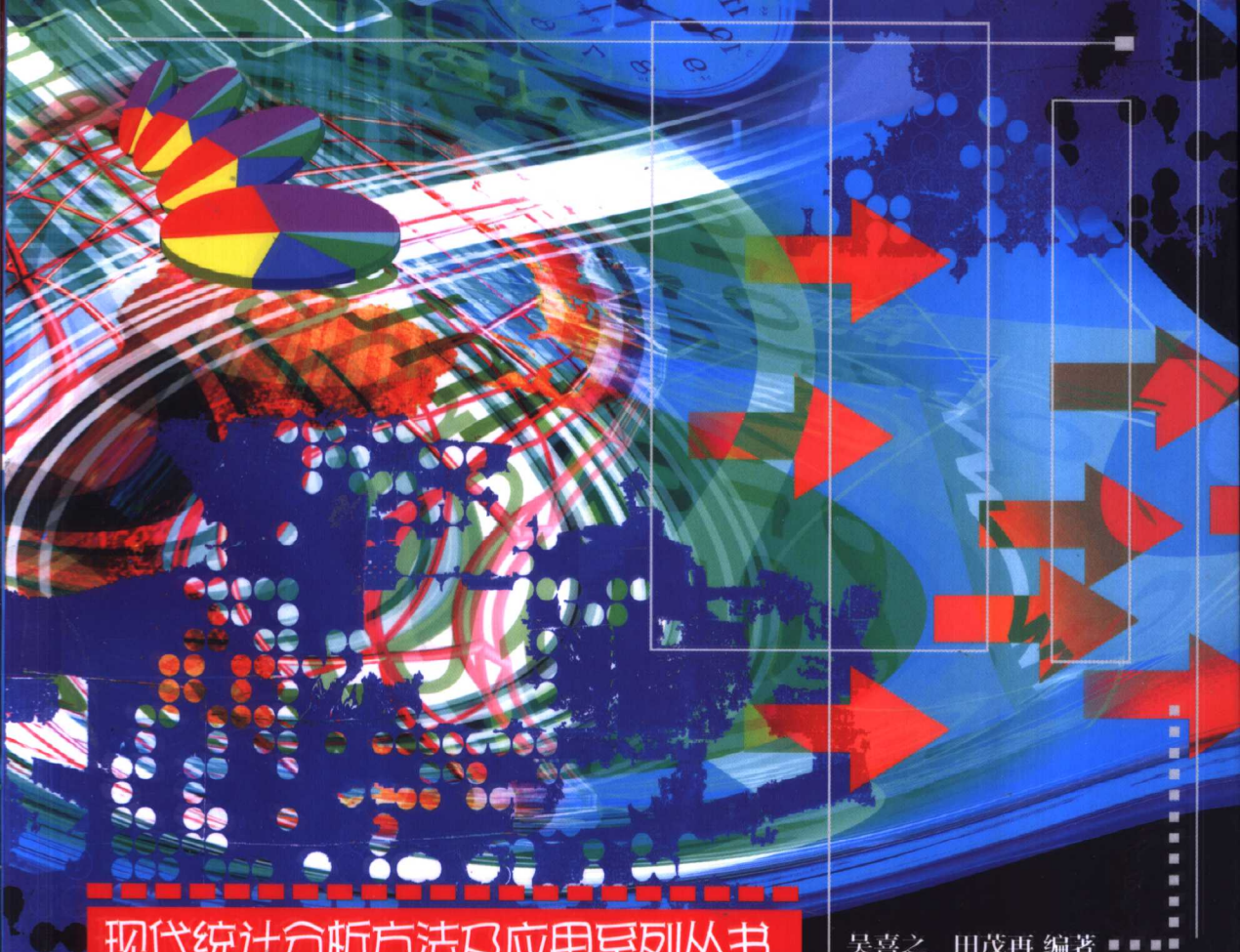


现代统计分析方法及应用系列丛书

吴喜之　田茂再 编著

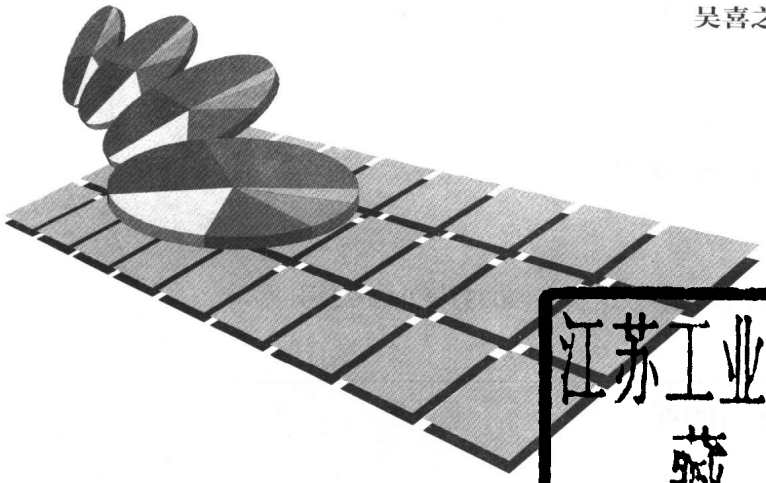
现代回归模型诊断

中国统计出版社
China Statistics Press

现代统计分析方法及应用系列丛书

现代回归模型诊断

吴喜之　田茂再 编著



中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

现代回归模型诊断/吴喜之、田茂再编著.
—北京:中国统计出版社,2003.6
ISBN 7-5037-4134-1

Ⅰ.现…
Ⅱ.①吴… ②田…
Ⅲ.回归分析－统计模型
Ⅳ.O212.1

中国版本图书馆 CIP 数据核字(2003)第 043022 号

作　　者/吴喜之　田茂再
责任编辑/吕　军
封面设计/刘国宁　张建民
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 75 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/(010)63459084,63266600－22500(发行部)
印　　刷/北京顺义兴华印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/18
印　　张/9.75
印　　数/1—3000 册
字　　数/170 千字
版　　别/2003 年 10 月第 1 版
版　　次/2003 年 10 月北京第 1 次印刷
书　　号/ISBN 7-5037-4134-1/O·48
定　　价/18.50 元

前　言

回归分析可以说是应用最广泛的统计方法之一。很多人的第一次实际统计应用就是回归。所有的数学软件都有不同深度的回归分析的计算函数或模块；甚至一些掌中计算器，也有简单回归的功能。一般来说，只要往计算机输入数据，点几下鼠标，就会有大量的计算结果输出。看上去简便快捷。但是，并不是所有的人都意识到在这些漂亮输出背后可能隐藏的潜在错误、危险甚至灾难。

所有统计模型的应用都要求数据满足某种条件。而其中许多条件在统计教科书或文献中仅仅是为了数学上或叙述上的方便而作的假定。任何实际数据在分析之前，都很难说可以对其作任何假定。在对数据的背景还没有弄清楚的时候就匆忙进行回归计算或任何统计计算，都是非常盲目的，要冒风险的。一个模型只反映了应用者对数据背景的认识程度，因而也仅仅是一个近似。当拥有更多的数据之后，人们对数据可能有较多的了解；这时，就会对模型作出改进，以适应对数据背景的新的认识。

国外某些应用领域的软件，比如法律或医药软件，都有大量的警告；在输入案情或病情之后，它绝对不会毫不犹豫地给出明确的指示；在这些软件的使用过程中，随时会提出不少问题和警示。然而，目前的统计软件，只要不出诸如除数是零或奇异矩阵求逆等纯粹数学问题或者数据形式和软件要求的不符之外，不会给予任何统计上的警告。你的命运部分或全部掌握在选项的恰当与否以及你是否能够理解眼花缭乱的计算机输出。更加不幸的是，基于种种原因，统计软件在不同程度上无法处理许多“非规范”数据。这时，

靠选项是无法解决问题的。唯一的办法是根据自己的知识和能力，利用可以编程的软件，对数据进行认识、建模和计算，并作出合理的结论。这种对数据和模型进行识别的过程，就是统计诊断的过程。

本书的目的就是给读者一些现代回归诊断的知识。

本书得到国家门然科学基金的资助。

目　录

第一章　回归诊断引论

第二章　局部影响分析

第三章　异方差性诊断检验

第四章　拟合欠佳诊断检验

1 回归诊断引论

§1.1 回归时怎么会犯错误？

我们先考虑一个简单回归的例子. 变量 (X, Y) 有 20 个观测值已经输入到计算机统计软件中 (比如 SPSS). 按照一些教科书的运作方式; 首先, 我们要进行相关分析. 根据 SPSS 输出, Pearson 相关系数为 0.977, p- 值为 0.000. 看来有很强的线性关系. 然后, 我们再继续作回归. 得到下面结果:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977[a]	.954	.951	.98765

a Predictors: (Constant), X; b Dependent Variable: Y

这里的 R(样本复相关系数) 和 R^2(样本决定系数) 与调整后的 R^2 都接近于 1, 按照标准统计教科书, 这说明拟合很好. 而计算机关于回归的 ANOVA 的输出为

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	361.480	1	361.480	370.579	.000[a]
	Residual	17.558	18	.975		
	Total	379.038	19			

a Predictors: (Constant), X; b Dependent Variable: Y

这从另一个角度说明模型的显著性. 下面另一个常用的关于系数的输出作出了和上面 F- 检验等价的对系数 $\beta_1 = 0$ 的 t 检验也十分显著， p- 值也为 0.000.

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	.436	.225		1.932	.069	-.038	.909
X	.958	.050	.977	19.250	.000	.854	1.063

a Dependent Variable: Y

该表也说明回归直线的截矩为 $\widehat{\beta}_0 = .436$, 斜率为 $\widehat{\beta}_1 = .958$. 还有什么可以担心的呢？正态性吗？好， SPSS 的正态 P-P 图没有表现出任何明显的两端偏离直线的情况. 也就是说，怀疑正态性的证据不足. 还不放心，再看看残差图 (图 1.1); 从残差图，看不出有什么异常；最大的残差为第 14 个观测值 (-2.42), 似

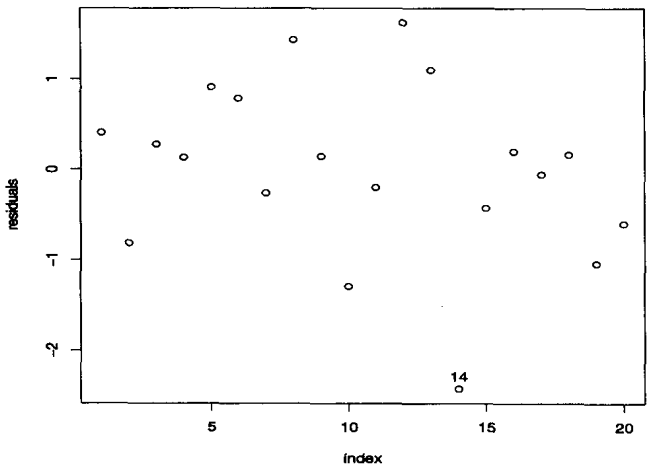


图 1.1 残差图

乎有些显眼. 现在可以得出回归直线为 $y = 0.436 + 0.958x$ 的结论了吗？实际上，这个回归一点意义都没有. 所有的显著性全部都是由一个观测值造成的. 这就是残差很小 (0.40) 的第一个观测值. 如果把这一个观测值删去， Pearson 相关系数就只有 0.161(和原先的 0.977 相比), 而 p- 值从原先的 0.000 升至 0.511. 再进行回归， R 从原来的 0.977 降为 0.161. F 从原来的 370.579(p- 值 =0.000)

降为 0.450(p- 值 =.511). 与这个 F 检验等价的对系数的检验也一点不显著了. 这时的残差最大的两个点为原先的第 12 点 (1.47) 和原先的第 14 点 (-1.46), 完全不突出了.

到底是怎么回事？其实, 只要点一下原始数据的散点图就可以看出原因. 删去第一个观测值的散点图为图 1.2 中的左图; 这里 X 和 Y 都是标准正态分布, 毫不相关. 而没有删去第 1 个观测值的散点图为图 1.2 中的右图;

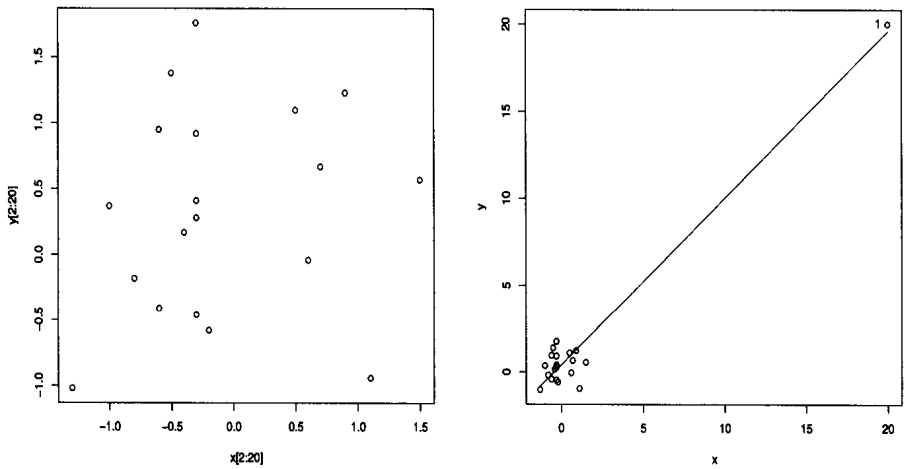


图 1.2 删去第 1 个观测值的原始数据 (左边) 和未删去第 1 个观测值的原始数据 (及拟合的回归直线) 的散点图

可以看出, 那第 1 个观测值就象杠杆一样决定了整个的回归神话. 显然, 如果该点在远离其他点的地方移动, 则回归直线也会随之移动. 这样的点实际上也被称为高杠杆点 (high-levarage-point). 它无疑是模型的影响点 (influential point), 但不是离群点 (outlier). 我们将会引进这些概念.

注意, 这里的例子仅仅是简单的回归, (X,Y) 为二维点阵, 可以用散点图在平面上点出原始数据. 如果是多维数据, 就没有这么方便了. 而高杠杆点是很容易找出来的一类影响点. 有许多类型的影响点根本无法用简单的方法找出.

这些有问题的点的来历很不一样. 有些是源于数据传录中发生的错误, 有些是由于模型不对而产生的 "问题点". 下面看一个真实数据例子. 这是 Ruppert and Carroll (1980) 的海盐 (salinity) 数据.

Obs	Salinity	Lagged salinity	Trend	Water flow	Obs	Salinity	Lagged salinity	Trend	Water flow
y	x_1	x_2	x_3	y	x_1	x_2	x_3		
1	7.6	8.2	4	23.005	15	10.4	13.3	0	23.927
2	7.7	7.6	5	23.873	16	10.5	10.4	1	33.443
3	4.3	4.6	0	26.417	17	7.7	10.5	2	24.859
4	5.9	4.3	1	24.868	18	9.5	7.7	3	22.686
5	5.0	5.9	2	29.895	19	12.0	10.0	0	21.789
6	6.5	5.0	3	24.200	20	12.6	12.0	1	22.041
7	8.3	6.5	4	23.215	21	13.6	12.1	4	21.033
8	8.2	8.3	5	21.862	22	14.1	13.6	5	21.005
9	13.2	10.1	0	22.274	23	13.5	15.0	0	25.865
10	12.6	13.2	1	23.830	24	11.5	13.5	1	26.290
11	10.4	12.6	2	25.144	25	12.0	11.5	2	22.932
12	10.8	10.4	3	22.430	26	13.0	12.0	3	21.313
13	13.1	10.8	4	21.785	27	14.1	13.0	4	20.769
14	12.3	13.1	5	22.380	28	15.1	14.1	5	21.393

其中海水含盐量 (y) 与河水入海量 (x_3) 的散点图为 (图 1.3): 看上去远在右上角的第 16 点是一个离群点, 而其余的点可以进行回归, 以得到一条自左上到右下的回归直线 (如图虚线所示). 其实, 第 16 点完全是一个真实的点, 只是我们的线性模型的假定把它变成了 "离群点." 如果我们用一个简单的开口向上的二次曲线来拟合 (如图中抛物线), 点 16 就不会是离群点了. 当然二次曲线对此模型也过于简单了. 但是这个第 16 点看上去的确有些单薄. 这是因为那一次的雨水和其他记录相比特别的大而造成的. 因此, 在任何情况下都不要对所选择的模型太确信了. 也不能对于不适合我们模型的点随意处理.

以上例子仅仅是回归中最简单的例子. 由于描述数据的模型只是对数据背景分布的某种程度的近似, 而数据也不一定都规范地属于某些简单模型. 在某种假定的模型下寻求对模型的影响点不仅使我们更加深刻地认识数据, 也为选择恰当的模型提供信息. 任何模型都是对现实世界的一种近似, 不存在完美的模型.

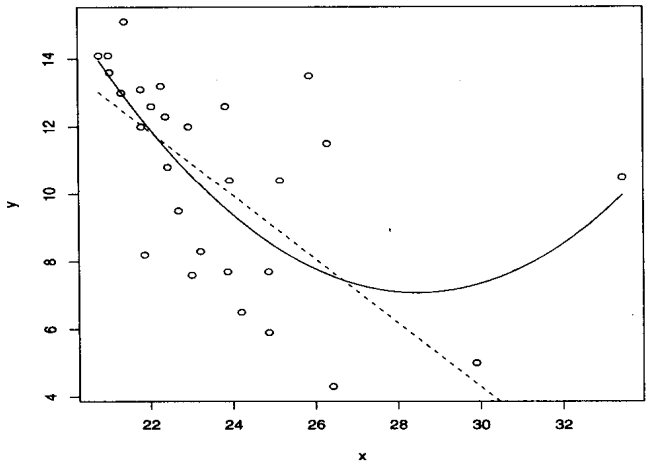


图 1.3 海水含盐量 (y) 与河水入海量 (x_3) 的散点图 (海盐数据)

§1.2 反映线性回归本质的投影矩阵

假定线性回归模型为

$$Y = X\beta + \epsilon.$$

这里 Y 为 $n \times 1$ 响应向量 (或称为因变量), X 为 $n \times k$ 设计矩阵 (X 被称为设计变量、协变量、自变量或解释变量), β 为 $k \times 1$ ($n > k$) 参数矩阵. 我们用 x_i 表示 X 的第 i 行. 称 (y_i, x_i) 为第 i 个观测值. 线性模型通常的假定除了线性之外, 还有为了计算而假定的逆矩阵 $(X^TX)^{-1}$ 存在, 或者等价地假定 $\mathrm{rank}(X) = k$, 即 X 满秩. 此外, 通常的简单线性回归模型还假定: (a) X 不是随机的 (否则考虑给定 X 时的条件概率), (b) ϵ_i 不依赖于 x_i, $(i = 1, ..., n)$, (c) ϵ_i 为独立同分布. 还往往假定正态性: $\epsilon \sim N_n(0, \sigma^2 I)$. 另外, 这里暗含的假定还有, 所有的观测值在决定最小二乘结果和对模型的影响是平等的. 这个线性模型两边取期望后可以得到 $E(Y) = X\beta$.

用最小二乘法, 估计的回归系数为 $\widehat{\beta} = (X^TX)^{-1}X^TY$. 最小二乘进行线性

回归实际上就是把向量 Y 投影到 X 张成的空间上. 投影矩阵为

$$P = X(X^TX)^{-1}X^T.$$

该投影阵又称为预测矩阵 (prediction matrix) 或帽子矩阵 (hat matrix), 也常记为 H (hat), 这是因为它给 Y 戴帽子: $PY = X(X^TX)^{-1}X^TY = X\widehat{\beta} = \widehat{Y}$. 而与预测矩阵正交的投影 $I-P$ 把 Y 投影到和 X 正交的空间上; 这样也就产生了残差 $e = (I-P)Y = Y - \widehat{Y}$. 因此, $Y = PY + (I-P)Y = \widehat{Y} + e$. 对于 σ^2 的估计通常为 $\widehat{\sigma}^2 = \frac{e^Te}{n-k}$.

预测矩阵反映了最小二乘回归的本质. 它有许多重要的性质. 它在非奇异的线性变换 ($E: X \mapsto XE$) 下是不变的. 显然, 预测矩阵 P 和 $I-P$ 是对称和幂等的矩阵. 它的秩 $\text{rank}(P) = k$. 如果 X 按照列划分成两部分: $X = (X_1, X_2)$ 则到 X_1 的投影阵 P_1 和到 X_2 中与 X_1 正交部分的投影阵 P_2 满足 $P = P_1 + P_2$. 记 p_{ij} 为矩阵 P 的第 ij 个元素, 对于 $i, j = 1, ..., n$, 我们有关于 p_{ij} 的一些性质:

(a) 对所有的 i, $0 \le p_{ii} \le 1$;

(b) 对所有的 $i \ne j$, $-0.5 \le p_{ij} \le 0.5$;

(c) 如果 X 有一个常数列, 则对所有的 i, $p_{ii} \ge n^{-1}$, $P\mathbf{1} = \mathbf{1}$. (这里 $\mathbf{1} = (1, ..., 1)^T$.)

(d) 如果 $p_{ii} = 1$ 或者 $p_{ii} = 0$, 则 $p_{ij} = 0$.

(e) $(1-p_{ii})(1-p_{jj}) - p_{ij}^2 \ge 0$.

(f) $p_{ii}p_{jj} - p_{ij}^2 \ge 0$.

(g) $p_{ii} + \frac{e_i^2}{e^Te} \le 1$.

读者可以从这些公式的数学意义来理解预测矩阵的一些性质. 此外, 对于固定的 n, p_{ii} 会随着变量的增加 (X 的列的增加) 而非减. 对于固定的 k, p_{ii} 会随着观测值的增加 (X 的行的增加) 而非增. 矩阵 P 和 $I-P$ 的特征值或者为 0, 或者为 1. 预测矩阵还有许多其他性质, 可以从线性代数的知识很容易得到. 如果对第 i 个观测值, 其相应的 p_{ii} 较大, 就称之为高杠杆点 (high-levarage-point). 高杠杆点是那些远离数据点主体的点. 而通常称残差 e_i 较大的点为离群点 (outlier). 离群点和回归直线的距离较远. 下面对基于残差的点图做一介绍.

§1.3 残差图诊断检验法

§1.3.1 点图诊断和删除法

本节要介绍一些用于回归诊断的点图，其中也包括前一节中已经介绍过的一些图形.

为了进行线性模型的诊断，人们设计了一些根据各种不同原理而得出的度量；这些度量各有侧重，各有其解释. 形成这些度量的统计量，例如残差，都多少从不同方面描述了用数据拟合模型的质量. 如果，所选择的度量对每一个观测值都有一个值，就可以将这些量做出点图. 那些在图中突出的点就可能对模型有较大的影响. 有些度量是通过删除该点之后拟合模型而算得的. 这是一种交叉验证. 如果删去该点，相应的度量变化很大，说明该点是影响点. 一般用下标 (i) 表示用删去该点后拟合出来的模型计算出来的人们感兴趣的统计量. 比如 $e_{(i)}$ 是利用删去第 i 个点之后所建立的模型 (通过删去 i 点后算出的拟合回归直线) 而得到的第 i 点的残差.

如果要删去多个点，也可以推导出相应的度量，以估计多个点的群体影响. 这种删除一个或多个点对模型的影响一般称为总体影响 (global influence). 在进行逐点删除法来寻找影响点时经常遇到两类问题. 一类为掩盖 (masking) 现象. 这时，在诊断中一些影响点互相掩盖. 删去其中的部分点时，看不出任何异常. 而只有把这些互相掩盖的点全部删除，才会发现问题. 另一类现象为淹没 (swamping) 现象. 这时，许多正常点在诊断中都显示出异常，无法区别真假影响点.

§1.3.2 残差与残差图

残差 (residual) 简单地说就是观察值与拟合值之差，它是研究回归诊断的最基本、最重要的工具之一. 回归中有一个重要分支就是残差分析. 除了最原始的残差之外，为了种种目的，人们定义了各种不同形式的残差. **残差图 (residul plot)**，就是以某种残差为纵坐标，某个变量 (如：因变量的拟合值，某个协变量，观察时间或观察序列指标等等) 为横坐标的散点图. 在传统的统计诊断中，残差图的作用主要包括：

(1) 诊断出实际数据是否与既定模型 (postulated Model) 有较大的偏离，试图找出影响拟合的离群点 (outlier) (远离其它绝大多数观测值的点) 、强影响点 (influential point) (对统计推断影响特别大的点) 和高杠杆点 (high leverage point).

(2) 诊断出既定模型本身的假设是否与实际相符合.

(3) 诊断出统计方法是否有问题.

在传统的模型诊断中不同的残差图有不同的作用，例如，残差 - 协变量图可能发现系统偏离；残差 - 序列指标图可能用来研究序列相关性；残差 - 拟合值图可以诊断异方差性等等. 当然，对不同残差图的解释也是一个很好的话题，这一点在下一节有讨论.

在做回归时，通常有一些假定. 当然在具体做散点图的时候，要考虑到这些假定：比如，假定残差向量的各个分量是 (或近似地是) 正态变量时，如果取残差为学生化残差来研究，它是对普通残差做某种标准化后所得的，其分量应该为为近似独立同分布的标准化正态变量.

下面我们以学生化残差为例，说明残差图的具体应用. 内部学生化残差 (internally Studentized residual) 定义为

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1-p_{ii}}}.$$

除了通常的残差 e_i 之外，它把高杠杆点的影响因素通过分母中的 $1-p_{ii}$ 考虑进来. 如果 p_{ii} 大，则 r_i 也会变大. 而外部学生化残差 (externally Studentized residual) 把内部学生化残差中的 σ 代替为删值后的 $\sigma_{(i)}$，它有许多等价达的表达式：

$$\begin{aligned} r^*_{(i)} &= \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-p_{ii}}} = r_i\sqrt{\frac{n-k-1}{n-k-r_i^2}} = \frac{e_i}{\sqrt{\dfrac{(1-p_{ii})SSE_{(i)}}{n-k-1}}} \\ &= \frac{e_i}{\sqrt{\dfrac{1-p_{ii}}{n-k-1}\left(SSE-\dfrac{e_i^2}{1-p_{ii}}\right)}} = \frac{\dfrac{e_i}{\sqrt{e^Te}}\sqrt{n-k-1}}{\sqrt{(1-p_{ii})-\dfrac{e_i^2}{e^Te}}} \end{aligned}$$

诊断原理： 我们以学生化残差为纵坐标，以拟合值 ($\widehat{y}$) 为横坐标作残差图；那么，由于学生化残差的分量 (按照模型假设) 为近似独立同分布的标准化

正态变量，这些残差可以看作是来标准正态变量的样本. 由标准正态分布的性质可知：大约有 68.3% 的学生化残差落在 $[-1,1]$ 中，有 95.5% 的学生化残差落在 $[-2,2]$ 中，有 99.7% 的学生化残差落在 $[-3,3]$ 中. 其次，学生化残差按假设与拟合值相关性很小. 所以，残差图中的点应该绝大部分落在学生化残差为 ± 2 之间，并且，不呈现任何有规律的趋势. 图 1.4 表示了四个学生化残差 r_i 对拟合值 $\hat{y}$ 散点图.

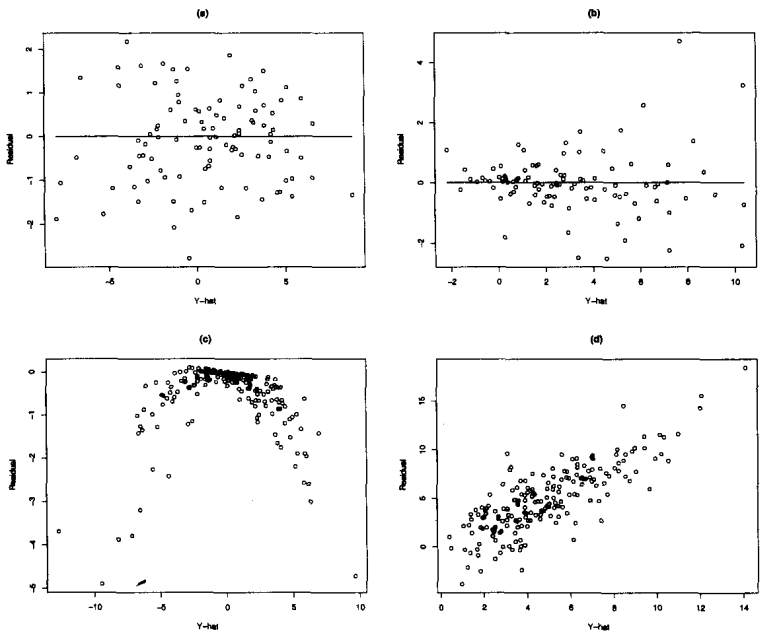


图 1.4 学生化残差 r 对拟合值 $\hat{y}$ 散点图：(a) 是正常的，看不出误差项明显违背假设的迹象；(b) 表明误差具有异方差性；误差方差随拟合值的增大而增大；(c) 表示模型本身具有非线性的趋势. 图 (f) 说明拟合值与学生化残差有线性正相关性.

§1.3.3 残差图误区

残差图是回归诊断的基本而且重要的工具，它具有直观明了、简单易行等优点. 其实，对残差以及其他诊断统计量的作图是一种综合方法，这些图有可能帮助人们在对数据进行分析的早期诊断中找出数据中的离群点、强影响点和