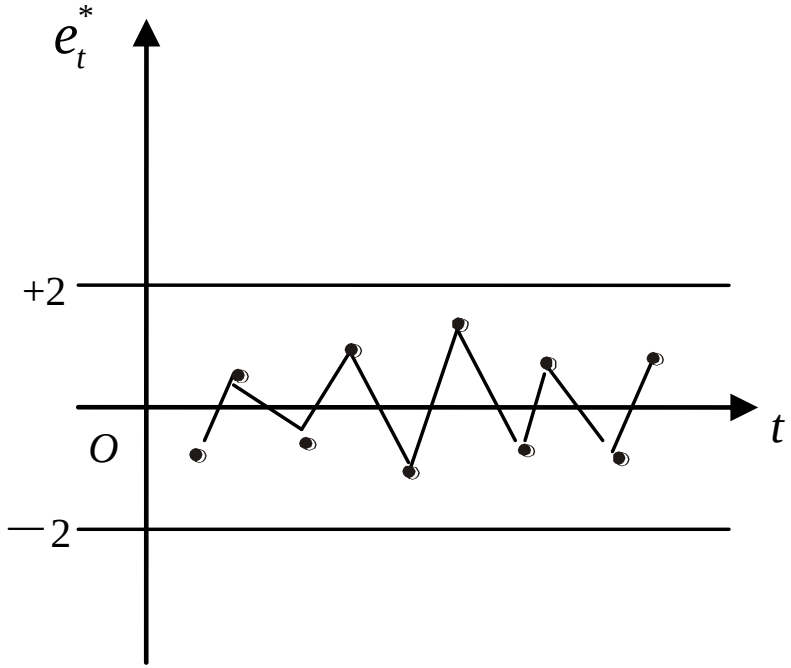
✓ 正相关：如果残差出现几个正的，接着又出现几个负的，呈现循环形状，则该自相关性是正相关。
✓ 负相关：如果残差逐次改变符号，呈现锯齿形状，则该自相关性是负相关。

自相关的检验方法：

1、残差图法：



正自相关



负自相关

自相关的检验方法：

2、Durbin-Watson 检验：

Durbin-Watson检验的统计量：$$d=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2}$$

- 软件Excel没有给出D-W统计量的值，
- Minitab给出D-W统计量的值，但没有给出D-W检验的p-值

自相关的检验方法：

查Durbin-Watson临界值表，查出上限$d_{U,\alpha}$ 和下限$d_{L,\alpha}$，把D-W统计量的计算值d与上下限进行对比，得出结论。

Durbin-Watson统计量的值在0~4之间。

✓一般在2附近表示不存在自相关，

✓在0附近表示有正的自相关，

✓在4附近表示有负的自相关。

自相关的检验方法：

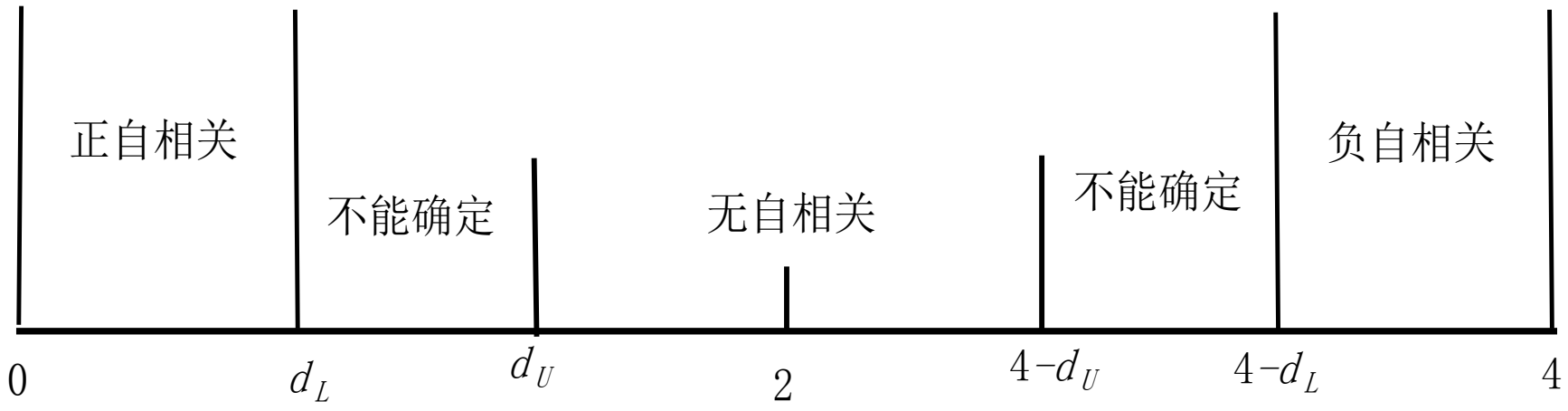
记：D-W统计量的计算值d，上限$d_{U,\alpha}$，下限$d_{L,\alpha}$ 则：

如果$0<d<d_{L,\alpha}$，正自相关；

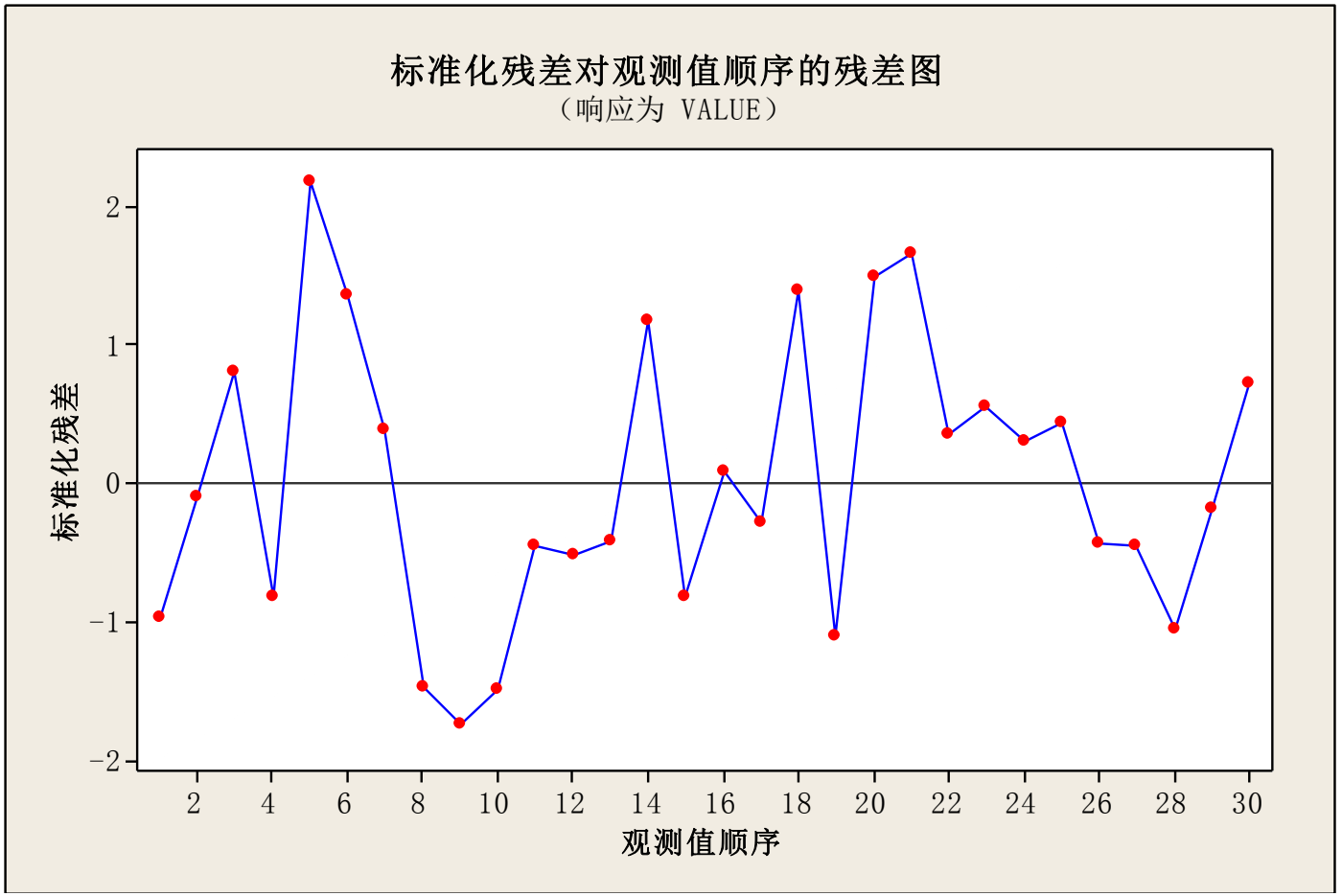
如果$4-d_{L,\alpha}<d<4$，负自相关；

如果$d_{U,\alpha}<d<4-d_{U,\alpha}$，无自相关；

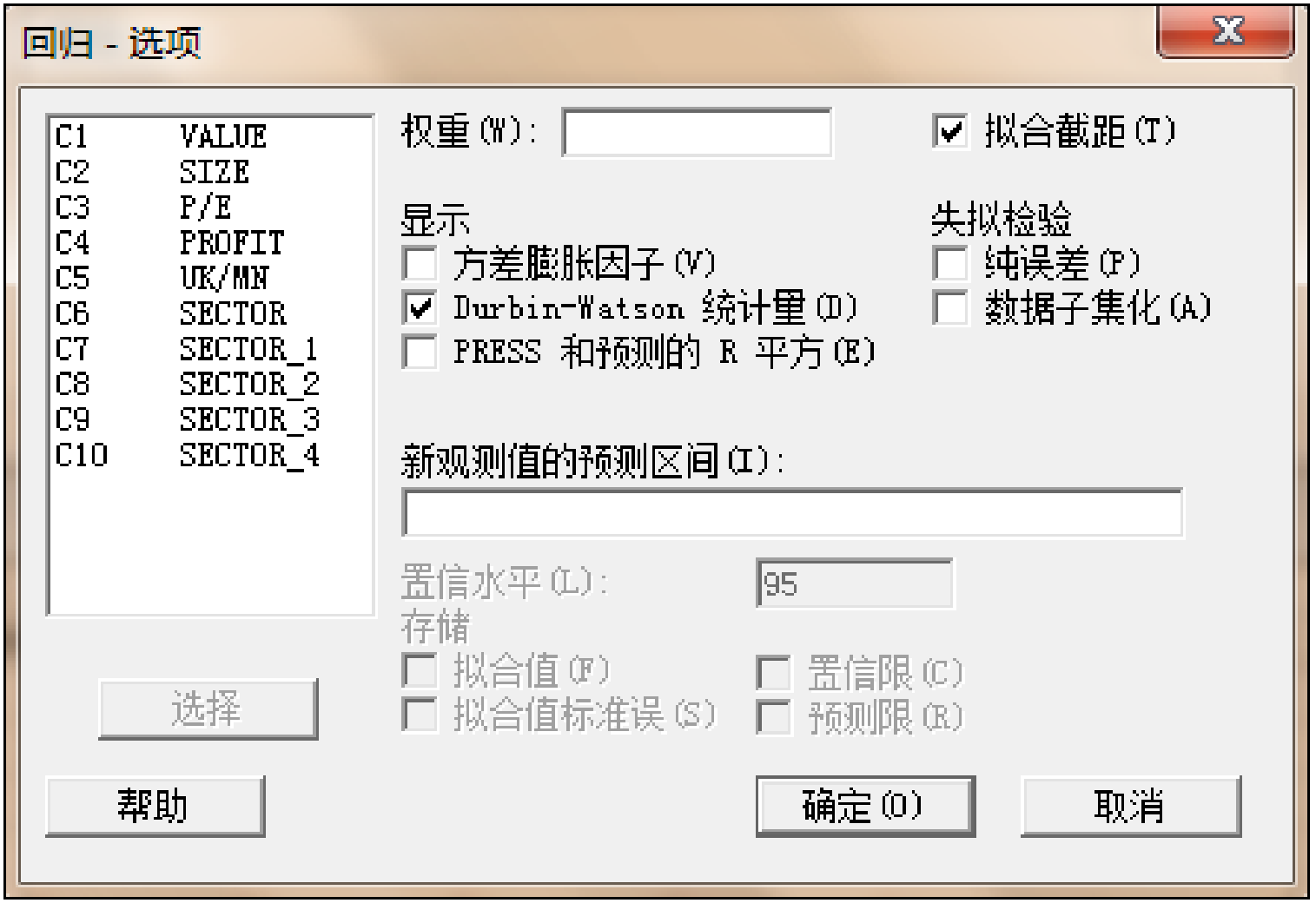
如果$d_{L,\alpha}<d<d_{U,\alpha}$，或$4-d_{U,\alpha}<d<4-d_{L,\alpha}$，不能确定是否存在自相关。



例（公司市值评估），标准化残差对观察值顺序的残差图:



在运行Minitab软件对数据做回归时，可以选择D-W统计量：计算结果d = 1.54987。：



检验：这是30个观察样本，3个自变量的回归问题，如果选取显著性水平为0.05，则通过查Durbin-Watson临界值表，可得临界值的下界dL=1.21，上界 dU=1.65。

计算结果d = 1.54987。

- 因为dL<d<dU，根据正相关的Durbin-Watson检验规则，可知不能得出是否存在正相关的结论。

- 又因为 d<4－dU =2.35，根据负相关的Durbin-Watson检验规则，可知误差项之间不存在负自相关。

第二节 回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

模型的误差项ε 是否服从正态分布?

需要进行检验。

假设： H_0：样本来自正态总体；

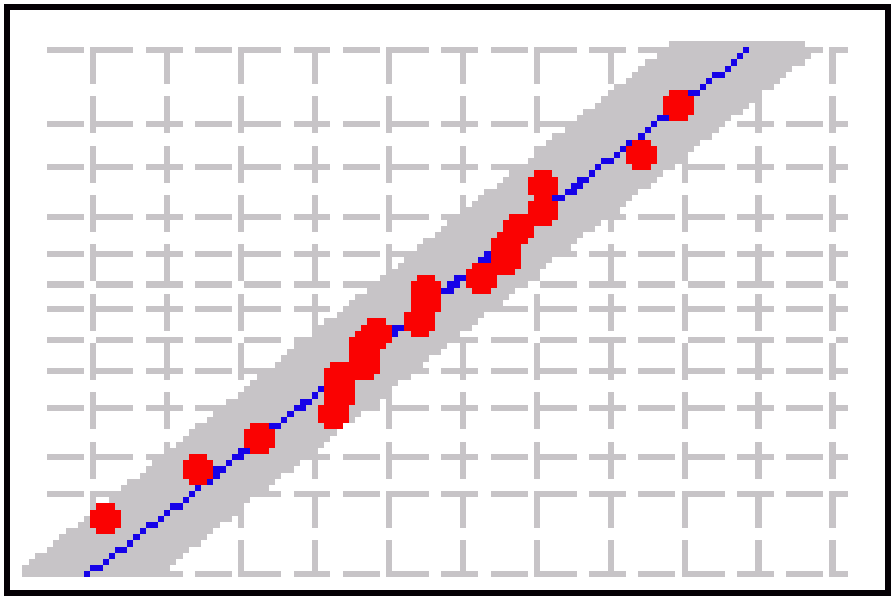
H_1：样本来自非正态总体。

(一)图示法（正态概率图）

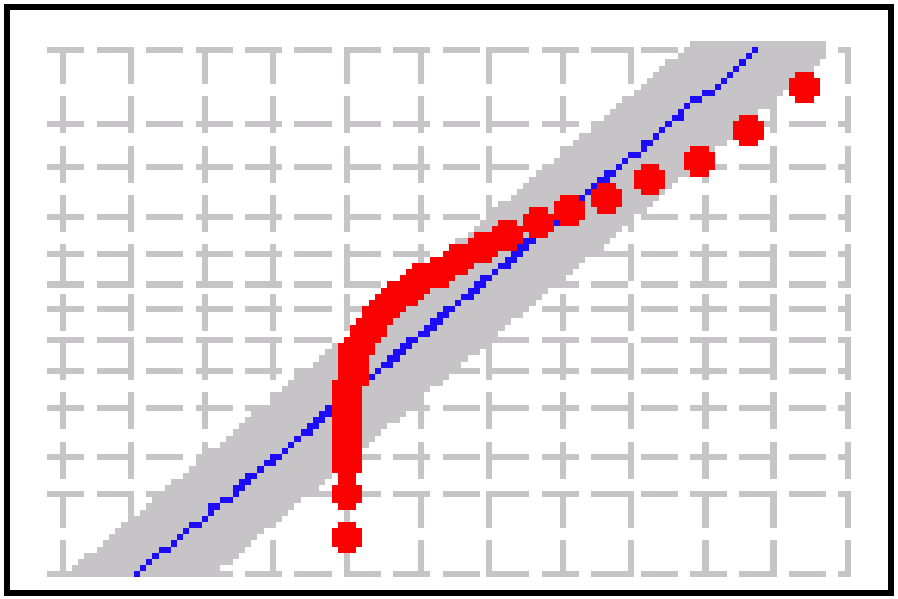
判断总体或残差的正态性，常用的图示法是正态概率图（也叫Q-Q图）。

如果总体呈正态分布，图中的点将大致形成一条直线。一种非正式的近似正态性检验，称为“粗笔检验”，常应用于概率图。想象有一支“粗笔”从拟合线上划过：如果它覆盖了图中的所有数据点，则数据可能为正态分布.

➢正态概率图在Minitab中可以在回归的图形选项中，选择残差的正态图，得到残差的正态性检验；
➢也可以在“统计”的“基本统计量”中选择“正态性检验”，得到变量（如因变量）的正态性检验。



正态数据的概率图



非正态数据的概率图

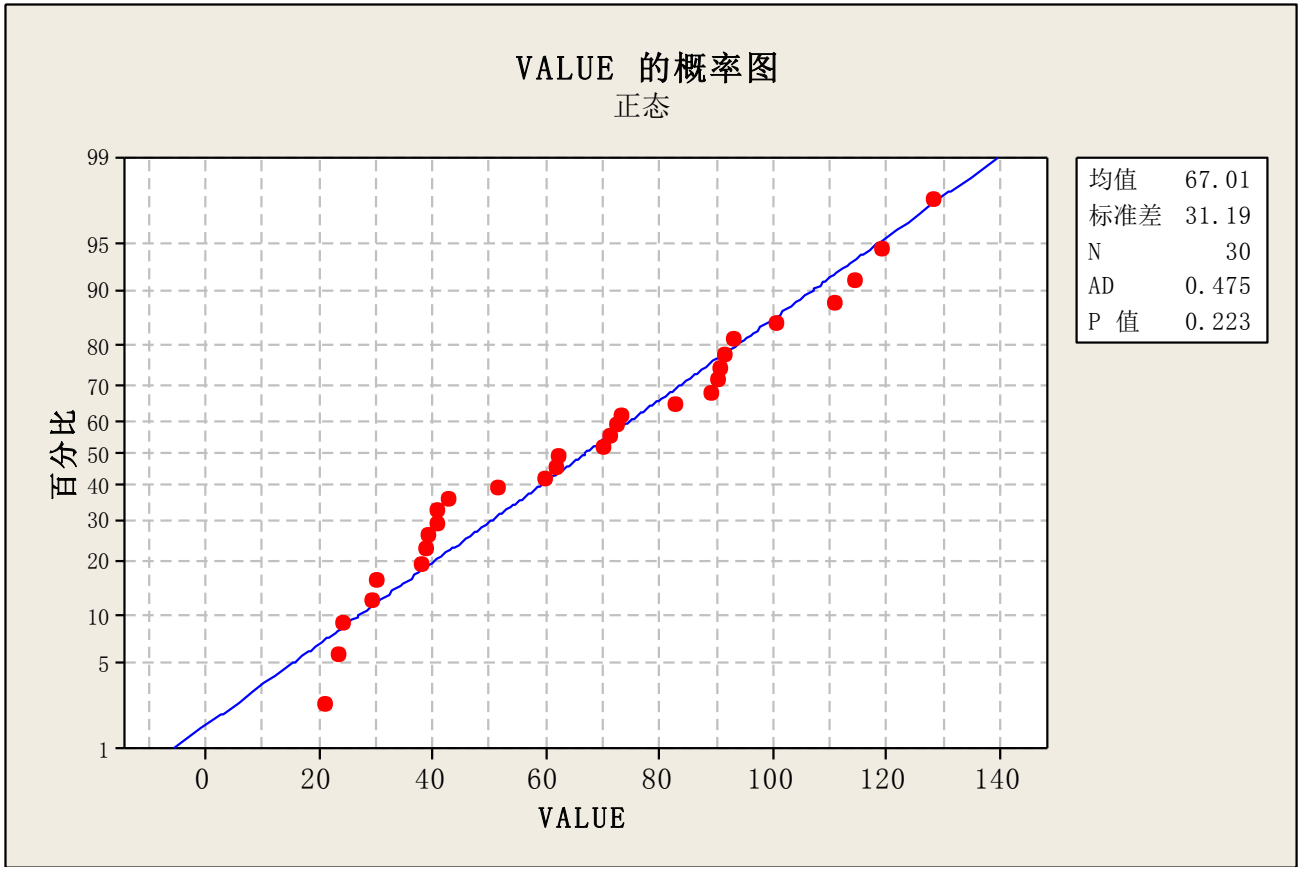
(二)检验法

常用的检验法有：

- Anderson-Darling检验，
- Ryan-Joiner检验
- Kolmogorov-Smirnov检验。

在Minitab中，这些检验都可以在“正态性检验”中选择。 Minitab软件可以给出正态概率图，同时给出检验统计量的值，以及检验的*p*-值。

例8（公司市值数据）：对因变量VALUE进行正态性检验。选择Anderson-Darling检验，结果如下：



所有的点基本上在一条直线上；A-D检验统计量AD=0.475，p-值为0.223，大于0.05，因此应接受H_0，即认为总体的公司市值（VALUE）服从正态分布。

第二节 回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

五、异常点与强影响点

除了模型假设，部分观测数据由于与既定模型有较大偏离，也会对模型的估计和检验有重大影响。

异常点（Outlier）：严重偏离既定模型的数据点；
强影响点（Influential point）：对统计推断的结果影响特别大的点

一个例子：

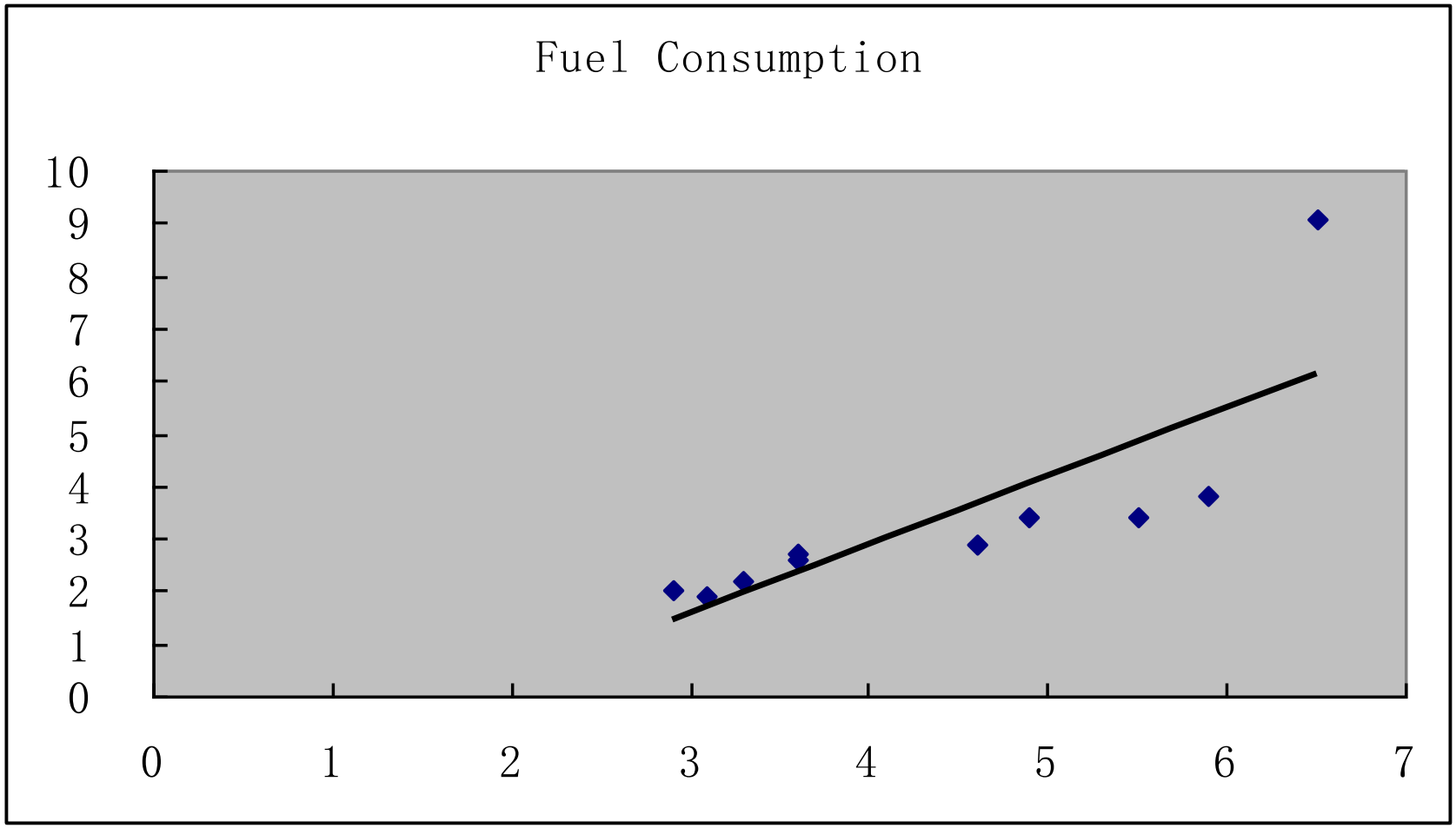
为了汽车自重(X)对油耗(Y)的影响，得到如下数据

Y	3.4	3.8	9.1	2.2	2.6	2.9	2.0	2.7	1.9	3.4
X	5.5	5.9	6.5	3.3	3.6	4.6	2.9	3.6	3.1	4.9

- 回归方程： $Y=-2.33+1.31X$
- t-统计量3.66，P-值 0.01，表示X对Y有显著影响
- 可决系数 R^2 为 0.63，Adjusted R^2 is 0.58.

	Coefficients	标准误差	t Stat	P-value
Intercept	-2.33	1.62	-1.44	0.19
X	1.31	0.36	3.66	0.01

散点图：有一个点明显偏离



删除该点，再做回归：

	Coefficients	标准误差	t Stat	P-value
Intercept	0.320	0.254	1.259	0.248
X	0.589	0.059	9.915	0.000

全部数据做回归：

	Coefficients	标准误差	t Stat	P-value
Intercept	-2.33	1.62	-1.44	0.19
X	1.31	0.36	3.66	0.01

回归方程变化巨大！
对于多元回归，如何判断奇异点和强影响点？
如何处理？

（一）异常点

检测异常点可以通过标准化残差来判断。

Minitab软件可以计算每个观测值的标准化残差，然后把标准化残差的值小于-2和大于2的点界定为异常点，并且用R单独标识。

标准化残差(Standardized residual)：

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}$$

如果数据中出现异常值，必然导致标准误s增大。因此，即使残差显示异常，由于分母s的增大，导致标准化残差偏小，从而不能检测出该异常点。

删除数据的学生化残差：

假设有n个观察数据，现在删除第i个数据，用剩下的（n−1）个数据重新进行回归，得到新的回归系数估计和标准误。把删除第i个数据后得到的标准差误记为$s_{(i)}$，则删除数据的学生化残差(Studentized Deleted Residuals)定义为：

$$e_{(i)}^{*}=\frac{y_i-\hat{y}_i}{s_{(i)}\sqrt{1-h_i}}$$

可以证明：

删除数据的学生化残差服从自由度为(n−p−2)的学生t−分布，其中n为观察值个数，p为自变量个数。

异常点的判断准则：

对于给定的显著性水平α，如果删除数据的学生化残差的绝对值大于$t_{\alpha/2}(n-p-2)$，则可认为该点是异常点。

在Minitab运行回归时，可以在“存储”选项中选择残差，标准化残差和删除数据的学生化残差等。

- 例(公司市值):n=30，p=2，n-p-2=26，以0.05为显著性水平，查表知$t_{0.025}(26)=2.056$，第5个观察值的删除数据的学生化残差等于2.42，大于2.056，所以应视为异常点。

（二）强影响点

要找出对模型结果有严重影响的数据点，首先要定义可以度量这种影响大小的指标。这样的指标常用的有杠杆值(Leverage)，Cook距离(Cook's Distance)和DFITS。

1、杠杆值

杠杆值（h_i）是用于度量某个观测值（x_i）到数据中心（数据集中所有观测值的平均值）之间的距离。

杠杆值h_i介于0和1之间。

高杠杆值点(High leverage):

如果杠杆值很大（接近于1），则标准化残差的分母为 ≈0，此时标准化残差就会很大。一般杠杆值很大的数据点称为高杠杆值点。

因此，杠杆值（h_i）越小，则反映了模型拟合得越好；反之，如果杠杆值（h_i）较大，则第i个观测值拟合的误差较大，对模型的拟合值影响较大。

杠杆值的经验判断标准：

如果杠杆值（h_i）大于$3(p+1)/n$或0.99两者中较小的一个（其中p为自变量数目，n为观测值的数目），则第i个观测值就是高杠杆值点。

Minitab会对每个数据的杠杆值进行计算并比较，对高杠杆值用X标识。

需要注意的是，高杠杆值点有可能是强影响点，但并不意味着一定是强影响点。

2、Cook 距离

删除第i个观测值后，回归方程的估计量会有多大差异？Cook 距离就是度量删除第i个观测值后，回归系数的估计值与原估计值之间的一种距离。

第i个观测值的Cook距离：

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Cook距离在度量第i个观测值对回归系数的影响时，同时考虑了各个观测值的杠杆值和标准化残差。较大的杠杆值或较大标准化残差都可以意味着较大Cook距离。

一个粗略的判断标准是：

如果Cook 距离$D_i>1$，则可认为第i个观测值为强影响点，应进行进一步的分析。

更一般的一个判断标准是：

如果*Cook* 距离$Di > F_{0.5}\,(p+1, n-p-1)$，即*Cook* 距离$D_i$大于 F–分布的中位数，则可认为第i个观测值为强影响点。

3、DFITS

DFITS (即Difference of fits)，是Welsch和Ku于1977年提出的，所以也叫W-K统计量。

删除第*i*个观测值后，Cook 距离比较了回归系数之间的差距，而DFITS则比较第*i*个拟合值之间的差距，通常定义为删除数据前后的第*i*个拟合值之差的标准化变量，也就是删除观测值并重新拟合模型时，拟合值改变的标准差的数量。

第*i*个观测值的DFITS：

$$DFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_i}}$$

DFITS的一个判断标准是：当 $DFITS_i > 2\sqrt{(p+1)/n}$ 时，可认为第 *i*个观测值为强影响点。

对例(美国国会中期选举数据)作两个自变量回归得到的影响诊断度量

年度	“标准选票损失”y	盖洛普对总统的评价x_1	人均年收入纯增量x_2（美元）	杆率值(h_i)	Cook 距离	DFITS
1946	7.3	32	-40	0.845	3.979*	-5.417*
1950	2.0	43	100	0.510	0.000	0.031
1954	2.3	65	-10	0.477	0.083	0.428
1958	5.9	56	-10	0.292	0.371	2.735*
1962	-0.8	67	60	0.447	0.257	0.869
1966	1.7	48	100	0.429	0.034	0.266

第二节 回归诊断

- 一、残差与残差图
- 二、异方差性
- 三、自相关
- 四、正态性检验
- 五、异常点与强影响点
- 六、多重共线性

多重共线性

所谓多重共线性(Multicollinearity)，是指自变量之间存在相关性。一般自变量之间都会有一定的相关性，但如果相关性很高，回归模型的分析就会出现问题。

（一）出现多重共线性的可能后果

如果有两个或以上变量之间完全线性相关，则回归方程的最小二乘估计无法求解。

如果有两个或以上变量之间接近线性相关，则：

- ✓回归系数的估计量会有很大的标准差，因而变得很不稳定，回归方程的结果不可靠；
- ✓回归系数的估计值有可能与实际参数出现相反的符号，从而得出与实际相反的意义解释，如实际上是正比例关系，回归结果却是反比例关系；
- ✓总体显著性F-检验为显著，但是所有单个参数的t-检验都不显著；
- ✓如果删除一个变量或者一个数据，回归方程受很大影响；
- ✓不能确定任一特定自变量对因变量的单独影响；
- ✓可能影响预测结果。

（二）判断是否存在多重共线性的方法

方差膨胀因子法(Variance inflation factor，VIF)

主要度量回归系数的方差的增加幅度。在Minitab软件的回归选项中为可选项。

✓如果VIF=1，表示不存在多重共线性，

✓如果VIF>1，该变量可能存在一定程度的相关性，

✓当VIF介于5到10之间时，表示存在严重多重共线性，回归系数的估计严重不准。

在美国中期国会选举数据的两个自变量回归模型中，两个自变量x_1和x_2的方差膨胀因子VIF都等于1.004，因此可认为不存在共线性。

（二）判断是否存在多重共线性的方法

相关系数法(Correlation coefficient)

计算两个自变量之间的样本相关系数r，粗略的判断准则为：如果$|r|>0.8$，则表示自变量之间可能存在有害的共线性关系。

在美国中期国会选举数据中，两个自变量x_1和x_2的相关系数$r=0.064$，不显著，也证实不存在共线性。

（二）判断是否存在多重共线性的方法

辅助回归法(Auxiliary regressions)

把其中一个自变量作为y，把其余自变量作为x，建立回归模型，如果复可决系数R^2 >0.8，则认为自变量之间可能存在有害的共线性关系。

条件数法(Condition number)

为样本相关阵的最大特征根与最小特征根之比，越大说明共线性越严重。超出范围，这里不讨论。

（三）出现多重共线性的处理办法

- ✓剔除一个或多个高度相关的自变量（如采取逐步回归法）；
- ✓对自变量做适当的变换；
- ✓增加更多与其余自变量不相关的变量；
- ✓增加一些观察数据；
- ✓其他方法：
- 👍主成份回归法(Principal component regression)；
- 👍特征根回归法(Latent root regression)；
- 👍岭回归法(Ridge regression)；
- 👍偏最小二乘估计法(Partial least squares method，PLS)

本节小结

（一）介绍了残差、标准化残差、删除数据的学生化残差以及残差图，这些是回归诊断的基本工具。

（二）介绍了如何对模型的误差假设进行检验

1、异方差性检验(Homoscedasticity)；

2、自相关检验(Autocorrelation)；

3、正态性检验(Normality test)。

（三）介绍了数据中异常点与强影响点的诊断

（四）多重共线性的判断与处理