回归模型的诊断

在本章中, 我们将讨论用于检查回归模型充分性的精确诊断. 本章讨论的几个重点是

- 为了计算精度而进行的再参数化
- 强影响观测值的剔除
- 多重共线性问题

回顾: 可以检查模型中预测变量的残差图来判断模型是否需要该变量的效应.

然而单纯使用残差图可能会: 考虑到模型中的其他预测变量, 这些残差图可能无法正确显示预测变量的边际效应的性质.

1 预测变量的模型充分性: 添加变量图

定义 (增加变量图)**.** 也称为偏回归图*(partial regression plots)*和调整变量图*(adjusted variable plot).* 目的是考虑到模型中已有的其他预测变量, 提供关于预测变量的边际重要性的图形信息.

注记. *R*语言`car`包中提供了大量的函数, 大大增强了拟合和评价回归模型的能力.

函数	目的
`qqPlot ()`	分位数比较图
`durbinWatsonTest ()`	对误差自相关性做 Durbin-Watson 检验
`crPlots ()`	成分与残差图
`ncvTest()`	对非恒定的误差方差做得分检验
`spreadievelPlot()`	分散水平检验
`outlierTest()`	Bonferroni离群点检验
`avPlots ()`	添加的变量图形
`inluencePlot ()`	回归影响图

这些残差相互之间的关系图

1. 显示了该变量在减少残差变异性方面的边际重要性
2. 可以提供有关预测因子的边际回归关系性质的信息, 以便纳入回归模型

为了使这些想法更具体, 我们考虑一个一阶多元回归模型, 它有两个预测变量X1和X2. 扩展到两个以上的预测变量是直接的. 假设我们关心X1的回归效应的性质, 假设X2已经在模型中, 对Y在X2上进行回归, 得到拟合值和残差:

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2}, \quad e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

这里的符号明确地表示了拟合模型中的响应变量和预测变量. 我们也将X1回归到X2，得到:

$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$
$$e_i(X_1 \mid X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

则预测变量X_1的附加变量图由Y残差$e(Y|X_2)$与X1残差$e(X_1|X_2)$的图组成。

包含了我们示例中的几个原型添加变量图，其中X2已经在回归模型中，X1正在考虑添加。

写成代码就是

```
fit2 = lm(Y~X2)
fit12 = lm(X1~X2)
plot(fit12$resi,fit2$resi, main='Added Variable Plot for X1')
abline(lm(fit2$resi~fit12$resi));
```

因此若图显示了一条水平带，表明X_1中除了X_2中包含的信息外，没有其他信息对预测Y有用，因此在这里将X_1添加到回归模型中是没有帮助的。

若图显示了斜率非零的线性带。这幅图表明，X_1中的线性项可能是对已经包含X_2的回归模型的有益补充。可以证明如果将该变量加入到已经包含X2的回归模型中, 经过原点的最小二乘线拟合到所绘残差的斜率为b_1，即X_1的回归系数.

若图显示了一条曲线带，表明在回归模型中加入X_1可能是有帮助的，并通过所示的模式表明了曲率效应的可能性质。

1.1 体脂含量的例子

- Y: 体脂含量百分比
- X_1: 褶皱厚度
- X_2: 大腿围
- X_3: 上臂围

用老师PPT中代码可以画出图片如图1所示.

若直接使用R语言中自带的`avPlots()`, 可以直接给出如下图片如图2所示.

1.2 算法

Step 1:

对响应变量Y和解释变量X_k, 若模型中已经含有$X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1}$, 得出残差

$$e_i(Y|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1}) = Y_i - (b_0 + b_1 X_{i1} + \cdots + b_{k-1} X_{i,k-1} + b_{k+1} X_{i,k+1} + \cdots + b_{p-1} X_{i,p-1})$$

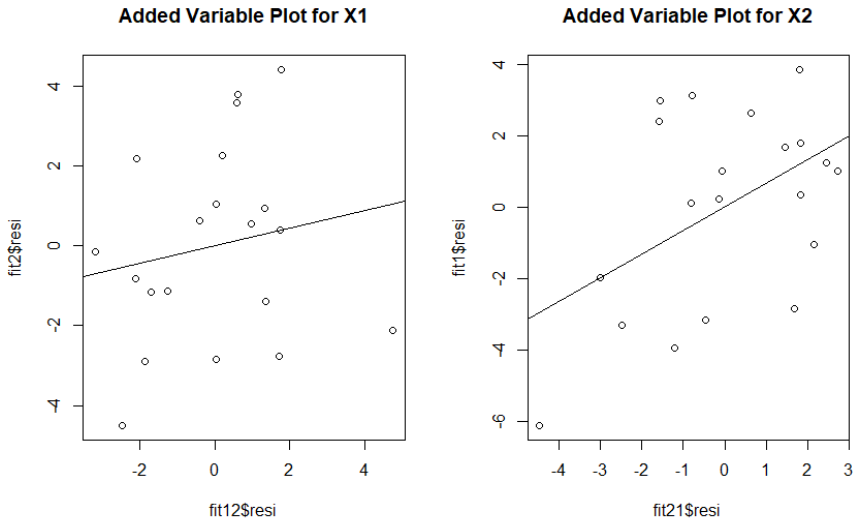

图 1: 用PPT中的代码实现体脂含量的增添变量图

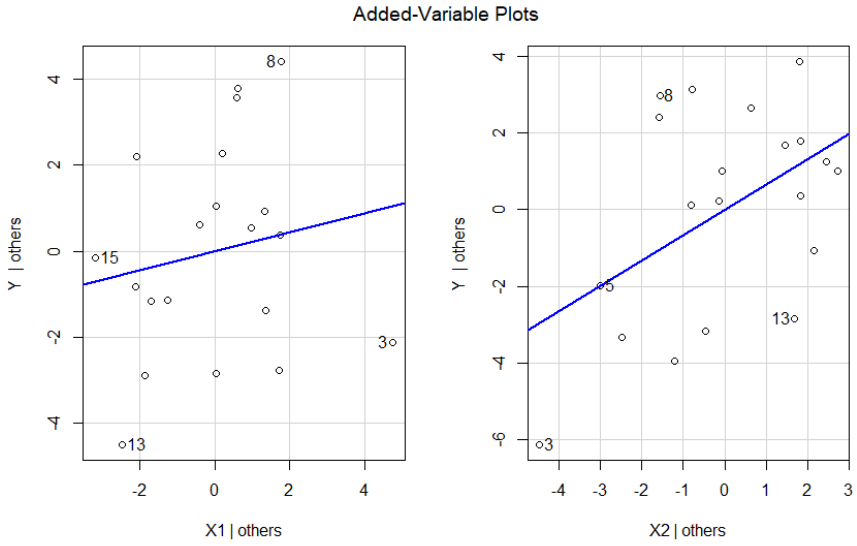

图 2: 用`avPlots()`代码实现体脂含量的增添变量图

Step 2:

以X_k为响应变量, $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1}$为解释变量进行回归, 得出残差

$$e_i(X_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1}) = X_{ik} - (b_0^* + b_1^* X_{i1} + \cdots + b_{k-1}^* X_{i,k-1} + b_{k+1}^* X_{i,k+1} + \cdots + b_{p-1}^* X_{i,p-1})$$

Step 3:

以$e_i(Y|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1})$为 Y轴, $e_i(X_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1})$为$X$轴画出残差图.

2 Y离群值的识别

2.1 学生化残差

回顾残差的定义与性质,

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}},\ \underset{n\times 1}{\boldsymbol{\varepsilon}} \sim N\left(\underset{n\times 1}{\mathbf{0}}, \sigma^2\ \underset{n\times n}{\mathbf{I}}\right)$$

模型本身的偏倚 (不能被观测到): $\varepsilon_i = Y_i - \left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right)$

残差 (可以被观测到)

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(b_0 + b_1 X_{i1} + \ldots + b_{p-1} X_{i,p-1}\right)$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \quad \text{where } \underset{n\times n}{\mathbf{H}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

$$\mathbf{E}\{e\} = \mathbf{E}\{(\mathbf{I} - \mathbf{H})\mathbf{Y}\} = (\mathbf{I} - \mathbf{H})\mathbf{E}\{\mathbf{Y}\} = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = (\mathbf{I} - \mathbf{H})\boldsymbol{\sigma}^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

$$\underset{n\times 1}{\mathbf{e}} \sim N\left(\underset{n\times 1}{\mathbf{0}}, \sigma^2(\underset{n\times n}{\mathbf{I}} - \underset{n\times n}{\mathbf{H}})\right)$$

$E\{e_i\} = 0$; 令 h_{ij} 为帽子矩阵 $\underset{n\times n}{\mathbf{H}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$的$(i,j)^{th}$元素

$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, $\sigma\{e_i, e_j\} = -h_{ij}\sigma^2 \quad \forall i \neq j$

$s^2\{e_i\} = \text{MSE}(1 - h_{ii})$, $s\{e_i, e_j\} = -h_{ij}MSE \quad \forall i \neq j$

学生化残差:

$$r_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

其中, $MSE = \dfrac{\sum e_i^2}{n - p}$

半学生化残差:

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

2.2 学生化去除残差

X_i, Y_i没有用来拟合模型, 定义

$$d_i = Y_i - \hat{Y}_{i(-i)}$$

其中, $\hat{Y}_{i(-i)}$表示去除i项后回归的拟合值.

$$b_{(-i)} = \left(X'_{(-i)}X_{(-i)}\right)^{-1} X'_{(-i)}Y_{(-i)} \sim N\left(\beta, \sigma^2\left(X'_{(-i)}X_{(-i)}\right)^{-1}\right)$$

$$\hat{Y}_{i(-i)} = b_{0(-i)} + b_{1(-i)}X_{i1} + \cdots + b_{p-1(-i)}X_{i,p-1} = x'_i b_{(-i)}$$

其中$x'_i = (1, X_{i1}, \cdots, X_{i,p-1})$, 方差为

$$\begin{aligned}
&\operatorname{var}\{d_i\} = \operatorname{var}(Y_i) + \operatorname{var}\left(\hat{Y}_{i(i)}\right) = \sigma^2 + \operatorname{var}\left\{\mathbf{x}'_i\mathbf{b}_{(i)}\right\} \\
&= \sigma^2 + \mathbf{x}'_i \operatorname{var}\left\{\mathbf{b}_{(i)}\right\}\mathbf{x}_i = \sigma^2\left[1 + \mathbf{x}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(\mathrm{i})}\right)^{-1}\mathbf{x}_i\right] \\
&s^2\{d_i\} = \mathrm{MSE}_{(i)}\left[1 + \mathbf{x}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(\mathbf{i})}\right)^{-1}\mathbf{x}_i\right]
\end{aligned}$$

学生化去除残差

$$t_i = \frac{d_i}{S(d_i)} = \frac{e_i}{\sqrt{MSE_{(-i)}(1-h_{ii})}},$$

其中 $d_i = Y_i - \hat{Y}_{i(-i)}$ 由于 $d_i = \dfrac{e_i}{1-h_{ii}}$, 有$MSE_{(i)} = \dfrac{(n-p)MSE - \frac{e_i^2}{1-h_{ii}}}{n-p-1}$ 故t_i可以写成

$$t_i = e_i\left(\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right)^{\frac{1}{2}}$$

由于学生化是用了$n-1$个观测值, 有p个参数, 故

$$t_i \sim t(n-p-1), i = 1, \cdots, n.$$

2.3 Y离群点的检验

单点判别准则: 当

$$|t_i| \geq t_{1-\frac{\alpha}{2}}(n-p-1)$$

表明第i个观测值是离群值.

注记. 上述*Bonferroni*方法可能会遗漏除最明显的异常值外的所有异常值，特别是当n很大时。因此，建议对任何带有删除的学生化残余> 3的观察都要持怀疑态度。

用*Bonferroni*方法的判别准则为

$$|t_i| \geq t_{1-\frac{\alpha}{2n}}(n-p-1) \quad \text{认为} Y_i \text{是离群值}$$

2.4 给出的代码

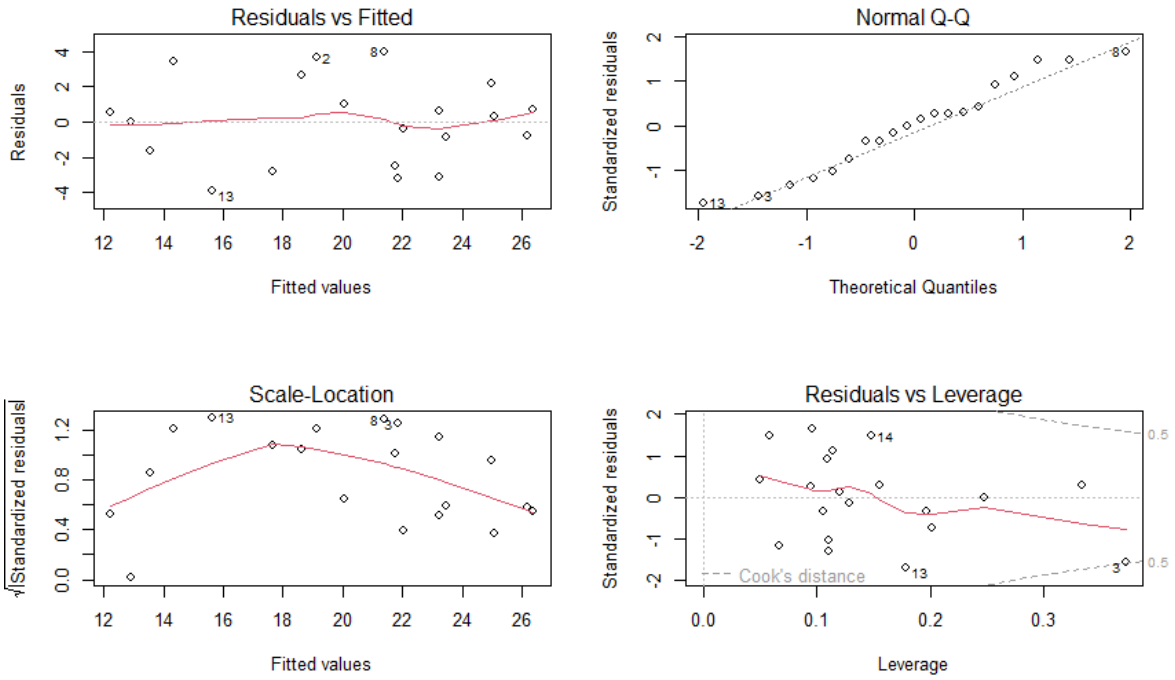

图 3: 标准方法的四种检验图

为理解这些图形, 我们来回顾一下OLS回归的统计假设。

- (正态性) 当预测变量值固定时, 因变量成正态分布, 则残差值也应该是一个均值为 0 的正态分布。“正态Q-Q图” (Normal Q-Q, 右上) 是在正态分布对应的值下, 标准化残差的概率图。若满足正态假设, 那么图上的点应该落在呈 45 度角的直线上; 若不是如此, 那么就违反了正态性的假设。

- (独立性) 你无法从这些图中分辨出因变量值是否相互独立, 只能从收集的数据中来验证。上面的例子中, 没有任何先验的理由去相信一位个体的体脂会影响另外一位个体的体脂。假若你发现数据是从一个特定的群体抽样得来的，那么可能必须要调整模型独立性的假设。

- (线性) 若因变量与自变量线性相关, 那么残差值与预测（拟合）值就没有任何系统关联。换

句话说, 除了白噪声, 模型应该包含数据中所有的系统方差。在“残差图与拟合图”（Residuals vs Fitted, 左上) 中可以清楚地看到一个直线关系, 这暗示着一次回归模型的正确性。

- (同方差性) 若满足不变方差假设, 那么在“位置尺度图”（Scale-Location Graph, 左下) 中, 水平线周围的点应该随机分布。该图似乎不满足此假设。

最后一幅“残差与杜杆图”（Residuals vs Leverage, 右下) 提供了你可能关注的单个观测点的信息。从图形可以鉴别出离群点、高杜杆值点和强影响点。下面来详细介绍。

- 一个观测点是离群点, 表明拟合回归模型对其预测效果不佳 (产生了巨大的或正或负的残差。
- 一个观测点有很高的杜杆值, 表明它是一个异常的预测变量值的组合。也就是说, 在预测变量空间中, 它是一个离群点。因变量值不参与计算一个观测点的杜杆值。
- 一个观测点是强影响点 (influential observation), 表明它对模型参数的估计产生的影响过大, 非常不成比例。强影响点可以通过Cook距离即Cook's D统计量来鉴别。其画的线是一条F分布中位数的等高线.

```
> p = 3
> elist = fit$resi;  SSE = sum(elist^2) #$
> X = cbind(1,X1,X2)
> hlist = diag(X%*%solve(t(X)%*%X)%*%t(X))
> tlist = elist*((n-p-1)/(SSE*(1-hlist)-elist^2))^(1/2)
> cbind(elist,hlist,tlist)
           elist      hlist          tlist
1  -1.6827093112 0.20101253 -0.7299854027
2   3.6429311788 0.05889478  1.5342541325
3  -3.1759701405 0.37193301 -1.6543295725
4  -3.1584651200 0.11094009 -1.3484842072
5  -0.0002886579 0.24801034 -0.0001269809
6  -0.3608155187 0.12861620 -0.1475490938
7   0.7161991891 0.15551745  0.2981276214
8   4.0147327554 0.09628780  1.7600924916
9   2.6551057360 0.11463564  1.1176487404
10 -2.4748115410 0.11024435 -1.0337284208
11  0.3358063798 0.12033655  0.1366610657
12  2.2255110139 0.10926629  0.9231785040
13 -3.9468613463 0.17838181 -1.8259027246
14  3.4474561945 0.14800684  1.5247630510
15  0.5705871038 0.33321201  0.2671500921
16  0.6422984777 0.09527739  0.2581323416
17 -0.8509464751 0.10559466 -0.3445090997
18 -0.7829198812 0.19679280 -0.3344080836
19 -2.8572887647 0.06695419 -1.1761712768
20  1.0404487275 0.05008526  0.4093564171
> max(abs(tlist));  qt(0.9975,n-p-1)
[1] 1.825903
[1] 3.251993
```

3 X离群值的识别: 依靠帽子矩阵对角线上的值

回顾帽子矩阵的若干性质,

帽子矩阵

$$H = X(X'X)^{-1}X' = (h_{ij})_{n\times n}$$

两参数简单线性回归的情况

$$h_{ij} = \mathbf{x}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_j = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)}{SS_{XX}}$$

$$h_{ii} = \mathbf{x}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_i = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{SS_{XX}} \Rightarrow h_{ii} > 0 \text{ 且 } \sum_{i=1}^{n} h_{ii} = 2, \text{ 当 } n > 1$$

一般多元情况

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}, \quad h_{ij} = \mathbf{x}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_j, \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

$$\sum_{i=1}^{n} h_{ii} = \text{trace}(\mathbf{H}) = \text{trace}\left(\mathbf{I}_p\right) = p$$

$$\underset{n\times n}{\mathbf{H}}\,\underset{n\times p}{\mathbf{X}} \times \mathbf{p} = \underset{n\times p}{\mathbf{X}} \Rightarrow \sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$$

$$\underset{n\times n}{\mathbf{H}} = \underset{n\times n}{\mathbf{H}}\,\mathbf{H} \times \mathbf{n} \Rightarrow h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 \geq 0$$

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H} \Rightarrow 1 - h_{ii} = \sum_{j=1}^{n} \left(I_{ij} - h_{ij}\right)^2 \geq 0$$

帽子矩阵中的对角线元素h_{ii}称为第i个观测值的杠杆值(Leverage Values), 它表示第i个观测值是否异常, 这是因为可以证明, h_{ii}表示的是 X_i与所有n个观测X的平均之间的距离度量, 这样**较大的杠杆值h_{ii}表明第i个观测值远离观测值的中心**.

注记. *若第i个观测值X_i是异常值, 即其具有较大的杠杆值h_{ii}, 则拟合值$\hat{Y}_{ii}$是一个实质性的杠杆, 这是因为*

1. $\hat{Y} = HY$, h_{ii}决定了Y_i的权重, h_{ii}越大,Y_i越重要, 而$h_{ii} = x_i'(X'X)^{-1}x_i$是$X$的函数, 因此$h_{ii}$度量的是$X$值决定$Y_i$在影响拟合值$\hat{Y}$中所起到的重要性程度的作用.

2. 由$\sigma^2(e_i) = \sigma^2(1 - h_{ii})$可知, h_{ii}越大, 残差e_i的方差越小, 在极端的情况下, $h_{ii} = 1, \sigma^2(e_i) = 0$, 这样拟合值等于观测值. 因此**不能用检测残差的方式发现异常值**.

一个杠杆值h_{ii}若大于平均杠杆的2倍, 就认为是大的, 平均杠杆值若用$\bar{h}$表示, 则

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

判别准则(经验法则): 当

$$h_{ii} > \frac{2p}{n}$$

认为X是异常的.

3.1 代码

```
> ###Identifying outlying X observations
> 2*p/n
[1] 0.3
> hlist  ##Case 3 and 15 larger than 2p/n
 [1] 0.20101253 0.05889478 0.37193301 0.11094009 0.24801034 0.12861620 0.15551745 0.09628780
 [9] 0.11463564 0.11024435 0.12033655 0.10926629 0.17838181 0.14800684 0.33321201 0.09527739
[17] 0.10559466 0.19679280 0.06695419 0.05008526
```

当然也可以自己编写

```
hat.plot <- function(fit) {
    # 找出参数个数
    p <- length(coefficients(fit))
    # 找出观测值个数
    n <- length(fitted(fit))
    # 给出h_ii
    plot(hatvalues(fit), main = "Index Plot of Hat Values", ylim = c(min(hatvalues(fit), 2*p/n),
     max(hatvalues(fit), 3*p/n)))
    # p/的n2或3倍
    abline(h=c(2,3)*p/n, col = "red", lty = 2)
    # 鼠标点击进行标注, 按推出ESC, 返回标注的点的编号
    identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
par(mfrow = c(1,1))
hat.plot(fit)
```

4 强影响点(Influential Observations)的识别(影响分析)

识别出关于X和Y的异常值后, 下一步就是弄清它们在拟合回归函数中是否具有可能得出严重歪曲结果的强影响.

在对一组数据拟合模型时, 我们希望保证拟合结果不要过度取决于一个或几个观测点.

第i个观察值对回归函数的拟合强影响的一种测度是: 根据n个观察值的估计回归系数向量b和根据$n-1$个观察值的估计回归系数向量$b_{(-i)}$之差,

$$b - b_{(-i)}$$

另一种可能的测度是

$$\hat{Y}_i - \hat{Y}_{(-i)}$$

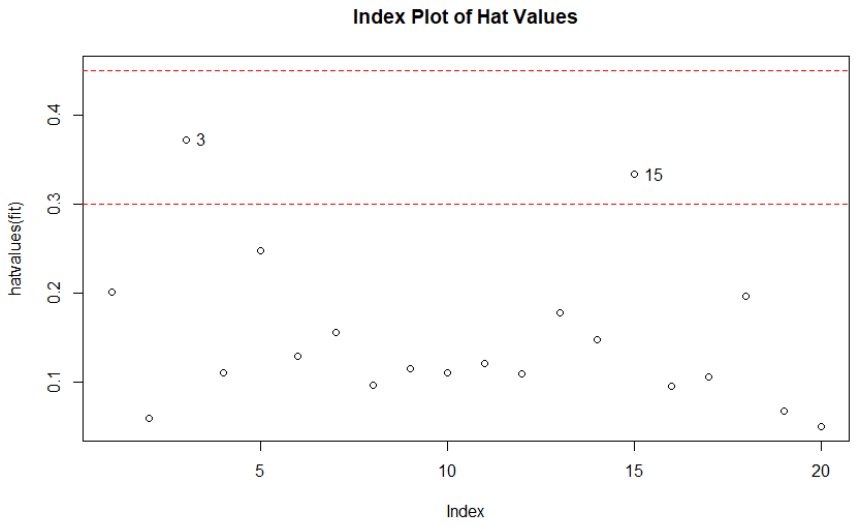

图 4: 画出的图像

4.1 影响度量: $DFFITS_i$

Belsley等建议识别强影响点的统计量为

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

其中 $t_i = \dfrac{d_i}{S\{d_i\}} = \dfrac{e_i}{\sqrt{MSE_{(i)}\,(1 - h_{ii})}} = \dfrac{e_i \sqrt{n - p - 1}}{\sqrt{SSE\,(1 - h_{ii}) - e_i^2}}.$

t_i是R-学生化残差,

4.2 Cook's D距离

第i个观测值对估计回归系数影响的总度量是Cook's D距离, 回顾回归模型的联合推断, 在$1 - \alpha$置信水平下, 所有p个回归参数β_k, $k = 0, 1, \cdots, p - 1$的联合置信区域的边界是

$$\frac{(b - \beta)' X' X (b - \beta)}{pMSE} \sim F_{1-\alpha}(p, n - p)$$

由于

$$\hat{\mathbf{Y}} = \mathbf{Xb}, \hat{\mathbf{Y}}_{(i)} = \mathbf{Xb}_{(i)} \Rightarrow \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)} = \mathbf{X}\left(\mathbf{b} - \mathbf{b}_{(i)}\right)$$

$$\Longrightarrow D_i = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)' \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}\right)}{pMSE} = \frac{\left(\mathbf{b} - \mathbf{b}_{(i)}\right)' \mathbf{X}'\mathbf{X} \left(\mathbf{b} - \mathbf{b}_{(i)}\right)}{pMSE}$$

Cook's D距离使用相同的形式度量剔除第i个观察值时的估计回归系数之差的联合影响:

$$
\begin{aligned}
D_i &= \frac{(b-b_{(-i)})' X'X (b-b_{(-i)})}{pMSE} \\
&= \frac{\left(\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_{(i)}\right)'\left(\hat{\mathbf{Y}}-\hat{\mathbf{Y}}_{(i)}\right)}{pMSE} \\
&= \frac{\sum_{j=1}^{n}\left(\hat{Y}_j-\hat{Y}_{j(i)}\right)^2}{pMSE} \\
&= \frac{\sum_{j=1}^{n}\left(\left(Y_j-\hat{Y}_j\right)-\left(Y_j-\hat{Y}_{j(i)}\right)\right)^2}{pMSE} \\
&= \frac{\sum_{j=1}^{n}\left(e_j-d_j\right)^2}{pMSE} \\
&= \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \\
&= \frac{h_{ii}}{p\left(1-h_{ii}\right)}\tilde{e}_i^2
\end{aligned}
$$

其中 $\tilde{e}_i = \dfrac{e_i}{\sqrt{MSE\,(1-h_{ii})}}$ 是学生化残差. 可以看出D_i与e_i, h_{ii}有关.

强影响点的识别可以用以下的方法判定,

$D_i > F_{0.50}(p, n-p)$　认为i为强影响点(用中位数)

4.3 影响度量: 估计回归系数的差值(DFBETAS)

DFB对应是按系数找强影响点,

$$
(\text{ DFBETAS })_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}, k = 0, 1, \cdots, p-1
$$

其中 c_{kk} 是 $C = (X'X)^{-1}$ 的第k个对角元, 且用 $MSE_{(i)}$ 估计 $\sigma^2\,(b_k)$.

4.4 代码实现

根据PPT上的代码

```
###Identifying influential cases
MSE = SSE/(n-p)
DFFITS = tlist * (hlist/(1-hlist))^0.5
Dlist = elist^2 /p/MSE*hlist/((1-hlist)^2)
clist = diag(solve( t(X)%*%X))
b = fit$coef;  DFBETAS = matrix(0,n,p)
for (i in 1:n){
    fiti = lm(Y[-i]~X1[-i]+X2[-i])
    bi = fiti$coef
    MSEi = sum(fiti$resi^2)/(n-1-p)   #$
```

```
    DFBETAS[i,] = (b-bi)/sqrt(MSEi*clist)  }
> cbind(Dlist,DFBETAS)
          Dlist
1  4.595055e-02 -3.051821e-01 -1.314856e-01  2.320319e-01
2  4.548118e-02  1.725732e-01  1.150251e-01 -1.426129e-01
3  4.901567e-01 -8.471013e-01 -1.182525e+00  1.066903e+00
4  7.216190e-02 -1.016120e-01 -2.935195e-01  1.960719e-01
5  1.883399e-09 -6.372122e-05 -3.052747e-05  5.023715e-05
6  1.136518e-03  3.967715e-02  4.008114e-02 -4.426759e-02
7  5.764939e-03 -7.752748e-02 -1.561293e-02  5.431634e-02
8  9.793853e-02  2.614312e-01  3.911262e-01 -3.324533e-01
9  5.313352e-02 -1.513521e-01 -2.946556e-01  2.469091e-01
10 4.395704e-02  2.377492e-01  2.446010e-01 -2.688086e-01
11 9.037986e-04 -9.020885e-03  1.705640e-02 -2.484518e-03
12 3.515436e-02 -1.304933e-01  2.245800e-02  6.999608e-02
13 2.121502e-01  1.194147e-01  5.924202e-01 -3.894913e-01
14 1.248925e-01  4.517437e-01  1.131722e-01 -2.977042e-01
15 1.257530e-02 -3.004276e-03 -1.247567e-01  6.876929e-02
16 2.474925e-03  9.308463e-03  4.311347e-02 -2.512499e-02
17 4.926142e-03  7.951208e-02  5.504357e-02 -7.609008e-02
18 9.636470e-03  1.320522e-01  7.532874e-02 -1.161003e-01
19 3.236006e-02 -1.296032e-01 -4.072030e-03  6.442931e-02
20 3.096787e-03  1.019045e-02  2.290797e-03 -3.314146e-03
```

利用car包中的influencePlot()函数还可以将离群点, 杠杆值, 强影响点整合到一张图中,

```
library(car)
par(mfrow=c(1,1))
influencePlot(fit, sub = "Circle size is proportional to Cook's distance", id=list(method="
    identify"))
```

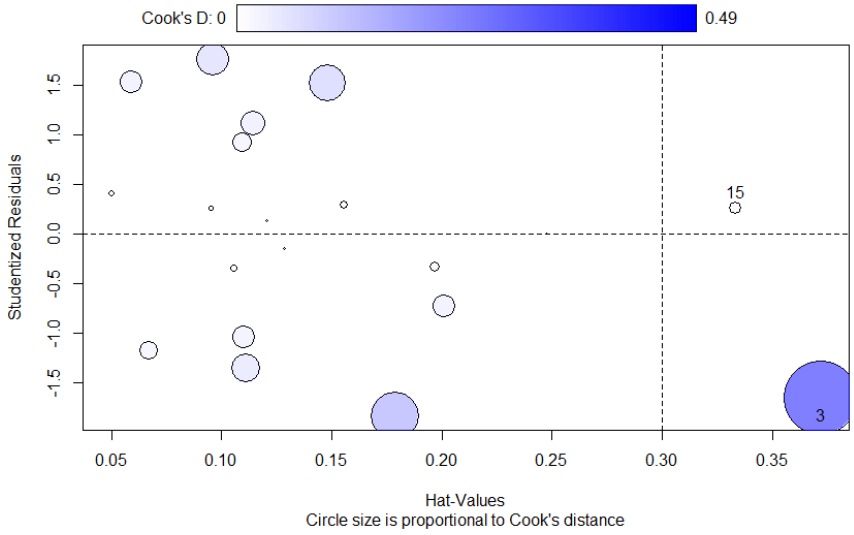

图 5: 画出的图像

图5被称为影响图, 纵坐标超过+2或小于−2的可以被认为是离群点, 水平轴超过0.2或0.3的有高杠杆值(通常为预测值的组合), 圆圈的大小与影响成比例, 圆圈很大的点可能是对模型参数的估计造成

的不成比例影响的强影响点.

5 多重共线性: 用方差膨胀因子(VIF)

回顾多重共线性:

- 当预测变量高度相关时, 回归系数的标准误差增大, 对偏相关系数而言,单独的回归系数不显著, 尽管可能总体模型是显著的.
- 导致行列式$(X^\top X)$接近0, 受较大的舍入误差和抽样方差的影响.

注意预测变量之间有较大的相关系数能说明可能存在多重共线性, 反之不一定成立, 这是因为多重共线性是指所有预测变量之间的线性.

5.1 多重共线性的识别

5.2 方差膨胀因子

方差膨胀因子(无量纲)估计的是: 估计回归系数的方差同自变量不线性相关时相比增加了多少.

定义 (方差膨胀因子)**.** 记R_j^2表示以X_j为预测变量, 其余的预测变量作为自变量的回归模型中的多重相关系数的平方.

$$E(X_j) = \beta_0 + \cdots + \beta_{j-1}x_{j-1} + \beta_{j+1}x_{j+1} + \cdots + \beta_{p-1}x_{p-1}$$

则X_j的方差膨胀因子定义为

$$VIF_j := \frac{1}{1 - R_j^2}$$

现在对模型进行标准化,

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$
$$X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_k}\right) \quad (k = 1, \ldots, p-1)$$

$$\underset{(p-1)\times(p-1)}{\mathbf{r}_{XX}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix}$$

$$\sigma^2\{b^*\} = (\sigma^*)^2\mathbf{r}_{XX}^{-1} \quad \sigma^2\{b_k^*\} = \sigma^2\{b^*\} = (\sigma^*)^2(VIF)_k$$

VIF_k是r_{XX}^{-1}的第k个对角元, 可以证明

$$(VIF)_k = \frac{1}{1 - R_k^2}$$

其中R_k^2是X_k与剩下的$p-2$个X拟合后的的决定系数, 对于两个预测变量的情况, 这时

$$R_k^2 = r_{12}^2$$

即X_1和X_2的单判定系数(即相关系数).

- 当$R_k^2 = 0$时, 即X_k与其他X变量不相关时, 方差膨胀因子$VIF_k = 1$.
- 当$R_k^2 \neq 0$时, $VIF_k > 1$, 表示b_k'的方差增大了, 这也可以由下式看出:

$$\sigma^2(B_k') = (\sigma')^2(VIF)_k = \frac{(\sigma')^2}{1-R_k^2}$$

- 当$R_k^2 = 1$时, 那么$VIF_k, \sigma^2(b_k')$就变成无穷大(∞).

由上可知VIF永远介于1到∞之间.

> 所有X中最大的$(VIF)_k$通常用作多重共线性严重程度的指标, 如果最大的VIF_k超过10, 即
>
> $$\max\left(VIF_1, \cdots, VIF_{p-1}\right) > 10$$
>
> 常常表示多重共线性可能过度的影响最小二乘估计值.

VIF_k的平均数也提供了关于多重共线性严重程度的信息, 它可以度量标准化回归系数偏离真实β_k'的程度, 可以证明, 这些误差平方和$(b_k' - \beta_k')^2$之和的期望值是,

$$E\left[\sum_{k=1}^{p-1}(b_k' - \beta_k')^2\right] = (\sigma')^2 \sum_{k=1}^{p-1} VIF_k$$

如果不存在X变量与模型中的其他变量线性相关, 那么$R_k^2 \equiv 0$, 因此$VIF_k \equiv 1$, 并且

$$E\left[\sum_{k=1}^{p-1}(b_k' - \beta_k')^2\right] = (\sigma')^2(p-1)$$

上述两式相除即得到VIF的平均数$\overline{VIF}$:

$$\overline{VIF} = \frac{\sum_{k=1}^{p-1} VIF_k}{p-1}$$

其中$p-1$为预测变量个数. 均值$\overline{VIF}$远大于1时, 认为存在严重的多重共线性.

5.3 条件数(数值分析)

度量多重共线性严重程度的另一个指标是方阵r_{XX}的条件数, 其定义为

$$k = \frac{\lambda_1}{\lambda_{p-1}}$$

也就是方阵r_{XX}的最大特征值与最小特征值之比. 一般的, 若$k < 100$, 则认为多重共线性较小; 若$100 \leq k \leq 1000$, 则认为有中等程度的多重共线性; 若$k > 1000$则认为存在严重的多重共线性.

变量填加理论

假设我们有模型

$$y = X_1\beta_1 + \epsilon$$

我们想要增加r个变量X_2到模型中. 则现在模型包含X_1, X_2可以被写为

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

若新的解释变量X_2与原来的解释变量是正交的, 即$X_1^\top X_2 = 0$且

$$X^\top X = \begin{pmatrix} X_1^\top X_1 & 0 \\ 0 & X_2^\top X_2 \end{pmatrix}$$

$$(X^\top X)^{-1} = \begin{pmatrix} (X_1^\top X_1)^{-1} & 0 \\ 0 & (X_2^\top X_2)^{-1} \end{pmatrix}$$

最小二乘估计量可以分别写为

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1} & 0 \\ 0 & \left(\mathbf{X}_2^T\mathbf{X}_2\right)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^T\mathbf{y} \\ \mathbf{X}_2^T\mathbf{y} \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{y} \\ \left(\mathbf{X}_2^T\mathbf{X}_2\right)^{-1}\mathbf{X}_2^T\mathbf{y} \end{pmatrix}$$

当新的预测变量与原来预测变量不正交, 即$X_1^\top X_2 \neq 0$, 则可以进行如下的分解.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \left(\mathbf{H}_1 + \mathbf{I} - \mathbf{H}_1\right)\mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \\ &= \mathbf{X}_1\left(\boldsymbol{\beta}_1 + \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{X}_2\boldsymbol{\beta}_2\right) + \left(\mathbf{I} - \mathbf{H}_1\right)\mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon \\ &= \mathbf{X}_1\boldsymbol{\theta} + \left(\mathbf{I} - \mathbf{H}_1\right)\mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon, \end{aligned}$$

其中

$$\mathbf{H}_1 = \mathbf{X}_1\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T$$

$$\boldsymbol{\theta} = \boldsymbol{\beta}_1 + \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{X}_2\boldsymbol{\beta}_2$$

这样矩阵X_1与$(I - H_1)X_2$变为正交的, 因此估计θ和β_2就可以分开进行估计, 并可以写成

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{y}$$

$$\hat{\boldsymbol{\beta}}_2 = \left[\mathbf{X}_2^T\left(\mathbf{I} - \mathbf{H}_1\right)\mathbf{X}_2\right]^{-1}\mathbf{X}_2^T\left(\mathbf{I} - \mathbf{H}_1\right)\mathbf{y}$$

从$\hat{\beta}_2$的表达式可以看出, 相当于是$(I - H_1)y$与$(I - H_1)X_2$进行建模, 进一步地, 由于

$$\hat{\beta}_1 = \hat{\theta} - \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{X}_2\hat{\beta}_2 = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}X_1^\top(y - X_2\hat{\beta}_2)$$

样本删除理论

使用的技巧是通过增加哑变量的方式删除某一样本. 令

$$u_{ij} = \begin{cases} 1, \text{if } i = j \\ 0, \text{if } i \neq j \end{cases}$$

新的模型被写为

$$Y_j = \beta_0 + \beta_1 X_{j1} + \cdots + \beta_{p-1} X_{j,p-1} + \gamma u_{ij} + \epsilon_j, \quad j = 1, \cdots, n$$

写成矩阵的形式

$$y = X\beta + u_i\gamma + \epsilon$$

因此拟合的均方误差(SSE)可以写成

$$SSE(\beta, \gamma) = \sum_{j \neq i}^{n} (y_j - x_j^\top \beta)^2 + (y_i - x_i^\top \beta - \gamma)^2$$

由于去除i项观测以后, 回归最小化了$\sum_{j \neq i}^{n} (y_j - x_j^\top \beta)^2$, 我们记去除$i$项观测后的OLS估计为$\hat{\beta}_{(-i)}$, 为使第二项$(y_i - x_i^\top \beta - \gamma)^2$为零, 则对参数$\gamma$的估计应为

$$\hat{\gamma} = y_i - x_i^\top \hat{\beta}_{(-i)}$$

而这正是去除残差$e_{(-i)}$, 我们使用变量添加理论有

$$\hat{\gamma} = \left[\mathbf{u}_i^T (\mathbf{I} - \mathbf{H}) \mathbf{u}_i\right]^{-1} \mathbf{u}_i^T (\mathbf{I} - \mathbf{H}) \mathbf{y} = \frac{e_i}{1 - h_{ii}} = e_{(i)}.$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \left(\mathbf{y} - \mathbf{u}_i \frac{e_i}{1 - h_{ii}}\right) = \hat{\boldsymbol{\beta}} - \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_i \frac{e_i}{1 - h_{ii}}$$

当$\gamma = 0$时, 模型被简化为

$$Y_j = \beta_0 + \beta_1 X_{j1} + \ldots + \beta_{p-1} X_{j,p-1} + \varepsilon_j, j = 1, 2, \cdots, n$$

因此

$$SSE(u_i|X) = SSE(R) - SSE(F) = SSE - SSE_{(-i)}$$

偏决定系数为

$$R^2_{u_i|X} = \frac{u_i^\top (I - H) Y}{sqrt u_i^\top (I - H) u_i \cdot Y^\top (I - H) Y} = \frac{u_i^\top e}{\sqrt{(1 - h_{ii}) SSE}} = \frac{e_i}{sqrt(1 - h_{ii}) SSE}$$

是两残差$(I - H)u_i$, $(I - H)Y$的Pearson相关系数.

$$R_{u_i|X} = \frac{SSR(u_i|X)}{SSE(X)} = \frac{SSE - SSE_{(-i)}}{SSE}$$

因此

$$SSE_{(-i)} = SSE - SSR(u_i|X) = SSE - R^2_{u_i|X} SSE = SSE - \frac{e_i^2}{1 - h_{ii}}$$