Robust Regression

유제진

Department of Statistics, Sungkyunkwan University

CONTENTS

Introduction

Outlier

- Given distribution is non-normal(skewed), when Normal distribution : Mean = Median = Mode
- Distributional(parametric) assumptions are violated in case of classical estimation.
- e.g. Normality of error, LLN, CLT etc.
- Existence of Outliers : Typical Problem.

Introduction

Outlier

- Outlier leads to bad performance of Statistical Procedure.
- Distributional outlier : classic statistical procedure is sensitive to "long-tailedness" of distribution
- Fatal Problem : Masking problem of modest outlier by larger outlier.

Introduction

Robust

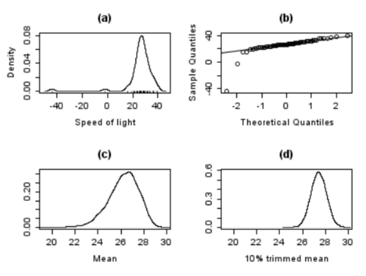
- Solution to 'Outlier Problem'
  $\Rightarrow$ insensitive to outliers and designed to not unduly affected by violation of distributional(or parametric) assumptions.

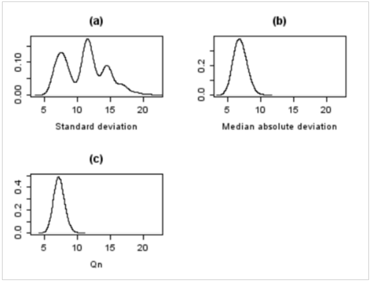
Typical robust measure

- Instead of Mean,
  $\Rightarrow$ Median or Trimmed Mean
  For estimation of location; central tendency.
- Instead of Standard Deviation,
  $\Rightarrow$ MAD(Mean Absolute Deviation) & IQR(Inter Quantile Range)
  For estimation of scale; statistical dispersion.

Introduction

Robust





Breakdown Point

Guideline for robustified approach

Breakdown Point and Robustness

- Finite Sample Breakdown Point :
  The fraction of data that can be given arbitrary values without making the estimator arbitrarily bad.
- Asymptotic Breakdown Point :
  The limit of the finite sample breakdown point as n goes to infinite.

Breakdown Point

Guideline for robustified approach

- Ex) Mean : Breakdown Point of 0

$$\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$$

  $\overline{X_n}$ can be arbitrarily large just by changing any of $X_1, \cdots, X_n$.

- In the same context, Median have Breakdown Point of 50%.
- Breakdown Point can't exceed 50%.
- The higher the Breakdown Point of an estimator, the more Robust it is.

M-ESTIMATION

M(Maximum likelihood type)-estimation

- Proposed by Huber(1964).
- General, dominant and widely used method in robust statistics
- Advantage
  (1) Generality
  (2) High Breakdown Point
  (3) Efficiency
- Disadvantage will be discussed later.
- Mean, median, trimmed mean, MLE(maximum likelihood estimator), LSE(least squares estimation) etc. are all special cases of M-estimators.

M-ESTIMATION

What is MLE?

$$\hat{\theta}_{MLE} = \underset{\theta}{argmax} \prod_{i=1}^{n} f(x_i, \theta)$$
$$= \underset{\theta}{argmin} \sum_{i=1}^{n} \{-log f(x_i, \theta)\}$$

- It can be generalized to minimization of $\sum_{i=1}^{n} \rho(x_i)$, where ρ is loss function.
- That is, M-estimator

$$\underset{\theta}{argmin} \sum_{i=1}^{n} \rho(x_i, \theta)$$

M-ESTIMATION

- M-estimator = $\underset{\theta}{argmin} \sum_{i=1}^{n} \rho(x_i, \theta)$
- Properties of function ρ
    1. $\rho(x) \geq 0$: Non-negativity
    2. $\rho(0) = 0$
    3. $\rho(x) = \rho(-x)$: Symmetry
    4. $\rho(x_i) \leq \rho(x_i')$ for $|x_i| \leq |x_i'|$: Monotonicity in $|x_i|$
- If ρ is differentiable, $\psi = \rho'$ is called the influence function.
- Minimizing $\sum_{i=1}^{n} \rho(x_i, \theta)$ can often be done by solving $\sum_{i=1}^{n} \psi(x_i, \theta) = 0$

Typical Examples of M-estimation

1. Mean

$$\rho(x_i, \theta) = (x_i - \theta)^2$$

$\Rightarrow$ M-estimator of θ =

$$\underset{\theta}{argmin} \sum_{i=1}^{n} (x_i - \theta)^2 = \bar{x} \text{ : Sample Mean}$$

2. Median

$$\rho(x_i, \theta) = |x_i - \theta|$$

$\Rightarrow$ M-estimator of θ =

$$\underset{\theta}{argmin} \sum_{i=1}^{n} |x_i - \theta| \text{ : Sample Median}$$

Objective Functions

- Objective Function = $\underset{\theta}{argmax} \sum_{i=1}^{n} \rho(x_i, \theta)$
- Another Definition of M-estimator =
  An extremum estimator whose objective function is the form of sample average ' $\frac{1}{n} \sum(\cdot)$ '

Objective Functions

Three typical objective functions

Method	Objective Function	Weight Functions
Least Squares	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 \\ k\lvert e\rvert - \frac{1}{2}k^2 \end{cases}$	$w_H(e) = \begin{cases} 1 \\ \frac{k}{\lvert e\rvert} \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\{1-[1-(\frac{e}{k})^2]^3\} \\ \frac{k^2}{6} \end{cases}$	$w_B(e) = \begin{cases} [1-(\frac{e}{k})^2]^2 \\ 0 \end{cases}$

where $w(e) = \psi(e)/e$.

- For Huber Bisquare, Range for $\rho.(e), w.(e)$ is $\begin{cases} \text{for } |e| \leq k \\ \text{for } |e| > k \end{cases}$

Objective Functions

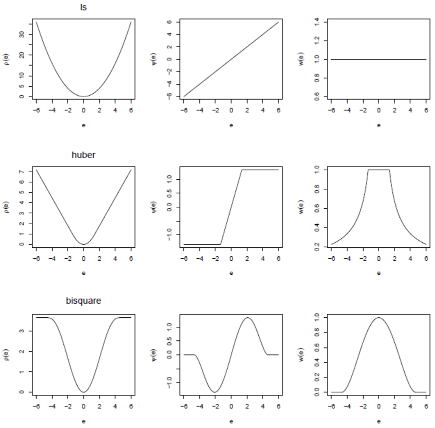


Figure 1: Objective, ψ, and weight functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are $k = 1.345$ for the Huber estimator and $k = 4.685$ for the bisquare. (One way to think about this scaling is that the standard deviation of the errors, σ, is taken as 1.)

- Usually, $k = 1.345\sigma$ for the Huber, $k = 4.685\sigma$ for the bisquare produce 95% efficiency when the errors are normal, and still offer protection against outliers.
- Common approach to take $\hat{\sigma}$

  $$\hat{\sigma} = MAR/0.6745$$

  where MAR is the "Median Absolute Residual".

  $$MAR = med|e_i - med(e)|$$

Robust Regression

- Emerged to overcome the limitation of traditional(=classical) regression analysis.
- Classical Regression methods such as Least square in Linear regression, need various assumptions on parameters or distributions.
- OLS and Robustness.

Robust Regression

When Robust Regression is needed?

1. Outliers are included in the model.
2. There exists "Heteroscedasticity" in the model.
   Typically, when the variance depends on explanatory variables, etc.

- Simplest method for Robust Regression :
  LAD(Least Absolute Deviation) Regression = L_1 regression
- M-estimation for Robust Regression :
  Robust to outlier of Y, but not robust to outlier of X.
  It has no advantages over LSE when outliers of X exist.

Robust Regression

Robust Regression with M-estimation

- Classical Linear Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i$$

- Fitted Model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x_i}^T \hat{\boldsymbol{\beta}}$$

- Residuals

$$e_i = y_i - \hat{y}_i$$

Robust Regression

Robust Regression with M-estimation

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{argmin} \sum_{i=1}^{n} \rho(e_i) \\ &= \underset{\beta}{argmin} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_{\boldsymbol{i}}^T \hat{\boldsymbol{\beta}})\end{aligned}$$

- If $\rho(e_i) = e_i^2$, Least Square Method using L_2-loss.

Robust Regression

Robust Regression with M-estimation

- If ρ is differentiable, $\psi = \rho'$.
- Then estimating equations can be written as

$$\sum_{i=1}^{n} \psi(y_i - \boldsymbol{x_i}^T \hat{\beta}) x_i^T = 0$$

- Solving equations is equivalent to a WLS(Weighted Least Squares) problem.
- IRLS(Iteratively Reweighted Least Squares) need to be used.

Quantile Regression

Quantile Regression(QR)

- It aims at estimating either a conditional 'median' or other quantiles of the response variable Y.
- Extension of linear regression. We use it when the conditions of linear regression are not applicable.
- Median function is more explainable than the mean regression function for 'asymmetric' conditional distribution.

Quantile Regression

Quantile

- r-th Quantile of random variable Y with Cdf F_y

$$Q_Y(\tau) = F_y^{-1}(\tau) = Inf\{y : F_Y(y) \geq \tau\}, \tau \in (0,1)$$

- Here, we define the loss function as

$$\rho_\tau(y) = y * (\tau - I(y < 0)) \text{ where } I : \text{Indicator Function}$$

Also known as "Check Function".

Quantile Regression

Quantile with Check Function

- Specific quantile can be found by minimizing the expected loss of $Y - \mu$ with respect to μ

$$\min_{\mu} E(\rho_\tau(Y-\mu)) = \min_{\mu}\{(\tau-1)\int_{-\infty}^{\mu}(y-\mu)dF_Y(y) + \int_{\mu}^{\infty}(y-\mu)dF_Y(y)\}$$

Quantile Regression

Quantile with Check Function

- Take the derivative of $E(\rho_\tau(Y - \mu))$ and set to 0. Then, let q_τ be the solution of the equation.

$$0 = (1-\tau)\int_{-\infty}^{q_\tau} dF_Y(y) - \tau \int_{q_\tau}^{\infty} dF_Y(y)$$

- This equation reduces to

$$0 = F_Y(q_\tau) - \tau \Rightarrow F_Y(q_\tau) = \tau$$

- Hence, q_τ is τ-th quantile of random variable Y.

QUANTILE REGRESSION

Conditional Mean / Quantile function

- Given Check Function for the

$$\rho_\tau(y) = y * (\tau - I(y < 0)) \text{ where } I : \text{Indicator Function}$$

1. Conditional mean of Y given $X = x$

$$E(Y|X = x) = \underset{\alpha}{argmin} E[(Y - \alpha)^2 | X = x]$$

2. Conditional quantile of Y given $X = x$

$$q_\tau(x) = Q_{Y|X}(\tau | X = x) = \underset{\alpha}{argmin} E[\rho_\tau(Y - \alpha) | X = x]$$

Quantile Regression

Two Important Applications of quantile regression

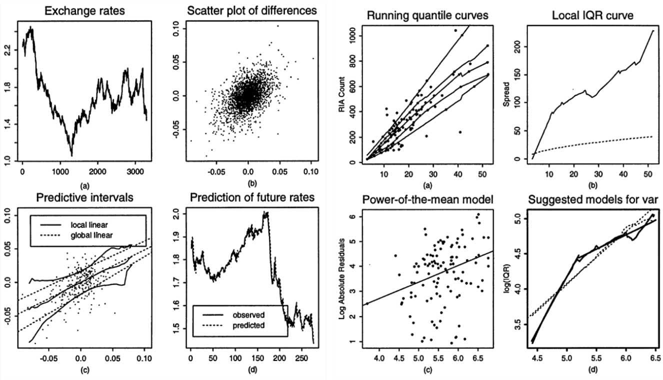
1. Constructing Prediction Intervals

$$\left[\, q_{\frac{\alpha}{2}}(x),\ q_{1-\frac{\alpha}{2}}(x) \right]$$

2. Detecting heteroscedasticity.

Quantile Regression

Plots of 'Predictive Intervals' and 'Heteroscedasticity' cases



QUANTILE REGRESSION

L_1 Regression

1. Also known as 'media regression'.
2. Special case of quantile regression.
3. Simplest method for robust regression.
4. $\tilde{\beta}$ is estimated by solving minimization problem.

$$\tilde{\beta} = argmin \frac{1}{n} \sum_{i=1}^{n} |y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}|$$

$$= argmin \frac{1}{n} \sum_{i=1}^{n} \rho_{0.5}(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}) \Rightarrow \text{Median of } \boldsymbol{x_i}^T \boldsymbol{\beta}$$

Quantile Regression

General form of parametric QR

- Suppose the τ-th conditional quantile function is

$$Q_{Y|X}(\tau) = \boldsymbol{X}\boldsymbol{\beta}_\tau$$

- Given the distribution function of Y,
  $\boldsymbol{\beta}_\tau$ can be obstained by solving

$$\boldsymbol{\beta}_{\boldsymbol{\tau}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{argmin} E(\rho_\tau(Y - \boldsymbol{X}\boldsymbol{\beta}))$$

Quantile Regression

General form of parametric QR

- Solving the sample analog gives the estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{argmin} \sum_{i=1}^{n} (\rho_\tau(Y_i - \boldsymbol{X}_i \boldsymbol{\beta}))$$

- $X_i \hat{\beta}_\tau$ is the estimate for quantifle function $Q_{Y|X}(\tau)$.

Quantile Regression

- The minimization problem can be reformulated as a linear programming problem.

$$\beta_\tau = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{argmin}(\rho_\tau(Y - X\beta))$$

$$\Rightarrow \min_{\beta^+,\beta^-,\mu^+\mu^-}\{\tau 1_n^T\mu^+ + (1-\tau)1_n^T\mu^- | X(\beta^+ - \beta^-)_+ u^+ - u^- = Y\}$$

$$\text{where } \beta_j^+ = max(\beta_j, 0), \mu_j^+ = max(\mu_j, 0)$$
$$\beta_j^- = -min(\beta_j, 0), \mu_j^- = -min(\mu_j, 0)$$

Quantile Regression

- For $\tau \in (0,1)$, under some regularity conditions, $\hat{\beta}_\tau$ is asymptotically normal.

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N(0, \tau(1-\tau)D^{-1}\Omega_x D^{-1})$$

- where $D = E(f_Y(X\beta)XX^T)$ and $\Omega_x = E(X^T X)$

Quantile Regression

Nonparametric QR

- Given $(X_1, Y_1), \cdots (X_n, Y_n)$,

$$\hat{\beta}(x_0) = \underset{\beta}{argmin} \sum_{i=1}^{n} \rho_\tau \{Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j\} * K_h(X_i - x_0)$$

- The local linear estimate of the quantile function $q_\tau(x) = Q_{Y|X}(\tau|X = x)$ is given by $\hat{q}_\tau(x) = \hat{\beta}_0$, as we learned before.