基于theta ipod的新的超参数选择方法

李　昊[1]　张玉环[2]
1.Hankuk University of Foreign Studies　韩国首尔　100071
2.北京市第十中学　北京　100039

摘　要：线性回归模型是应用广泛的一类模型估计回归系数的方法是最小二乘法，然而最小二乘法很容易受到异常点的影响不稳健的。真实数据中都会存在异常点很难避免，这时候利用最小二乘法进行估计估计结果会比较不理想，在一定程度上限制了线形模型在实际科研中的应用。面对有异常值的数据的处理方法，大部分的处理方法为利用cook’s distance找出异常值后，删除再进行回归，这种方法在一些情况下是不合理的，因为在实际科学领域中异常值也是包含有一定的信息，不能完全删除，特别是在医学药学等领域。在此方法中我们可以利用不同的loss function，来决定留下多少有关于异常值的信息，theta ipod（Outlier Detection Using Nonconvex Penalized Regression（Yiyuan She & Art B. Owen，2011）是一个基于最小二乘法下的一个探测异常值并且估计出具有鲁棒性的参数。他的基本思想是在最小二乘法估计的基础上将一个新的转换参数，并且为这个转换因子添加一个L1惩罚项函数，当这个点的转换参数变为0则认为这个点不是异常值。并且为惩罚项函数会依赖于一个超参数lambda，最后使用最小bic得分的lambda来建造模型，但是在论文中出现的bic方法是非常不严谨的，因为bic在 lambda端点值时会存在极端最小值。所以我提出了一种新的利用bic来选择lambda的方法，我们考虑了方差对bic值的影响，先讲总体方差利用样本估计出来。之后添加到新的模型里，可以得到更加准确的超参数，进而得到很准确的参数估计值并且准确判断出异常值和异常值对模型影响的大小。在最后一节数据模拟章节给出了具体例子方便参考。

关键词：theta ipod；BIC；异常值

1　theta–ipod with Huber’s loss算法介绍

theta ipod（Outlier Detection Using Nonconvex Penalized Regression（Yiyuan She & Art B. Owen，2011）是一个探测异常值并且估计出具有鲁棒性的参数的方法，他的模型是在原来的一般回归模型中引入一个新的参数γ，用于判断数据的距离当前离散程度并且利用阈值函数修正，为了防止过度拟合我们还为新的参数γ添加了一个惩罚项λ，使用迭代算法就可以减小异常值影响并且估计出一个具有鲁棒性的回归参数，并且有效的防止过度拟合，并且式1.1为凸函数可以利用迭代算法就可以很容易找出使得式1.1最小的参数估计。

数学模型为用1.1来估计1.2中的β。其中1.2的ρ()为Huber’s损失函数

$$g(\underline{\beta},\underline{\gamma},\sigma)=\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i-\underline{x_i}^{\top}\underline{\beta}-\lambda_i)+\frac{1}{2}cn\sigma+\lambda\sum_{i=1}^{n}|\gamma_i| \quad (1.1)$$

$$l(\underline{\beta},\sigma)=\sum_{i=1}^{n}\rho(\frac{y_i-\underline{x_i}^{\top}\beta}{\sigma};\lambda)+\frac{1}{2}cn\sigma \quad (1.2)$$

并且可进一步证明在$\Psi(t:\lambda)=t-\Theta(t:\lambda)$条件下两个模型的所估计的β参数是一致的

证明如下

$$\Psi(t:\lambda)=\rho'(t:\lambda)$$

$$X^{\top}X\hat{\underline{\beta}}=X^{\top}(\underline{y}-\hat{\underline{\gamma}})$$

$$X^{\top}X\hat{\underline{\beta}}=X^{\top}(\underline{y}-\Theta(\hat{\underline{r}}:\lambda\hat{\sigma}))$$

$$X^{\top}(\underline{y}-\Theta(\hat{\underline{r}}:\lambda\hat{\sigma}))=X^{\top}(\underline{y}-\hat{\underline{r}}+\psi(\hat{\underline{r}}:\lambda\hat{\sigma}))$$

$$X^{\top}(\underline{y}-\hat{\underline{r}})+X^{\top}(\psi(\hat{\underline{r}}:\lambda\hat{\sigma}))$$

$$\hat{\underline{r}}=\underline{y}-X\hat{\underline{\beta}}$$

$$X^{\top}X\hat{\underline{\beta}}+X^{\top}(\psi(\hat{\underline{r}}:\lambda\hat{\sigma}))=X^{\top}X\hat{\underline{\beta}}$$

$$X^{\top}(\Psi(\hat{\underline{r}}:\lambda\hat{\sigma}))=\underline{0}$$

$$\Psi(\hat{\underline{r}}:\lambda)=\underline{0}$$

Theta– ipod迭代算法

将有异常值的数据进行最小二乘估计所求出的β_{ols}作为迭代起始点

$\gamma^{(i)}=y-X^{T}(\hat{\beta}_{ols})$

直到γ不再变化循环下列

$y_{adj}=y-\gamma^{(j)}$

$\beta^{(j)}=(X^{T}X)^{-1}X^{T}y_{adj}$

$r^{(j)}=y-X^{T}\beta^{(j)}$

$\gamma^{(j+1)}=\theta(r^{(j)},\ \lambda)$

最后返回γ和β的值

2　theta–ipod with Huber’s loss算法新的超参数选择方法

在Outlier Detection Using Nonconvex Penalized

Regression（Yiyuan She & Art B. Owen，2011）一文中提到的λ选择方法是创建一个从0开始到y的最大值的向量之后分别计算

BIC（λ）= mlog(RSS/m)−k(log(m)+1)　　（2.1）

选择最小的BIC所对应的λ，利用此λ求出β

2.1式中 $RSS=\left\|\left(I-X\left(X^TX\right)^{-1}X^T\right)\left(y-\hat{\gamma}\right)\right\|$

2.1式中m=样本数−参数个数

2.1式中k=当前λ下的γ不为0的个数

利用bic作为参数选择方法是很明智的他可以在大样本的情况下很快计算出结果

但　是Outlier Detection Using Nonconvex Penalized Regression（Yiyuan She & Art B. Owen，2011）论文中的bic计算方法有一个致命缺点，在他的论文中选择超参数λ的部分中也提到：“曲线有时在λ范围的末端附近具有狭窄的局部最小值。为了抵消这种影响，我们将平滑样条拟合到数据点集（DF（λ），BIC＊（λ））并选择具有最大邻域的局部最小值。可以使用平滑样条的局部最大值来确定邻域大小。当然，可能还有其他合理的方法来解决这个问题”。经过推导实验，发现他并没有包括原来样本的方差信息，所以这导致了他在λ在靠近0时有极端最小值（图2.1）。我们将总体的方差的预测值设置为是最小一乘法中的样本残差减去样本残差中位数的绝对值的中位数除以0.6745（0.6745是一个校正因子，当误差服从正太分布时可以保证方差估计的一致性）。

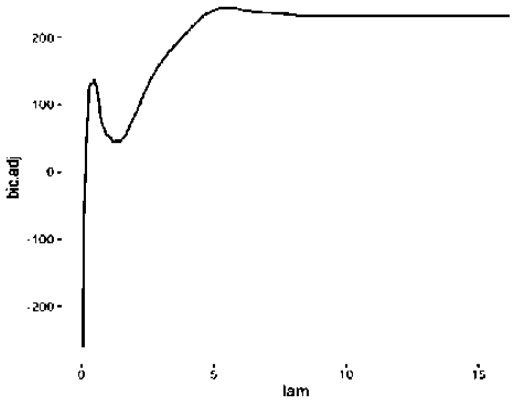

图2.1

我们提出一种新的bic计算方法，计算公式如下：

NEWBIC(λ)=nlog(σ²)+RSS/(σ²)+k(log(m)+1)　（2.2）

2.2式中σ是最小一乘法中的样本残差减去样本残差中位数的绝对值的中位数除以0.6745

2.2式中n为样本个数

2.2式中 $RSS=\left\|\left(I-X\left(X^TX\right)^{-1}X^T\right)\left(y-\hat{\gamma}\right)\right\|$

2.2式中m=样本数 − 参数个数

2.2式中k=当前λ下的γ不为0的个数

利用此方法就可以解决Outlier Detection Using Nonconvex Penalized Regression（Yiyuan She & Art B. Owen，2011）论文中计算bic的漏洞，在bic中添加了数据中的y的离散信息，非常有效的解决了λ范围的末端附近具有狭窄的局部最小值，利用此newbic计算出的bic和λ的折线图（图2.2），可以看出bic符合典型机器学习参数的选择曲线。

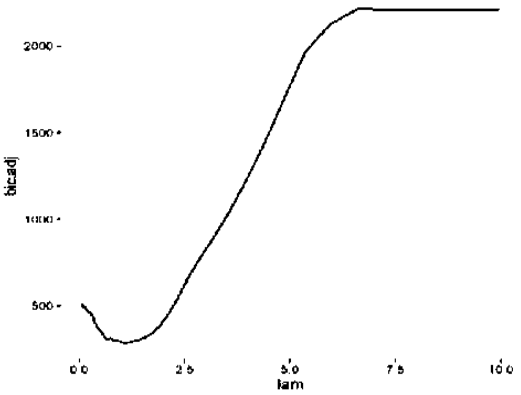

图2.2

3　数据模拟

生成100个服从平均分布的自变量x将β的真值设为（0，2），利用$X^T\beta$计算出y值并将其中百分之10的点加上6作为异常值。运行theta−ipod算法并利用newbic筛选出最合适的λ，画出每一个样本根据不同λ的值的γ路径（图3.1）和根据不同λ的值β值的变化路径，其中的虚线为最优λ时β和每个样本的γ值。

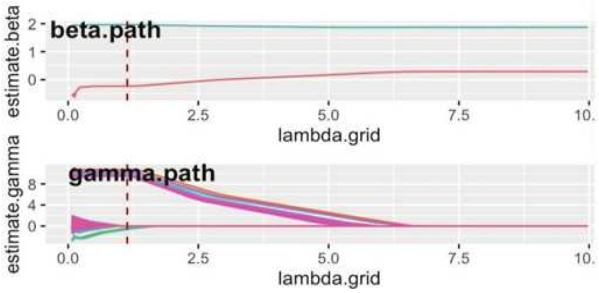

图3.1

从图3.1可以看出原论文的所筛选出最小的bic时的最小值可以发现lambda无限接近于0，这会导致将所有样本点变为异常值，所估计的参数也会变得不准确。在r语言中利用lm求出的OLS估计值为（1.412，2.096）可以看出OLS估计收到了很大的影响，尤其是$\beta_{\{0\}}$。通

过上述算法我们可以得出当newbic最小时所对应的λ为1.134418，在当前λ下β的预测值（蓝色线条）为（–0.2320448，1.9688299），可以发现与真值（粉色虚线）的（0，2）所差无几并且将异常值准确的判断了出来如图（3.2）所示。

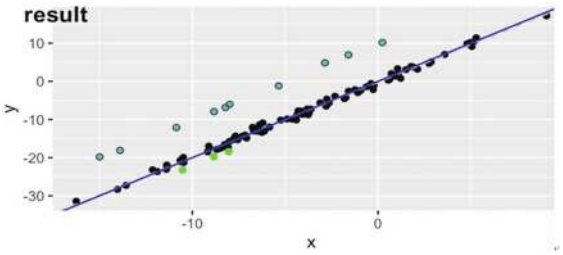

图3.2

上图中的空心原点为异常值，绿色点为λ值为1.134418时所探测出的异常值，虽然有一些点被污染但是还是将异常值充分的找了出来。

当然我们也可以利用交叉验证的方法来选择λ值，经过实验验证，交叉验证的方法所需时间非常久，而且抽样方差也会非常大（当验证数据中有太多的异常值时和没有异常值时），所以mse和λ的曲线会非常的混乱而且偏差非常严重，所以并不推荐使用。当然应该也会有更加准确的λ选择方法来避免正常点被污染。

参考文献：

[1] [She and Owen，2011] She，Y.，Owen，A. B.（2011）. Outlier detection using nonconvex penalized

[2] regression. Journal of the American Statistical Association，106，626–639.