稳健简单线性回归

蒋冰

November 26, 2013

1 问题介绍

大多数传统的统计方法是在对研究的总体的分布形式和其他特征做出一些假设的情况下的，而实际总体不满足假设条件，或观察数据中包含不能代表总体特性的异常点（Outliers）时，传统的统计方法可能就会出现较大误差。由于仪器故障、操作失误等原因，异常点是很容易出现的。因此，偏离假设是常见的现象。通常，数据量比较少和自变量个数不多时，可由散点图或残差图等找出异常点，但当样本量增大或自变量个数增多时，检测出异常点就会变得很困难。而且，即使找出了异常点，若异常点不是由于记录、录入等人为误差造成的，剔除异常点就没有充分理由。所以需要一种估计方法能既不剔除由随机误差造成的客观存在的异常点，又不会对回归系数影响太大的回归方法，即具有稳健性。

估计的稳健性 (Robustness) 指的是在估计过程中产生的估计量对模型误差的不敏感性。因此稳健估计是在比较宽的资料范围内产生的优良估计。如在独立同分布正态误差的线性模型中，最小二乘估计 (Ordinary Least Square , OLS) 是有效无偏估计。然而当误差是非正态分布时，OLS 不一定是最有效的。但误差分布事先不一定知道，故有必要考虑稳健回归的问题。稳健回归 (Robust Regression) 估计，如误差为正态时，它比 OLS 稍差一点，但误差非正态时，它比 OLS 要好得多。这种对误差项分布的稳健特性，常能有效排除异常值干扰。

2 方法

2.1 基本思想

稳健估计讨论问题的方式是: 对于实际问题有一个假定模型，同时又认为这个模型并不准确，而只是实际问题理论模型的一个近似。它要求解决这类问题的估计方法应达到以下目标：

1. 假定的观测分布模型下，估值应是最优的或接近最优的。
2. 当假设的分布模型与实际的理论分布模型有较小差异时，估值受到粗差的影响较小。
3. 当假设的分布模型与实际的理论分布模型有较大偏离时，估值不至于受到破坏性影响。

稳健估计的基本思想是：在粗差不可避免的情况下，选择适当的估计方法，使参数的估值尽可能避免粗差的影响，得到正常模式下的最佳估值。稳健估计的原则是要充分利用观测数据（或样本）中的有效信息，限制利用可用信息，排除有害信息。由于事先不大准确知道观测数据中有效信息和有害信息所占比例以及它们具体包含在哪些观测中，从抗差的主要目标着眼是要冒损失一些效率的风险，去获得较可靠的、具有实际意义的、较有效的估值。

2.2 M 估计

在最小二乘法估计中，回归系数 $\hat{\beta}$ 是使 $\sum_{i=1}^{n} e_i^2$ 达到最小，其中 $e_i = y_i - \sum_{j=1}^{m} \beta_i x_{ij}$ ，

若令 $\rho(x) = x^2$, 则 $\hat{\beta}$ 使 $\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} e_i^2$ 达到最小。显然函数 $\rho(e_i) = e_i^2$ 随 |x| 的增大而增大，为使上式达到最小，就必须照顾某些点，特别是一些异常点。因此最小二乘法估计会往往使得那些远离数据群体的数据 (很可能是异常值) 对残差平方和影响比其他数据大得多。这是因为最小二乘估计为了达到极小化残差平方和的目的，必须迁就远端的数据，所以异常值对于参数估计相当敏感。

M 估计可以看作是对最小二乘估计的一种改善。它用一个函数 $\rho(x)$ 来代替每个观测值与估计值的残差平方和 x^2，此函数除了随 |x| 的增大速度比 x^2 慢外，其他性质与 x^2 类似。并且通过对 $\sum_{i=1}^{n} \rho(e_i)$ 取极小而得到的。对于函数 ρ 和样本，M 估计是使得目标函数 $\sum_{i=1}^{n} \rho(e_i)$ 达到最小的估计。这便是 M 估计的想法。

$\rho(x)$ 是 x 的函数，它随 x 值的增大而增大，通常取 $\rho(x) = -lnf(x)$，其中 $f(x)$ 是 x 的分布密度函数。$\rho(x)$ 对取极小就相当于对 x 的似然函数的对数 $\sum_{i=1}^{n} lnf(x)$ 取极大。故 M 统计值就成为最大似然估计。若 $\rho(x) = x^2$，M 估计量等价于最小二乘估计量。

为了消除测量单位对估计的影响，类似于传统的将数据标准化的做法，在 M 估计中用 $z_i = \frac{x_i}{s}$ 来代替 $_i$。这里 s 是待估尺度（参数）(scale)。尺度 s 的稳健估计是另一个重要的稳健统计学问题。因此此文中将不考虑标准化。

令 ρ 关于 β 的导数为：$\psi(z_j) = \frac{\partial \rho(e_i)}{\partial \beta_j}$，则对原式子求导有，$\sum_{i=1}^{n} \psi(x_i) x_{ij} = 0, j = 1, 2, \cdots, m$，解此联立方程就可以解出 β 的值。

因为此方程较为复杂，要直接借出来比较困难，所以可以采用迭代法来求解。

令 $w_i = \frac{\psi(e_i)}{e_i}$, 则原式可以转换为 $\sum_{i=1}^{n} \psi(x_i) x_{ij} = \sum_{i=1}^{n} e_i w_i x_{ij} = \sum_{i=1}^{n} (y_i - \sum_{i=1}^{n} \beta_i x_{ij}) w_i x_{ij} = 0$，

所以有 $\sum_{i=1}^{n} w_i x_{ij} y_i = \sum_{i=1}^{n} \beta_i x_{ij} w_i x_{ij}$，可以看出这是一个加权的最小二乘法。

为了能与最小二乘法做比较，因而在 M 估计中常将最小二乘估计作为初值，即取 $\beta^{(0)}=\hat{\beta}$。

设 $\beta^{(l)}$ 是第 l 次迭代值，其迭代公式为：$\beta^{(l+1)}=\beta^{(l)}+\Delta\beta^{(l)}, l=0,1,2,\cdots$，这里$\Delta\beta^{(l)}$ 是增量。为求 $\Delta\beta^{(l)}$ 好处假设 $\psi(x)$ 在 $\beta^{(l)}$ 处可以做泰勒展开，并取下面的近似表达式：$\psi(e_i)\approx\psi(e_i^l)-\sum_{j=1}^{m}\psi'(e_i^l)x_{ij}\Delta\beta_j^{(l)}, i=1,2,3,\cdots,n$ 其中，$e_i^{(l)}=y_i-\sum_{j=1}^{m}\beta_i^{(l)}x_{ij}$, 将展开式代入方程组，可得:

$\sum_{i=1}^{n}\psi(e_i)x_{ij}=\sum_{i=1}^{n^{i}}(\psi(e_i^l)-\sum_{j=1}^{m}\psi'(e_i^l)x_{ij}\Delta\beta_j^{(l)})x_{ij}, j=1,2,\cdots,m$，上述方程租是关于 $\Delta\beta_1^{(l)},\Delta\beta_2^{(l)},\cdots,\Delta\beta_m^{(l)}$ 的线性方程租。因而可以解出$\Delta\beta^{(l)}$。为了让迭代不会无限循环，需要事前设定好收敛原则：事先确定一个非常小的 $\varepsilon>0$，要求 $\underset{j}{max}|\Delta\beta_j^{(l)}|<\varepsilon$，便停止迭代。

下面列出几种常用的 ρ 函数、ψ 函数：

1. Huber 函数

$$\rho(x)=\begin{cases}\frac{1}{2}x^2 & |x|\le k\\ k|x|-\frac{1}{2}x^2 & |x|>k\end{cases}$$

$$\psi(x)=\begin{cases}x & |x|\le k\\ 0 & |x|>k\end{cases}$$

$$w(x)=\begin{cases}1 & |x|\le k\\ \frac{k}{|x|} & |x|>k\end{cases}$$

2. Hampel 三段截尾函数

$$\rho(x)=\begin{cases}\frac{1}{2}x^2 & |x|\le a\\ k|x|-\frac{1}{2}x^2 & a<|x|\le b\\ ab-\frac{1}{2}a^2+(c-b)\frac{a}{2}\left[1-\left(\frac{c-|x|}{c-b}\right)^2\right] & b<|x|\le c\\ ab-\frac{1}{2}a^2+(c-b)\frac{a}{2} & |x|>c\end{cases}$$

$$\psi(x)=\begin{cases}x & |x|\le a\\ a\bullet sign(x) & a<|x|\le b\\ a\bullet\frac{c-|x|}{c-b}\bullet sign(x) & b<|x|\le c\\ 0 & |x|>c\end{cases}$$

$$w(x)=\begin{cases}1 & |x|\le a\\ \frac{a}{|x|} & a<|x|\le b\\ a\bullet\frac{c-|x|}{c-b}\bullet sign(x) & b<|x|\le c\\ 0 & |x|>c\end{cases}$$

3. Tukey 的双二次函数

$$\rho(x)=\begin{cases}\frac{1}{6}\left[1-\left(1-x^2\right)^3\right] & |x|\le 1\\ \frac{1}{6} & |x|>1\end{cases}$$
$$\psi(x)=\begin{cases}x\left(1-x^2\right)^2 & |x|\le 1\\ 0 & |x|>1\end{cases}$$
$$w(x)=\begin{cases}\left(1-x^2\right)^2 & |x|\le 1\\ 0 & |x|>1\end{cases}$$

3 数值模拟

为了比较稳健估计和最小二乘法，本文随机生成了一些数据，然后分别用两种估计方法进行回归分析。

本文首先随机选了 n =10 个 x 值，然后用 $y=3.5x+6+\varepsilon$ 来生成 10 个对应的 y 值。再将最后的 2 个 y 值改成符合函数 $y=3.5x+12+\varepsilon$ 的 2 个异常点。

```
n <- 10
m <- 2
r <- rnorm(n, 0, 1)
x <- runif(n, -10, 50)
beta1 <- 3.5
beta0 <- 6
betax <- 12
y <- beta1 * x + beta0 + r   #  y=b0+b1*x
y[(n - m + 1):n] <- betax * x[(n - m + 1):n] + betax + r[(n - m + 1):n]
```

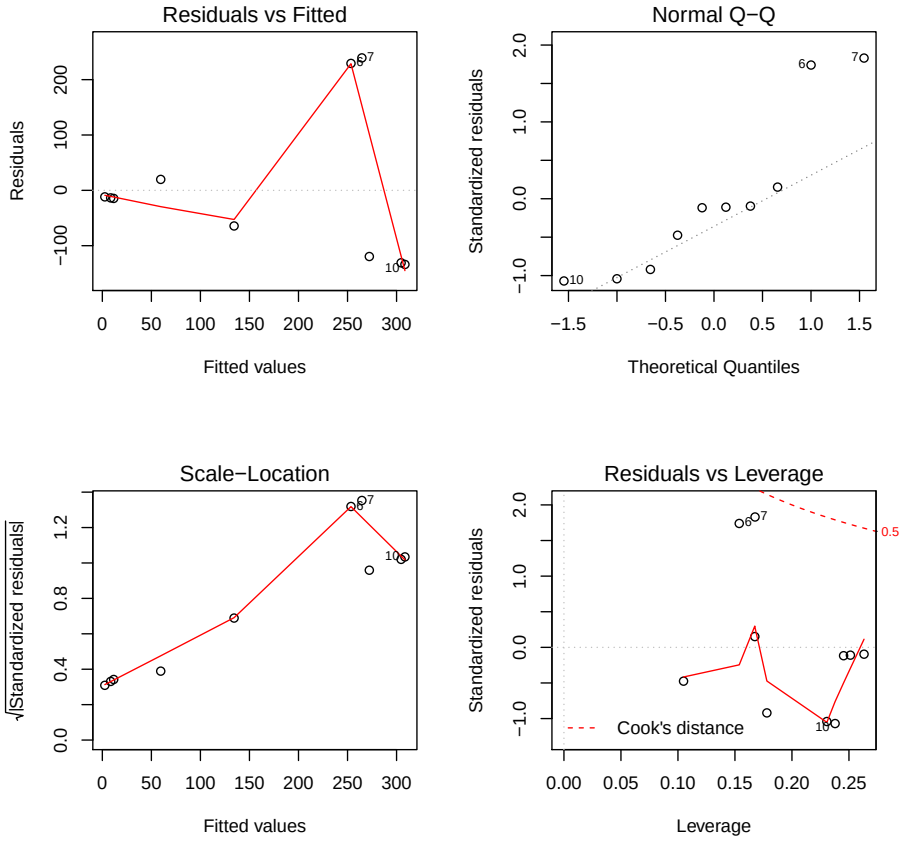
为重复实验方便，我们保存了随机生成的 x 和 y，下表是按升序排序后的 x 和 y，其中第 6 点和第 7 点是异常点。

Order	x	y
1	-4.282575	-9.211033
2	-3.251418	-5.083544
3	-2.722361	-3.00885
4	5.576017	79.347572
5	18.518049	69.986128
6*	**39.13665**	**482.892625**
7*	**41.05978**	**503.99225**
8	42.375578	152.674542
9	47.983068	173.75062
10	48.655117	174.669358

然后进行回归分析和诊断，其中 model_1 即为最小二乘法回归。

```
x <- c(-4.282575, -3.251418, -2.722361, 5.576017, 18.518049, 39.13665, 41.05978,
    42.375578, 47.983068, 48.655117)
y <- c(-9.211033, -5.083544, -3.00885, 79.347572, 69.986128, 482.892625, 503.99225,
```

```
    152.674542, 173.75062, 174.669358)
model_1 <- lm(y ~ x)
par(mfrow = c(2, 2))
plot(model_1)
```



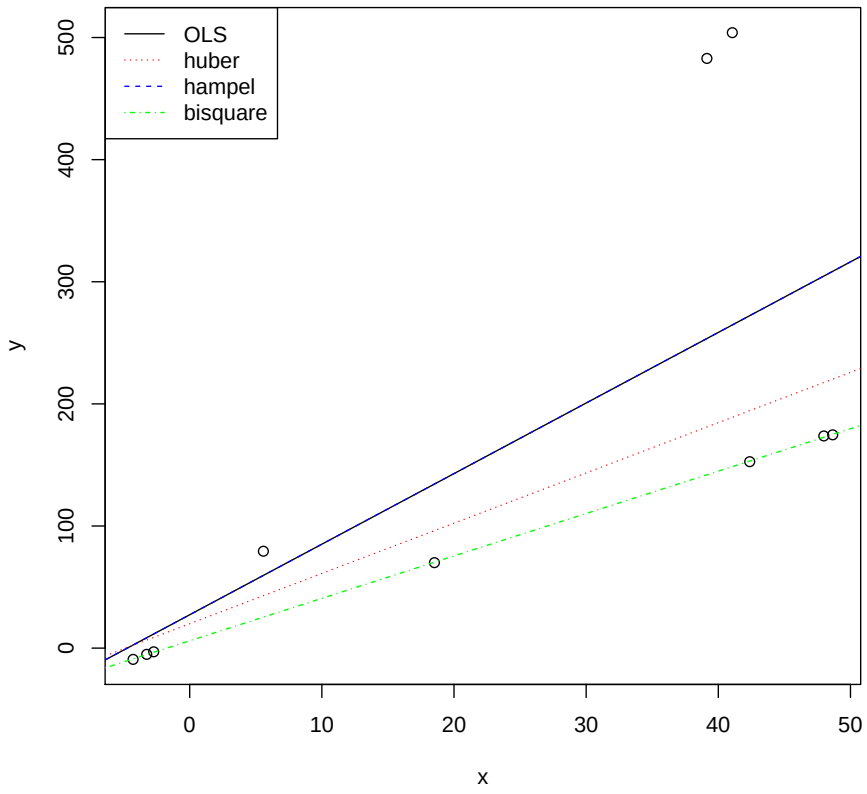
左上的图为，Residual vs fittedji 为拟合值 $\hat{y}$ 对残差的图形, 可以看出, 除了 6 号点和 7 号点外，数据点都基本均匀地分布在直线 $y = 0$ 的两侧，无明显趋势; 从右上角的 Normal QQ-plot 图，除了 6 号点和 7 号点外，可以看到数据点分布趋于一条直线, 说明如果没有一场点，残差是服从正态分布的; 左下方的图 Scale — Location 图显示了标准化残差 (standardized residuals) 的平方根的分布情况. 号点和 7 号点为残差最大值点; 右下方的图显示了 Cook 距离 (Cook's distance)，它表示各点对回归的影响点。通过作图和比较，我们可以基本确定，第 6 点和第 7 点是异常点。

因此，我们用 M 估计对原数据进行回归。其中 model_2 为采用 huber 函数的稳健回归 M 估计，model_3 为采用 Hampel 三段截尾函数的稳健回归 M 估计，model_4 为采用 Tukey 的双二次函数的稳健回归 M 估计。

```
library("MASS")
model_2 <- rlm(y ~ x, psi = psi.huber)
model_3 <- rlm(y ~ x, psi = psi.hampel)
model_4 <- rlm(y ~ x, psi = psi.bisquare)
```

下图即为数据的散点图和回归直线。

```
plot(y ~ x)
abline(model_1, col = "black", lty = "solid")
abline(model_2, col = "red", lty = "dotted")
abline(model_3, col = "blue", lty = "dashed")
abline(model_4, col = "green", lty = "dotdash")
legend("topleft", c("OLS", "huber", "hampel", "bisquare"), lty = c("solid",
  "dotted", "dashed", "dotdash"), col = c("black", "red", "blue", "green"))
```



从上图可以看出，由于异常点的出现，最小二乘法显然为照顾异常点，与预期有所偏离。而稳健回归却很好地削弱了异常点的影响，但选择不同的 ψ

函数，对异常点的削弱情况也不同。在本例中，采用 Hampel 函数的 M 估计，几乎没有削弱异常点的影响，其回归线与最小二乘法的回归线很靠近；而采用 bisquare 函数的 M 估计几乎完全消除了异常点的影响，其与去掉异常点的最小二乘法差不多；而采用 huber 函数的 M 估计则兼顾了所有点的影响。它们回归得拟合直线系数汇总如下图所示（其中 RSE 为 Residual standard error）：

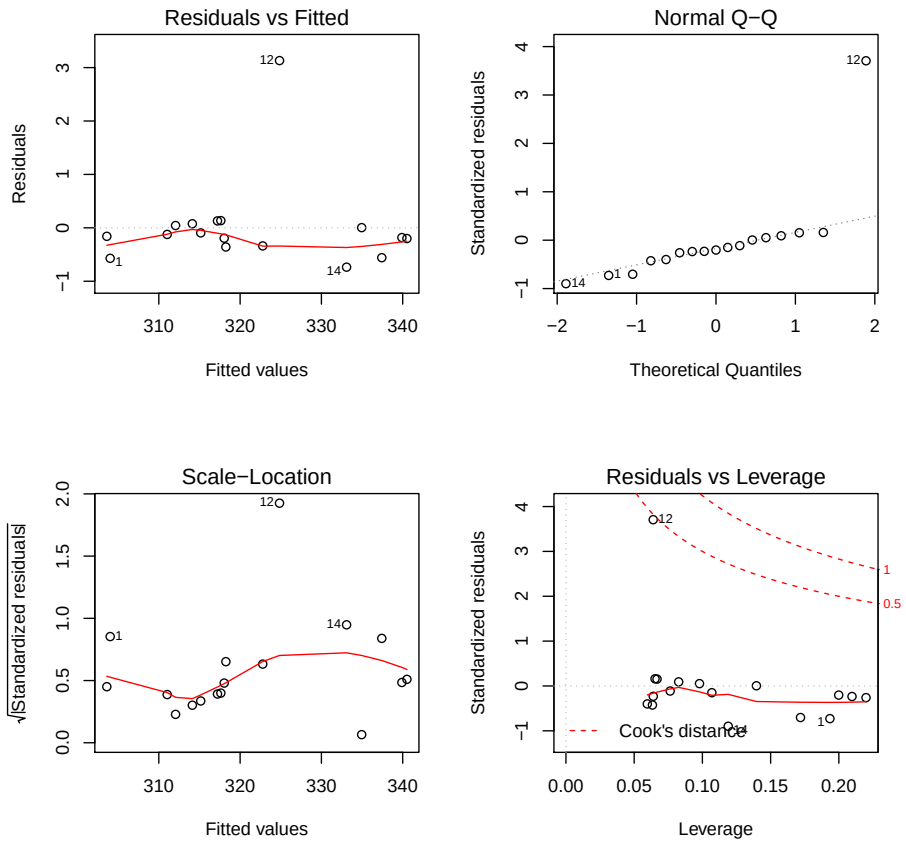
	$\hat{\beta}_0$	$se(\hat{\beta}_0)$	t value	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	t value	RSE	R^2
OLS	27.303	66.685	0.409	5.780	2.099	2.754	0.409	0.4866
huber	19.9676	30.9381	0.6454	4.1171	0.9738	4.2278	57.96	
hampel	27.3032	66.6853	0.4094	5.7798	2.0990	2.7536	136.3	
bisquare	6.0185	0.3788	15.8877	3.4728	0.0119	291.2480	0.7013	

4 实际数据

本文在实际数据中比较稳健估计和最小二乘法，在此选择的数据是 Forbes 数据。

在 19 世纪四五十年代，苏格兰物理学家 James D.Forbes 试图通过水的沸点来估计海拔高度。他在阿尔卑斯山及苏格兰手机数据。R 语言的 MASS 包选取了他 1857 年的论文中的 17 个数据，本文选用的就是这个数据，然后用最小二乘法来分析原是数据，其中 model_1 即为最小二乘法回归。

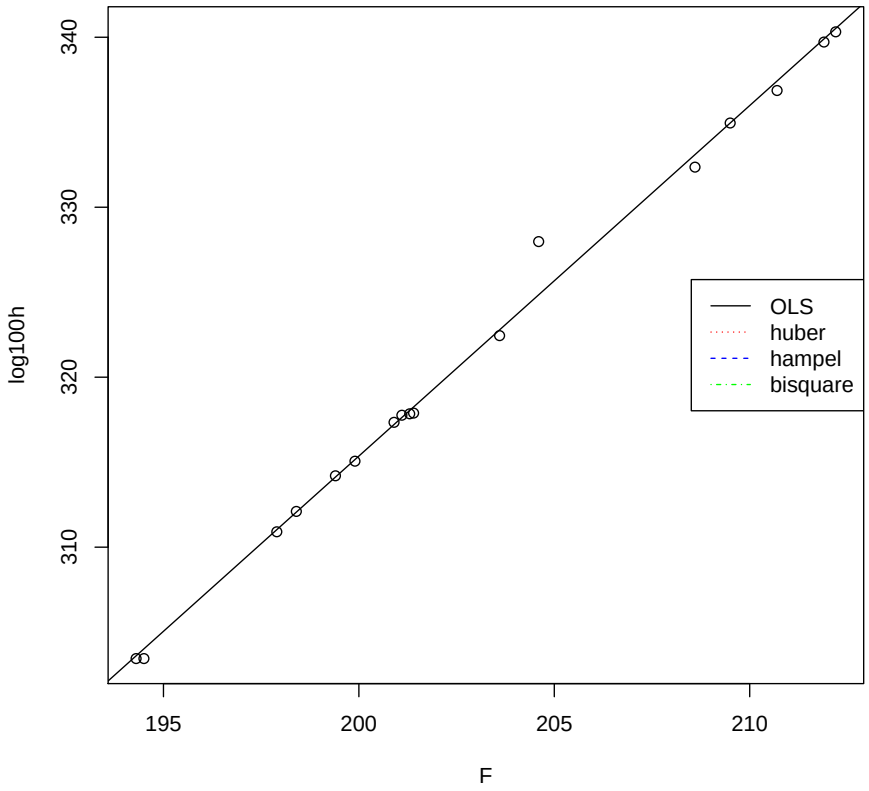
```
library(MASS)
F <- forbes$bp
h <- forbes$pres
myforbes <- data.frame(F = F, h = h, logh = log(h), log100h = 100 * log(h))
model_1 <- lm(formula = log100h ~ F, data = myforbes)
par(mfrow = c(2, 2))
plot(model_1)
```



从图上可以较明显地，原数据的第 12 点应为异常点。

因此，我们用 M 估计对原数据进行回归。其中 model_2 为采用 huber 函数的稳健回归，model_3 为采用 Hampel 三段截尾函数的稳健回归，model_2 为采用 Tukey 的双二次函数的稳健回归。

```
plot(log100h ~ F, data = myforbes)
abline(model_1, col = "black")
abline(model_2, col = "red", lty = "dotted")
abline(model_3, col = "blue", lty = "dashed")
abline(model_4, col = "green", lty = "dotdash")
legend("right", , c("OLS", "huber", "hampel", "bisquare"), lty = c("solid",
  "dotted", "dashed", "dotdash"), col = c("black", "red", "blue", "green"))
```



从上图可以看出，由于异常点的出现，最小二乘法显然为照顾异常点，与预期有所偏离。而稳健回归却很好地削弱了异常点的影响，但选择不同的 ψ 函数，对异常点的削弱情况也不同。在本例中，采用 Hampel 函数的我估计，几乎没有削弱异常点的影响，其回归线与最小二乘法的回归线很靠近；而采用 bisquare 函数的 M 估计几乎完全消除了异常点的影响，其与去掉异常点的最小二乘法差不多；而采用 huber 函数的 M 估计则兼顾了所有点的影响。它们回归得拟合直线系数汇总如下图所示（其中 RSE 为 Residual standard error）：

	$\hat{\beta}_0$	$se(\hat{\beta}_0)$	t value	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	t value	RSE	R^2
OLS	-97.08662	7.69377	-12.62	2.06224	0.03789	54.42	0.409	
huber	-95.4058	2.7701	-34.4418	2.0532	0.0136	150.4864	0.2999	
hampel	-95.1766	2.3846	-39.9138	2.0519	0.0117	174.7026	0.2858	
bisquare	-95.1816	2.5214	-37.7490	2.0519	0.0124	165.2250	0.2812	

上图可以看出，虽然有异常点的存在，但由于异常点的影响较小，采用 M 估计得到回归系数与最小二乘法得到回归系数差不多。但采用 M 估计得到的回归系数的显著性却比用最小二乘法得到回归系数有显著提升。

5 总计

从上面的分析我们可以看出，由于异常点的存在，最小二乘法在处理这些数据时，会受到不同程度的影响。最小二乘法要剔除这些异常点，需要进行比较回归诊断，而且在实际情况中，不一定允许剔除数据。而进行稳健回归时，我们不需检测或剔除异常点，便可得到比较理想的结果。而且，如果异常点较多或影响较大，最小二乘法的结果很能令人满意，而稳健回归却能得到合适的结果。而且，即使异常点的影响很小，使用稳健回归能提高回归系数的显著性。在实际数据中，根据情况采用适当形式的稳健回归是很有必要的。

References

[1] Huber, P. J. (2011). Robust statistics (pp. 1248-1251). Springer Berlin Heidelberg.

[2] Fox, J. (2002). An R and S-Plus companion to applied regression. Sage.

[3] 王谷, & 过秀成. (2011). 包含异常数据的居民出行稳健回归分析. 武汉理工大学学报 (交通科学与工程版), 3, 019.

[4] 孙士兵, 赵欢, & 贺宗梅. (2008). 基于稳健回归技术的软件成本估计方法. 科学技术与工程, 8(17), 4864-4868.

[5] 郭亚帆, & 杜金柱. (2010). 经典回归与稳健回归方法的应用比较研究. 市场研究, (009), 17-21.