# STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

*yinfeng@cuhk.edu.cn*

April 14, 2020

# Overview

1. Outlier

2. Heavy-Tailed Distributions

3. Robust Statistics

4. Robust Regression

# Scope

We first consider robust regression for multiple linear regression model when the random error terms mainly follow Gaussian i.i.d. with zero mean, but there also exist a small amount of outliers that make the overall distribution appear heavy-tailed.

Robust regression can also be applied for nonlinear models.

# Definition of Outlier

An outlier differs considerably from the majority of the data, more precisely,

- differs considerably in terms of input, $\mathbf{x}$;
- differs considerably in terms of output, $y$;
- differs considerably in terms of both input $\mathbf{x}$ and output $y$.

Outlier is connected to influential point. Recall previous lecture.

# Outlier Might Be Informative

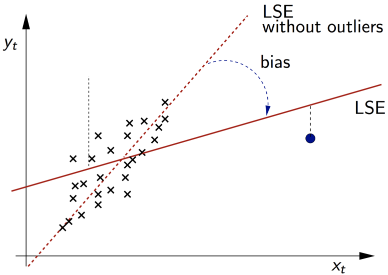
It would, however, be misleading to always think of outliers as “bad” data. They may well contain unexpected relevant information.

## A True Story (according to [Kandel, 91])

*The discovery of the ozone hole was announced in 1985 by a British team working on the ground with conventional instruments and examining its observations in detail. Only later, after reexamining the data transmitted by the TOMS instrument on NASAs Nimbus 7 satellite, was it found that the hole had been forming for several years. Why had nobody noticed it? The reason was simple: the systems processing the TOMS data, designed in accordance with predictions derived from models, which in turn were established on the basis of what was thought to be reasonable, had rejected the very (excessively) low values observed above the Antarctic during the Southern spring. As far as the program was concerned, there must have been an operating defect in the instrument.*

# Adverse Effect of Outlier

1. Data collected in a broad range of applications frequently contain one or more outliers; that is, observations that deviate from the general pattern of the data.
2. Classical estimates, such as the sample mean, sample variance and ordinary least-squares fit of a regression model, can be very adversely influenced by outlier(s) and often fail to provide good fits to the majority of the data.

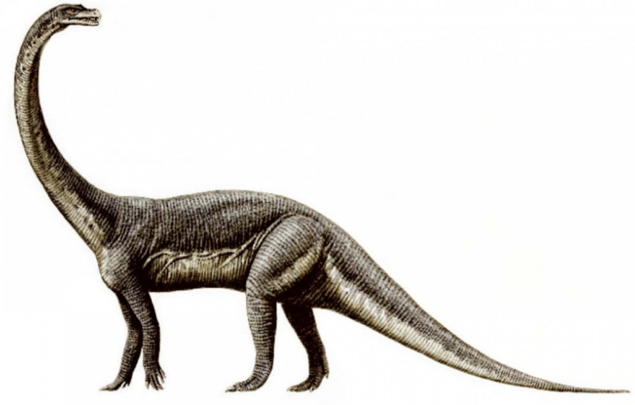


# How to Handle Outlier

1. Delete outliers based on outlier diagnostic. Diagnostic of outliers, using for instance Cook's measure, can be time consuming for multiple linear regression (with large number of inputs $k$ and large data samples $n$).
2. Delete outliers based on residual analysis. For instance, use the plot of scaled residuals versus fitted values and regressors to identify outliers.
3. Retain outliers and use robust statistics and robust regression!

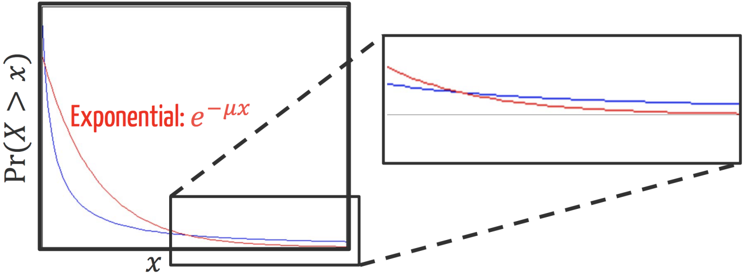
# Heavy-Tailed Distribution

I am heavy-tailed!



# Heavy-Tailed Distribution: Definition

Formal definition: A distribution with a “tail” that is “heavier” than an exponential (decay slower than an exponential).



For right heavy-tailed distribution: $\lim_{x\to\infty} \frac{Pr\{X>x\}}{e^{-\mu x}} = \infty$ for all $\mu > 0$ and $\mu \neq \infty$.

But in practice, we use Gaussian distribution (the one selected to represent the majority) as the reference.

# Heavy-Tailed Distribution: Example-I

Contaminated Gaussian distribution:

$$p(\varepsilon_i) = (1-\alpha)\mathcal{N}(\varepsilon_i; 0, \sigma^2) + \alpha\mathcal{H}, \quad i = 1, 2, ..., n \tag{1}$$

where $\alpha$ is called contamination coefficient and $\mathcal{H}$ can be any heavy-tailed distribution, such as $t$-distribution, exponential distribution, etc.

Among others, the most famous contaminated Gaussian is:

$$p(\varepsilon_i) = (1-\alpha)\mathcal{N}(\varepsilon_i; 0, \sigma^2) + \alpha\mathcal{N}(\varepsilon_i; 0, \sigma_c^2), \quad i = 1, 2, ..., n \tag{2}$$

where $\sigma_c^2$ is a large value, it is common to assume $\sigma_c^2 > 10\sigma^2$.

# Heavy-Tailed Distribution: Example-I

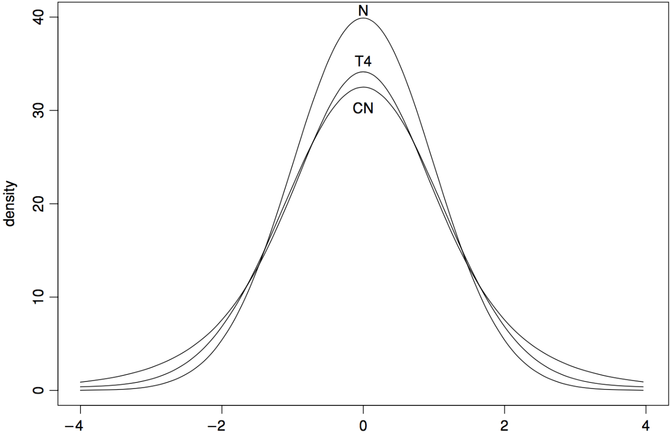
How to generate i.i.d. samples from a contaminated Gaussian, namely, $p(\varepsilon_i) = (1-\alpha)\mathcal{N}(\varepsilon_i; 0, \sigma^2) + \alpha\mathcal{N}(\varepsilon_i; 0, \sigma_c^2)$, with $\alpha > 0$ ?

Follow the steps:

- Generate a number $v_i$ from the uniform distribution $\mathcal{U}[0,1]$.
- If $v_i$ falls in the range $[0,\alpha)$, then generate $\varepsilon_i$ from $\mathcal{N}(\varepsilon_i; 0, \sigma_c^2)$; otherwise generate $\varepsilon_i$ from $\mathcal{N}(\varepsilon_i; 0, \sigma^2)$.
- Repeat step-1 and step-2 independently for $n$ times and obtain $\{\varepsilon_i\}_{i=1}^n$.

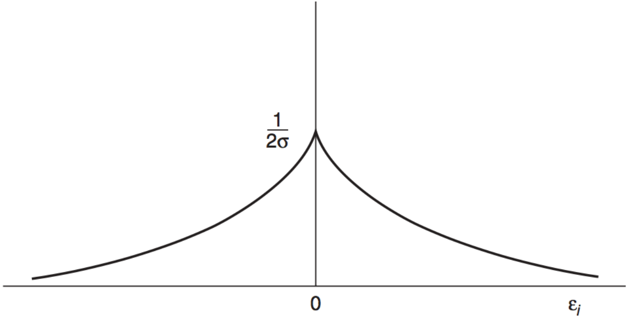
# Heavy-Tailed Distribution: Example-II

$t$-distribution (e.g., $t_4$) versus contaminated Gaussian distribution:



# Heavy-Tailed Distribution: Example-III

Double exponential distribution $p(\varepsilon_i) = \frac{1}{2\sigma} e^{-|\varepsilon_i|/\sigma}, i = 1, 2, ..., n$:



Does it decay slower than Gaussian with $e^{-(\varepsilon_i)^2/\sigma^2}$?

# Robust Statistics: Sample Mean vs. Sample Median

1. Sample mean is defined for the following example to be $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Sample median is more conveniently defined in terms of the order statistics $x_{(1)}, x_{(2)}, ..., x_{(n)}$, obtained by sorting the observations $\mathbf{x} = (x_1, x_2, ...., x_n)$ in increasing order.
    - If $n$ is odd, then $n = 2m - 1$ for some integer $m$, and in that case the sample median is calculated as $Med(\mathbf{x}) = x_{(m)}$.
    - If $n$ is even, then $n = 2m$ for some integer $m$, and then any value between $x_{(m)}$ and $x_{(m+1)}$ satisfies the definition of a sample median, and it is customary to take the average $Med(\mathbf{x}) = \left(x_{(m)} + x_{(m+1)}\right)/2$

# Robust Statistics: Sample SD vs. Sample MAD

1. Sample standard deviation (SD) is calculated as the square-root of $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. SD tells about the dispersion of the data.
2. Sample median absolute deviation (MAD) is calculated as:

$$MAD(\mathbf{x}) = Med(|\mathbf{x} - Med(\mathbf{x})|). \tag{3}$$

To make the MAD comparable to the SD, we define the normalized MAD ("MADN") as

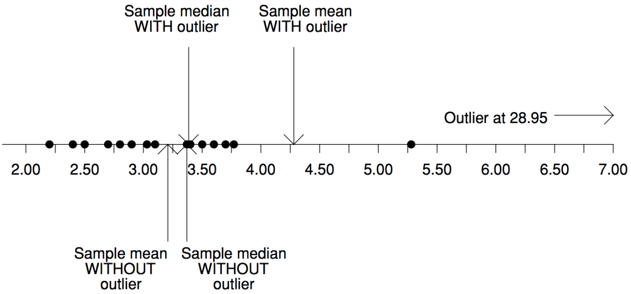
$$MADN(\mathbf{x}) = \frac{MAD(\mathbf{x})}{0.6745}. \tag{4}$$

The tuning constant 0.6745 makes $MADN(\mathbf{x})$ an approximately unbiased estimator of $\sigma$ if $n$ is large and $x_i \sim \mathcal{N}(0, \sigma^2)$.

# Robust Statistics: A Convincing Example

Example: Consider the following 24 determinations of the copper content in wholemeal flour (in parts per million)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2.20 | 2.20 | 2.40 | 2.40 | 2.50 | 2.70 | 2.80 | 2.90 |
| 3.03 | 3.03 | 3.10 | 3.37 | 3.40 | 3.40 | 3.40 | 3.50 |
| 3.60 | 3.70 | 3.70 | 3.70 | 3.70 | 3.77 | 5.28 | $\boxed{28.95}$ |



# Robust Statistics: A Convincing Example

For the "copper content" example, we have

- with all the data points, sample mean = 4.28, sample std. = 5.30;
- with the outlier deleted, sample mean = 3.21, sample std. = 0.69;
- with the outlier "corrected" from 28.95 to 2.895, sample mean = 3.195, sample std. = 0.675; (used as ideal result)
- with all the data points, sample median = 3.38, sample MADN = 0.53;
- with the outlier deleted, sample median = 3.37, sample MADN = 0.50;

Question: Why not always use the median and MADN?

Answer: If the data follows ideal condition, i.e., Gaussian i.i.d., then these robust estimators have “poorer” statistical performance than that of the classical estimators designed for the ideal condition.

# Robust Estimation in Essence [Maronna, 06]

- The classical estimators are in some sense “optimal” when the data are exactly distributed according to the assumed model, but can be very sub-optimal when the distribution of the data differs from the assumed model by a “small” amount.
- Robust estimators on the other hand maintain approximately optimal performance, not just under the assumed model, but under “small” perturbations of it too.

# Robust Estimation in Essence

optimal under perfect condition

fail under adverse condition

near-optimal under perfect condition

work well under adverse condition

# Robust Regression

## Definition

Robust regression is an alternative to ordinary least squares regression when data is contaminated with outliers or demonstrates itself to be heavy-tailed but without precise knowledge of the underlying distribution.

Robustness is required!

# Robust Regression

We define a maximum-likelihood type estimator (M-estimator) as:

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(e_i(\boldsymbol{\beta})\right) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right). \tag{5}$$

How to choose the function $\rho(\cdot)$?

We can follow two different strategies:

1. Strategy-I: we can take the robust criterion function $\rho(\cdot) = -ln\left[p_{\varepsilon}(e_i(\beta))\right]$, where $p_{\varepsilon}$ is a selected heavy-tailed distribution, although it may mismatch the true one.
2. Strategy-II: we can let $\rho(\cdot)$ be any of the well acknowledged robust criterion functions that
   1. should be non-negative, $\rho(z) \geq 0$;
   2. should be equal to zero when its argument is zero, $\rho(0) = 0$;
   3. should be symmetric, $\rho(z) = \rho(-z)$;
   4. should be monotone in $|z|$.

   Here, $z$ is implicitly assumed to have been normalized.

# Robust Regression: Strategy-I

Strategy-I: Using the double exponential distribution in the M-estimation framework, yields the classic L1-norm (least-absolute-value) regression.

$$\begin{aligned}\arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ln p_{\varepsilon}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ln(2\sigma) + \left(\frac{|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|}{\sigma}\right) \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \boldsymbol{\beta}| \end{aligned} \tag{6}$$

# Robust Regression: Strategy-II

Use some well acknowledged robust criterion functions

Note: 1. $\Psi(z) = \partial\rho(z)$, $w(z) = \Psi(z)/z$; 2. $t, a, b, c$ are tuning constants.

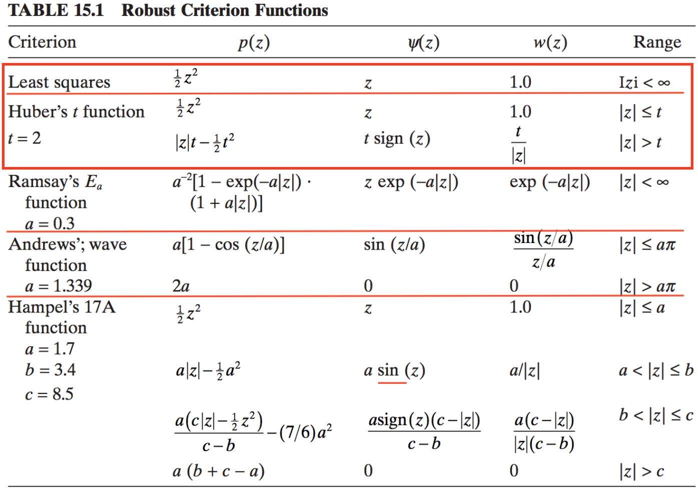
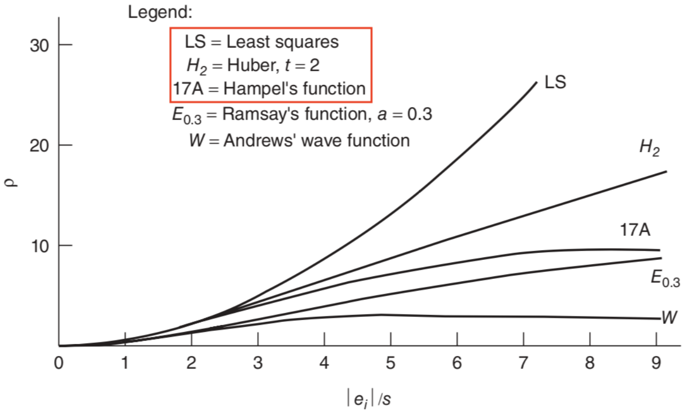


**TABLE 15.1 Robust Criterion Functions**

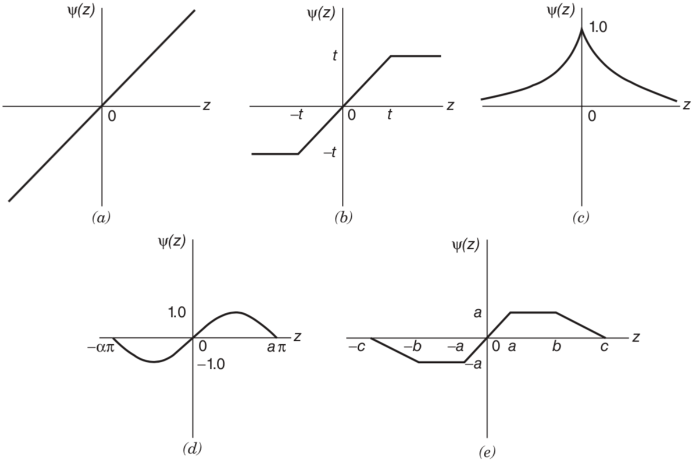
| Criterion | $p(z)$ | $\psi(z)$ | $w(z)$ | Range |
|---|---|---|---|---|
| Least squares | $\frac{1}{2}z^2$ | $z$ | 1.0 | Izi $< \infty$ |
| Huber's $t$ function | $\frac{1}{2}z^2$ | $z$ | 1.0 | $\lvert z\rvert \le t$ |
| $t = 2$ | $\lvert z\rvert t - \frac{1}{2}t^2$ | $t$ sign $(z)$ | $\frac{t}{\lvert z\rvert}$ | $\lvert z\rvert > t$ |
| Ramsay's $E_a$ function<br>$a = 0.3$ | $a^{-2}[1 - \exp(-a\lvert z\rvert) \cdot (1 + a\lvert z\rvert)]$ | $z \exp(-a\lvert z\rvert)$ | $\exp(-a\lvert z\rvert)$ | $\lvert z\rvert < \infty$ |
| Andrews'; wave function | $a[1 - \cos(z/a)]$ | $\sin(z/a)$ | $\frac{\sin(z/a)}{z/a}$ | $\lvert z\rvert \le a\pi$ |
| $a = 1.339$ | $2a$ | 0 | 0 | $\lvert z\rvert > a\pi$ |
| Hampel's 17A function<br>$a = 1.7$ | $\frac{1}{2}z^2$ | $z$ | 1.0 | $\lvert z\rvert \le a$ |
| $b = 3.4$ | $a\lvert z\rvert - \frac{1}{2}a^2$ | $a \sin(z)$ | $a/\lvert z\rvert$ | $a < \lvert z\rvert \le b$ |
| $c = 8.5$ | $\frac{a(c\lvert z\rvert - \frac{1}{2}z^2)}{c-b} - (7/6)a^2$ | $\frac{a\,\text{sign}(z)(c-\lvert z\rvert)}{c-b}$ | $\frac{a(c-\lvert z\rvert)}{\lvert z\rvert(c-b)}$ | $b < \lvert z\rvert \le c$ |
| | $a(b + c - a)$ | 0 | 0 | $\lvert z\rvert > c$ |

# Robust Regression: Strategy-II

Illustration of the above robust criterion functions:



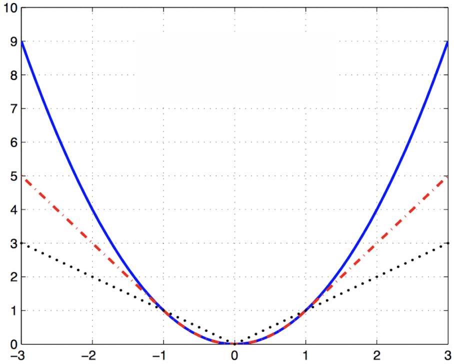
# Robust Regression: Strategy-II



Robust influence functions: (a) least-squares; (b) Huber's $t$ function; (c)Ramsay's $E_a$ function; (d) Andrews' wave function; (e) Hample's 17A function.

# Robust Regression: Strategy-II

Comparison between the least-squares, the least-absolute-value, and the Huber criterion functions.



# Robust Regression: IRLS Algorithm

To get an approximate scale-invariant M-estimator, we use instead

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{s}\right), \tag{7}$$

where $s$ is a robust estimate of the std of the random error terms.

A popular choice of s is the normalized median absolute deviation.

How to get the solution?

Equate the first-order derivatives of $\rho$ with respect to $\beta_j, j = 0, 1, 2, ..., k$ to zero, yielding

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{s}\right) = 0, \quad j = 0, 1, 2, ..., k, \tag{8}$$

where $\psi(\cdot)$ is the first-order derivative of $\rho(\cdot)$, $x_{ij}$ is the $j$-th input value of the $i$-th data point, and $x_{i0} = 1$ particularly.

$\psi(\cdot)$ controls the weight given to each residual, which influences the parameter fit.

Recall:

- The $\psi(\cdot)$ function for the least-squares is unbounded, thus least-squares tends to be non-robust when used with data arising from a heavy-tailed distribution.
- The Huber function has a monotone $\psi(\cdot)$ function but does not weight large residuals as heavily as the least-squares.

In general, the $\psi(\cdot)$ function is nonlinear and the robust-M estimate must be solved by iterative methods, e.g., iteratively reweighted least squares (IRLS) attributed to Beaton and Tukey.

Approximate

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{s}\right) = 0, \quad j = 0, 1, 2, ..., k \tag{9}$$

at the $(\eta + 1)$-th iteration by

$$\sum_{i=1}^{n} x_{ij} w_{i,\eta}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) = 0, \quad j = 0, 1, 2, ..., k \tag{10}$$

where

$$w_{i,\eta} = \begin{cases} \frac{\psi\left[\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\eta}\right)/s\right]}{\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\eta}\right)/s}, & y_i \neq \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\eta} \\ 1, & y_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{\eta}. \end{cases} \tag{11}$$

In matrix form:

$$\mathbf{X}^T\mathbf{W}_\eta\mathbf{X}\beta = \mathbf{X}^T\mathbf{W}_\eta\mathbf{y}, \tag{12}$$

where $\mathbf{W}_\eta$ is an $n \times n$ diagonal matrix of weights with diagonal elements $w_{1,\eta}, w_{2,\eta}, \ldots, w_{n,\eta}$ given beforehand.

## Complete Algorithm

- Find an initial estimate $\hat{\beta}^0$
- Select a threshold $\delta_T$
- For $\eta = 0, 1, \ldots,$ do
  1. Compute $\hat{\beta}^{\eta+1} = \left(\mathbf{X}^T\mathbf{W}_\eta\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_\eta\mathbf{y}$
  2. Compute $\delta = ||\hat{\beta}^{\eta+1} - \hat{\beta}^{\eta}||_2$
  3. If $\delta < \delta_T$, then terminate the iterative process; otherwise $\eta = \eta + 1$ and repeat the above process.

How to choose a good starting point $\hat{\beta}^0$?

# Robust Regression: Asymptotic Properties of $\hat{\boldsymbol{\beta}}_M$

Huber [1973] showed that $\hat{\boldsymbol{\beta}}_M$ approaches asympototically to normal distribution with the covariance matrix equal to

$$\sigma^2 \frac{E\left[\psi^2(\varepsilon/\sigma)\right]}{E^2\left[\psi'(\varepsilon/\sigma)\right]}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{13}$$

A numerical approximation of (13) is given by

$$\frac{ns^2}{n-p}\frac{\sum_{i=1}^n \psi^2((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/s)}{\sum_{i=1}^n \psi'((y_i-\mathbf{x}_i^T\boldsymbol{\beta})/s)}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{14}$$

# Robust Regression: Performance Metrics

Two important properties of a robust estimator:

1. Breakdown Point *bp*: the maximum fraction of outliers in the dataset that a robust estimator can handle.
2. Efficiency, $\eta = \frac{\sum_i^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{LS})^2}{\sum_i^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_M)^2}$, under perfect random error condition, i.e., Gaussian i.i.d.. We desire $\eta \approx 1$.

# Robust Regression for Nonlinear Model

Similarly, we define the M-estimator as:

$$\hat{\boldsymbol{\theta}}_M = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho\left(\frac{e_i(\boldsymbol{\theta})}{s}\right) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - f(\mathbf{x}_i^T; \boldsymbol{\theta})}{s}\right). \qquad (15)$$

But we have to rely on numerical method such as gradient descent to solve the parameter estimate, $\hat{\boldsymbol{\theta}}_M$.

# Summary

- Outlier versus influential point
- Heavy-tailed distribution, such as contaminated Gaussian, double exponential, etc.
- Robust statistics and robust regression
- Robust criterion functions and the associated influence functions
- IRLS algorithm (iterative procedure)
- Break point and efficiency of a robust estimator