연구논문집 제10권 제 2호 p.123~124 1979

Robust Regression Estimates

Chung, In-Ha
Dept. of General Education

〈**Abstract**〉

Robust regression estimates are studied and some new aspects of robustness are proposed.

Robust 회귀추정

정 인 하
교 양 과 정 부

〈요 약〉

Robust 회귀추정에 대하여 연구하였으며 새로운 Robustness의 관점이 제안되었다.

I. Introduction

Consider the classical least squares problem:

$$y_i=\sum_{j=1}^{p} x_{ij}\beta_j+e_i, \quad i=1, \cdots, n \qquad (1.1)$$

where x_{ij} are known constants and e_i are independently and identically distributed random errors. We are going to estimate p unknown parameters $\beta_1, \cdots, \beta_p$ on the basis of n observations $y_1, \cdots, y_n$.

Least squares Estimate (*LSE*) $\hat{\beta}$ can be obtained by minimizing

$$\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p} x_{ij}\beta_j)^2 \qquad (1.2)$$

or solving the system of p equations

$$\sum_{i=1}^{n}(y_i-\sum_{k=1}^{p} x_{ik}\beta_k)x_{ij}, \quad j=1, \cdots, p \qquad (1.3)$$

It is a well-known fact that *LSE* is optimum when errors are independently and identically distributed normal.

When the variances of e_i are not homogeneous and they have long-tailed distribution, efficiency of *LSE* reduces. Even a single wild observation may make the *LSE* undesirable. This is the justification for studying the robust estimates as the alternatives to *LSE*.

II. Robust Regression Estimates

There are three categories of robust estimates:

(a) R-estimates

Estimates derived from rank tests

(b) L-estimates

Linear combinations of order statistics

(c) M-estimates

Maximum-Likelihood-type estimates

Jaeckel establishes a relationship between the members of the above three categories.

He found that each member of one category corresponds to a unique member of each of the other two categories.

This enables him to conclude that properties proved for some members of these categories also hold for the corresponding members of the other class.

Replace (1.1) by minimizing

$$\sum_{i=1}^{n}\rho(y_i-\sum_{j=1}^{p}x_{ij}\beta_j) \qquad (2.1)$$

and (1.2) by the system of p equations

$$\sum_{i=1}^{n}\Psi(y_i-\sum_{k=1}^{p}x_{ik}\beta_k)x_{ij}=0 \qquad (2.2)$$

$j=1, \cdots, p$, where ρ is a convex function and $\phi=\rho'$.

Huber(1973) proposed

$$\begin{aligned}\rho(x)&=\frac{1}{2}x^2 && if|x|<k \\ & kx-\frac{1}{2}k^2 && if|x|\geqq k \qquad (2.3)\end{aligned}$$

Denote residuals by Δ_i, then

$$\Delta_i=y_i-\Sigma_j x_{ij}\beta_j \qquad (2.4)$$

Replace (2.1) by minimizing

$$\sum_{i=1}^{n}a_n(R_i)\Delta_i \qquad (2.5)$$

where $R_i=\text{Rank}(\Delta_i)$, $i=1, \cdots, n$ and $a_n(\cdot)$ is a monotone score function satisfying

$$\sum_i a_n(i)=0 \qquad (2.6)$$

An estimate obtained from (2.1) or (2.2) is an M-estimate.

The estimate obtained from (2.5) is an R-estimate.

Ⅲ. Restrictions on Independent Variables

Let $\hat{\beta}=(\hat{\beta}_1, \cdots, \hat{\beta}_p)$ be the solution of the normal equation

$$X^T X\beta=X^T y \qquad (3.1)$$

Then for some i, $\hat{y}_i=\Sigma_j x_{ij}\hat{\beta}_j$ may be so close to y_i, even a gross error in y_i does not appear in the residual $y_i-\hat{y}_i$.

Let $\Gamma=X(X^TX)^{-1}X^T=(\gamma_{il})$ be the projection matrix. Then Γ is symmetric and $\Gamma^T\Gamma=\Gamma$.

For the *LSE* $\hat{\beta}$, $\hat{y}_i=\Sigma_l\gamma_{il}y_l$ and the *i*th residual is

$$y_i-\hat{y}_i=(1-\gamma_{ii})y_i-\sum_{l\neq i}\gamma_{il}y_l$$

It is easy to prove that

$$\Sigma_l\gamma_{il}^2=\gamma_{ii}, \text{ and } 0\leqq\gamma_{ii}\leqq1,\ ave(\gamma_{ii})=\frac{1}{n}\Sigma\gamma_{ii} =p/n.$$

To make sure that we can spot outlying observations, max γ_{ii} should be considerably larger than 1.

A large value of γ_{ii} corresponds to an outlying values $(x_{i1}, \cdots, x_{ip})$ of the independent variables.

Thus always calculate the diagonal element of the projection matrix $\Gamma=X(X^TX)^{-1}X^T$. If a particular γ_{ii} is close to 1, then decrease it by replication of the observations.

Ⅳ. Conclusions

We have thought the robust estimation in the three categories, *M, L, R*-estimate. But we can obtain the robust estimate by putting restrictions on the independent variables. This may not belong to the above three categories. Another aspect of robust regression estimate is the so called model robustness (Huber). This problem arises when we are not quite sure whether a straight line is a proper curve to fit. This problem needs further study.

References

1. Huber, P.J. (1973), "Robust regression; Asymptstics, Conjectures and Monte Carlo." The Annals of Statistics V.1 : 799—821
2. Huber, P.J. (1975), "Robustness and Designs." A survey of Statistical Designs and Linear models, North Holland Publishing company.
3. Andrews, D.F. (1975) : "Alternative Calculations for Regression and Analysis of Variance Problems." Applied Statistics, North Holland Publishing Company.