Robustness and Other Things

Ruben Zamar
Department of Statistics, UBC

September 29, 2012

PART I

PART I

ROBUST STATISTICS

CLASSICAL STATISTICS TRIES

TO FIT WELL

ALL THE DATA

ROBUST STATISTICS TRIES

TO FIT WELL

MOST OF THE DATA

J. W. Tukey (1979)

" ... It is perfectly proper to use both

classical and robust/resistant

methods routinely ..."

J. W. Tukey (1979)

"...and only worry when

they differ enough to matter"

J. W. Tukey (1979)

“...when they differ, you should think hard”

STACK LOSS DATA EXAMPLE

Stack Loss Data

- **Data available in R (dataset name = stackloss)**

Stack Loss Data

- **Data available in R (dataset name = stackloss)**
- **21 daily observations of the oxidation of ammonia to nitric acid**

Stack Loss Data

- **Data available in R (dataset name = stackloss)**
- **21 daily observations of the oxidation of ammonia to nitric acid**
- **First published by Brownlee (1965)**

Stack Loss Data

- Data available in R (dataset name = stackloss)
- 21 daily observations of the oxidation of ammonia to nitric acid
- First published by Brownlee (1965)
- Extensively studied in the statistical literature

Stack Loss Data

- **Data available in R (dataset name = stackloss)**
- **21 daily observations of the oxidation of ammonia to nitric acid**
- **First published by Brownlee (1965)**
- **Extensively studied in the statistical literature**
  - **Daniel and Wood, 1980, Chapters 5 and 7**

Stack Loss Data

- Data available in R (dataset name = stackloss)
- 21 daily observations of the oxidation of ammonia to nitric acid
- First published by Brownlee (1965)
- Extensively studied in the statistical literature
  - Daniel and Wood, 1980, Chapters 5 and 7
  - Atkinson, 1985, pp. 129-136, 267-8

Stack Loss Data

- Data available in R (dataset name = stackloss)
- 21 daily observations of the oxidation of ammonia to nitric acid
- First published by Brownlee (1965)
- Extensively studied in the statistical literature
  - Daniel and Wood, 1980, Chapters 5 and 7
  - Atkinson, 1985, pp. 129-136, 267-8
  - Venables and Ripley, 1997

Input Variables

Input Variables

The rate flow of cooling air **(Air.Flow)**

Stack Loss Data (continued)

Input Variables

The rate flow of cooling air	**(Air.Flow)**
The temperature of the cooling inlet water	**(Water.Temp)**

Input Variables

The rate flow of cooling air	(**Air.Flow**)
The temperature of the cooling inlet water	(**Water.Temp**)
The concentration of acid	(**Acid.Conc**.)

Input Variables

The rate flow of cooling air	**(Air.Flow)**
The temperature of the cooling inlet water	**(Water.Temp)**
The concentration of acid	**(Acid.Conc.)**

Output Variable

Input Variables

The rate flow of cooling air	**(Air.Flow)**
The temperature of the cooling inlet water	**(Water.Temp)**
The concentration of acid	**(Acid.Conc.)**

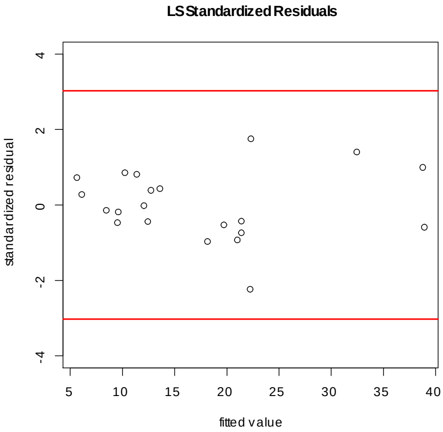
Output Variable

An inverse measure for the overall efficiency of the plant	**(stack.loss)**

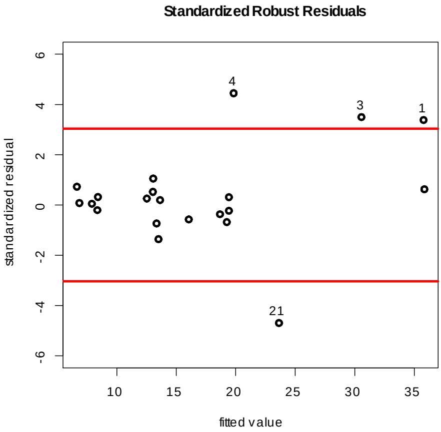
Classical and Robust Linear Models

Regression Coefficient Estimate	LS	Robust
Intercept	-39.9	-37.6
Air Flow	0.72	0.80
Water Temperature	**1.3**	**0.6**
Acid Concentration	**-0.15**	**-0.07**
Residual SE	**3.2**	**1.8**

LS Residual Plot



Robust Residual Plot



Thinking Hard...

- Daniel and Wood (1971, Chapter 5, page 81) noticed a different behavior in the response variable whenever the water temperature was over 60 degrees.

Thinking Hard...

- Daniel and Wood (1971, Chapter 5, page 81) noticed a different behavior in the response variable whenever the water temperature was over 60 degrees.
- The plant needs to stabilize after the water temperature reaches 60 degrees.

Thinking Hard...

- Daniel and Wood (1971, Chapter 5, page 81) noticed a different behavior in the response variable whenever the water temperature was over 60 degrees.
- The plant needs to stabilize after the water temperature reaches 60 degrees.
- They concluded that observations obtained with Water Temperature $\geq$ 60 degrees require special attention, and should be removed from the analysis.

Thinking Hard...

- Daniel and Wood (1971, Chapter 5, page 81) noticed a different behavior in the response variable whenever the water temperature was over 60 degrees.
- The plant needs to stabilize after the water temperature reaches 60 degrees.
- They concluded that observations obtained with Water Temperature $\geq$ 60 degrees require special attention, and should be removed from the analysis.
- These are **cases 1, 3, 4 and 21** directly uncovered by the robust fit.

PART II

PART II

SOME TECHNICAL CONCEPTS

Statistical Models

- Many data can be modeled as follows:

$$\text{OUTPUT DATA} = \text{SIGNAL(INPUT DATA, } \boldsymbol{\theta}) + \text{NOISE}$$

A Closer Look at the Noise

- We distinguish two types of noise

A Closer Look at the Noise

- We distinguish two types of noise

1. **“TYPICAL” NOISE**

A Closer Look at the Noise

- We distinguish two types of noise
  1. **"TYPICAL" NOISE**
  2. **ATYPICAL NOISE**

Sources of Typical Noise

- Typical noise comes from

NATURAL FLUCTUATIONS

MEASUREMENT ERRORS

ITEM TO ITEM VARIABILITY, ETC

Sources of Typical Noise

- Typical noise comes from

NATURAL FLUCTUATIONS
MEASUREMENT ERRORS
ITEM TO ITEM VARIABILITY, ETC

- Not necessarily "Gaussian Noise"

Sources of Typical Noise

- Typical noise comes from

NATURAL FLUCTUATIONS

MEASUREMENT ERRORS

ITEM TO ITEM VARIABILITY, ETC

- Not necessarily "Gaussian Noise"
- Other classical parametric models such as Gamma, Weibull, Poisson, etc

Sources of Atypical Noise

WHERE DOES ATYPICAL NOISE COME FROM?

Sources of Atypical Noise

WHERE DOES ATYPICAL NOISE COME FROM?

OUTLIERS AND GROSS ERRORS
MEASUREMENTS OF UNEVEN QUALITY (mixture)
DATA CONTAMINATION (mixture)
MISSING DATA (declared or unsuspected)
DATA DUPLICATIONS, ETC

STATISTICIAN TASKS (A VERY SIMPLIFIED VIEW)

- **Filter the noise in data (both typical and atypical)**

STATISTICIAN TASKS (A VERY SIMPLIFIED VIEW)

- **Filter the noise in data (both typical and atypical)**
- **Extract the signal from data**

- **Filter the noise in data (both typical and atypical)**
- **Extract the signal from data**
- **Measure the noise strength**

STATISTICIAN TASKS (A VERY SIMPLIFIED VIEW)

- **Filter the noise in data (both typical and atypical)**
- **Extract the signal from data**
- **Measure the noise strength**
- **Assess uncertainty**

STATISTICIAN TASKS (A VERY SIMPLIFIED VIEW)

- **Filter the noise in data (both typical and atypical)**
- **Extract the signal from data**
- **Measure the noise strength**
- **Assess uncertainty**
- **Predict likely future data**

Output from Statistical Analysis

- **Point Estimates**

$$\theta$$

Output from Statistical Analysis

- **Point Estimates**

$$\theta$$

- **Confidence Regions**

$$Cov\left(\hat{\theta}\right), \qquad \text{Confidence Region for } \boldsymbol{\theta}$$

Output from Statistical Analysis

- **Point Estimates**

$$\boldsymbol{\theta}$$

- **Confidence Regions**

$$Cov\left(\hat{\boldsymbol{\theta}}\right), \qquad \text{Confidence Region for } \boldsymbol{\theta}$$

- **Prediction / Interpolation**

$$\widehat{SIGNAL} \quad \pm \quad 2 \times SE\left(\widehat{SIGNAL}\right)$$

Only Typical Noise (OLD CLASSICAL STATISTICS)

- TYPICALLY

$$\hat{\theta} \to \theta$$

Only Typical Noise (OLD CLASSICAL STATISTICS)

- TYPICALLY

$$\hat{\theta} \longrightarrow \theta$$

- AND

$$Cov\left(\hat{\theta}\right) = \frac{1}{n} C_{\hat{\theta}} \longrightarrow 0$$

Only Typical Noise (OLD CLASSICAL STATISTICS)

- TYPICALLY

$$\hat{\theta} \to \theta$$

- AND

$$Cov\left(\hat{\theta}\right) = \frac{1}{n} C_{\hat{\theta}} \to 0$$

- BETTER RESULTS WHEN $C_{\hat{\theta}}$ IS "SMALL" $\Longrightarrow$ USE EFFICIENT PROCEDURES

Only Typical Noise (OLD CLASSICAL STATISTICS)

- TYPICALLY

$$\hat{\theta} \rightarrow \theta$$

- AND

$$Cov\left(\hat{\theta}\right) = \frac{1}{n} C_{\hat{\theta}} \rightarrow 0$$

- BETTER RESULTS WHEN $C_{\hat{\theta}}$ IS “SMALL” $\Longrightarrow$ USE EFFICIENT PROCEDURES
- I BELIEVE THAT TOO MUCH ATTENTION IS GIVEN TO THE PROBLEM OF MINIMIZING $C_{\hat{\theta}}$

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\hat{\theta} \to \Delta, \quad \Delta \neq \theta$$

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\hat{\boldsymbol{\theta}} \rightarrow \Delta, \qquad \Delta \neq \boldsymbol{\theta}$$

- The difference between Δ and $\boldsymbol{\theta}$ is called "contamination bias" (cb)

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\hat{\boldsymbol{\theta}} \to \Delta, \qquad \Delta \neq \boldsymbol{\theta}$$

- The difference between Δ and $\boldsymbol{\theta}$ is called "contamination bias" (cb)
- $cb\left(\hat{\boldsymbol{\theta}}\right)$ is of order 1 while $Cov\left(\hat{\boldsymbol{\theta}}\right)$ of order $1/n$

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\hat{\boldsymbol{\theta}} \to \Delta, \quad \Delta \neq \boldsymbol{\theta}$$

- The difference between Δ and $\boldsymbol{\theta}$ is called "contamination bias" (cb)
- $cb\left(\hat{\boldsymbol{\theta}}\right)$ is of order 1 while $Cov\left(\hat{\boldsymbol{\theta}}\right)$ of order $1/n$
- Therefore, for large n, $cb\left(\hat{\boldsymbol{\theta}}\right)$ should be the leading concern

A (Classical) Robustness Model

- Let F_θ be the joint distribution for the data

A (Classical) Robustness Model

- Let F_θ be the joint distribution for the data
- Let H be an arbitrary distribution on the data space

A (Classical) Robustness Model

- Let F_{θ} be the joint distribution for the data
- Let H be an arbitrary distribution on the data space
  - H represents the "contamination generating mechanism"

A (Classical) Robustness Model

- Let F_{θ} be the joint distribution for the data
- Let H be an arbitrary distribution on the data space
  - H represents the "contamination generating mechanism"
- Let $0 \leq \epsilon < 1$

A (Classical) Robustness Model

- Let F_{θ} be the joint distribution for the data
- Let H be an arbitrary distribution on the data space
  - H represents the "contamination generating mechanism"
- Let $0 \leq \epsilon < 1$
  - ϵ represents the fraction of contamination

A (Classical) Robustness Model

- Let $F_{\boldsymbol{\theta}}$ be the joint distribution for the data
- Let H be an arbitrary distribution on the data space
  - H represents the "contamination generating mechanism"
- Let $0 \leq \epsilon < 1$
  - ϵ represents the fraction of contamination
- The robustness model

$$\mathcal{F}_{\epsilon} = \{F : F = (1 - \epsilon)\, F_{\boldsymbol{\theta}} + \epsilon H\}$$

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including
  - Empirical distributions F_n [in this case $T_n = T(F_n)$]

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including
  - Empirical distributions F_n [in this case $T_n = T(F_n)$]
  - The robustness neighborhood $\mathcal{F}_\epsilon$

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including
  - Empirical distributions F_n [in this case $T_n = T(F_n)$]
  - The robustness neighborhood $\mathcal{F}_\epsilon$
- Suppose also that T is consistent

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including
  - Empirical distributions F_n [in this case $T_n = T(F_n)$]
  - The robustness neighborhood $\mathcal{F}_\epsilon$
- Suppose also that T is consistent
  - $T(F_n) \rightarrow T(F)$ $a.s.[F]$ for all $F \in \mathcal{F}_\epsilon$

Estimating Functional

- Let $T(F)$ be an estimating functional for $\boldsymbol{\theta}$
- Suppose $T(F)$ is defined on a set of distributions including
  - Empirical distributions F_n [in this case $T_n = T(F_n)$]
  - The robustness neighborhood $\mathcal{F}_\epsilon$
- Suppose also that T is consistent
  - $T(F_n) \to T(F)\ a.s.[F]$ for all $F \in \mathcal{F}_\epsilon$
- A robust estimate would satisfy $T(F) \approx T(F_{\boldsymbol{\theta}})$ when ϵ is relatively small

Maxbias and Other Robustness Measures

- Consider an appropriate distance d on the parameter space Θ

Maxbias and Other Robustness Measures

- Consider an appropriate distance d on the parameter space Θ
- Contamination bias:

$$b_T(\epsilon, F) = d\left[T(F), T(F_{\boldsymbol{\theta}})\right], \qquad F \in \mathcal{F}_\epsilon$$

Maxbias and Other Robustness Measures

- Consider an appropriate distance d on the parameter space Θ
- Contamination bias:

$$b_T(\epsilon, F) = d\left[T(F), T(F_{\boldsymbol{\theta}})\right], \qquad F \in \mathcal{F}_\epsilon$$

- Contamination maxbias

$$B_T(\epsilon) = \sup_{F \in \mathcal{F}_\epsilon} d\left[T(F), T(F_{\boldsymbol{\theta}})\right]$$

The Breakdown Point (BP) and Gross Error Sensitivity (GES)

- The BP of and estimating functional $T(F)$ is defined as follows

$$BP_T = \sup \{\epsilon : B_T(\epsilon) < \infty\}$$

The Breakdown Point (BP) and Gross Error Sensitivity (GES)

- The BP of and estimating functional $T(F)$ is defined as follows

$$BP_T = \sup\{\epsilon : B_T(\epsilon) < \infty\}$$

- The GES is defined as follows

$$GES_T = \left.\frac{d}{d\epsilon} B_T(\epsilon)\right|_{\epsilon=0} = B_T'(0)$$

The Breakdown Point (BP) and Gross Error Sensitivity (GES)

- The BP of and estimating functional $T(F)$ is defined as follows

$$BP_T = \sup\{\epsilon : B_T(\epsilon) < \infty\}$$

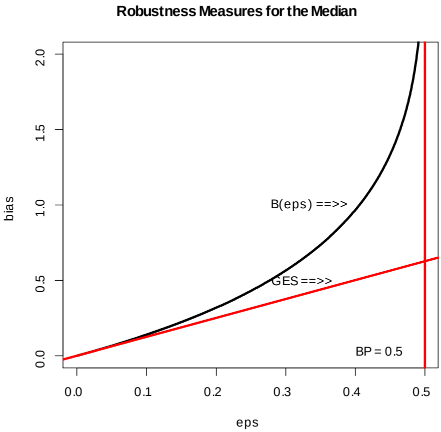
- The GES is defined as follows

$$GES_T = \left.\frac{d}{d\epsilon} B_T(\epsilon)\right|_{\epsilon=0} = B_T'(0)$$

- Therefore

$$B_T(\epsilon) = \epsilon GES_T + o(\epsilon)$$

The Median - Gaussian Case



Examples of the Types of Results One Obtains in this Setting

- Huber (1964) showed that

$$B_{\text{Median}}\left(\epsilon\right) \leq B_{T}\left(\epsilon\right)$$

for all translation equivariant estimate T, and for all $\epsilon > 0$

Examples of the Types of Results One Obtains in this Setting

- Huber (1964) showed that

$$B_{\text{Median}}(\epsilon) \leq B_T(\epsilon)$$

  for all translation equivariant estimate T, and for all $\epsilon > 0$

  - $T_n(X_1 + c, X_2 + c, ..., X_n + c) = T_n(X_1, X_2, ..., X_n) + c$

Examples of the Types of Results One Obtains in this Setting

- Huber (1964) showed that

$$B_{\text{Median}}(\epsilon) \leq B_T(\epsilon)$$

  for all translation equivariant estimate T, and for all $\epsilon > 0$
  - $T_n(X_1 + c, X_2 + c, ..., X_n + c) = T_n(X_1, X_2, ..., X_n) + c$
- Huber (1964) also showed that

$$B_{\text{Median}}(\epsilon) = F_0^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$$

Examples of the Types of Results One Obtains in this Setting

- Huber (1964) showed that

$$B_{\text{Median}}(\epsilon) \leq B_T(\epsilon)$$

  for all translation equivariant estimate T, and for all $\epsilon > 0$

  - $T_n(X_1 + c, X_2 + c, ..., X_n + c) = T_n(X_1, X_2, ..., X_n) + c$

- Huber (1964) also showed that

$$B_{\text{Median}}(\epsilon) = F_0^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$$

- Martin, Yohai and Zamar (1989) obtained minimax-bias results for multiple linear regression

More General (and Realistic) Robust Model

- Typical robust methods work as follows

More General (and Realistic) Robust Model

- Typical robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)

More General (and Realistic) Robust Model

- Typical robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases

More General (and Realistic) Robust Model

- Typical robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures

More General (and Realistic) Robust Model

- Typical robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures
  - Percentage of unusual data points is relatively small

More General (and Realistic) Robust Model

- Typical robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures
  - Percentage of unusual data points is relatively small
  - Hopefully way below 50%

The Shape of Data Tables

- Most statistical applications

D
A
T
A
T
A
B
L
E

LARGE and Shallow Data Tables

- In some applications we deal with datasets like this

D A T A T A B L E

Some Examples of This Type of Data

- Microarray data

p	number of genes	several thousands
n	number of patients	at best a few hundreds

Some Examples of This Type of Data

- Microarray data

p	number of genes	several thousands
n	number of patients	at best a few hundreds

- Asthenosphere Data

p	number of locations	about 5000
n	number of days	3650 days

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"
- Rows may be only partially spoiled

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"
- Rows may be only partially spoiled
- Consider "cell contamination" as opposed to "row contamination"

Implications for Robustness Methods

- Downweighting entire rows may be too “wasteful”
- Rows may be only partially spoiled
- Consider “cell contamination” as opposed to “row contamination”
- Need for more flexible robustness methods and models

An Example: Cell-wise Contamination Model

- From Alqallaf's PhD thesis and Alqallaf et al (2009)

$$\mathbf{X} = (I - B)\,\mathbf{Y} + B\mathbf{Z}$$

$$B = diag\,(B_1, B_2, ..., B_p)$$

$$P\,(B_i = 1) = 1 - P\,(B_i = 0) = \epsilon_i$$

An Example: Cell-wise Contamination Model

- From Alqallaf's PhD thesis and Alqallaf et al (2009)

$$\mathbf{X} = (I - B)\,\mathbf{Y} + B\mathbf{Z}$$

$$B = diag\,(B_1, B_2, ..., B_p)$$

$$P\,(B_i = 1) = 1 - P\,(B_i = 0) = \epsilon_i$$

- Lot's of room for research at the MSc and PhD levels on this area

PART III

PART III

DATA MINING

Definition of Data Mining

Data mining is the analysis of (large) observational datasets

Definition of Data Mining

Data mining is the analysis of (large)
observational datasets
to find **unsuspected relationships**

Definition of Data Mining

Data mining is the analysis of (large) observational datasets
to find **unsuspected relationships**
to **sumarize the data in novel ways**

Definition of Data Mining

Data mining is the analysis of (large)
observational datasets
to find **unsuspected relationships**
to **sumarize the data in novel ways**
understandable and useful
to the data owner.

WAL-MART EXAMPLE


WAL-MART
EXIT
EXIT
EXIT

Wal-Mart Data Warehouse

- Wal-Mart captures all the sale transactions in their 8,500 stores in 15 countries

Wal-Mart Data Warehouse

- Wal-Mart captures all the sale transactions in their 8,500 stores in 15 countries
- Continuously transmits these data to its massive 500 terabytes data warehouse.

Wal-Mart Data Warehouse

- Wal-Mart captures all the sale transactions in their 8,500 stores in 15 countries
- Continuously transmits these data to its massive 500 terabytes data warehouse.
- Over 3,500 suppliers access data on their products and perform data analyses

Wal-Mart Data Warehouse

- Wal-Mart captures all the sale transactions in their 8,500 stores in 15 countries
- Continuously transmits these data to its massive 500 terabytes data warehouse.
- Over 3,500 suppliers access data on their products and perform data analyses
- Identify **customer buying patterns** at the store display level

Wal-Mart Data Warehouse

- Wal-Mart captures all the sale transactions in their 8,500 stores in 15 countries
- Continuously transmits these data to its massive 500 terabytes data warehouse.
- Over 3,500 suppliers access data on their products and perform data analyses
- Identify **customer buying patterns** at the store display level
- **Goal:** To manage local store inventories and to identify new merchandising opportunities

Wal-Mart Famous Example

- men in their 20s who purchase beer on Fridays after work are also likely to buy a pack of diapers

Wal-Mart Famous Example

- men in their 20s who purchase beer on Fridays after work are also likely to buy a pack of diapers

- put beer and diapers near each other to increase sales for both

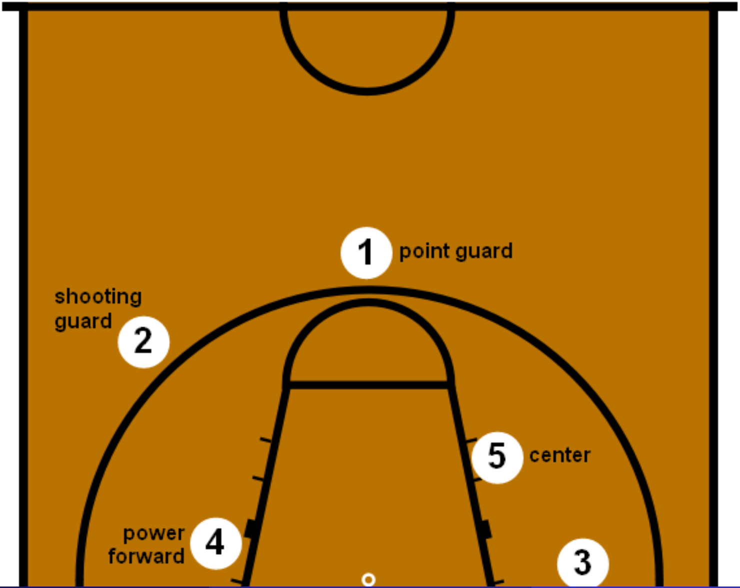
Wal-Mart Famous Example

- men in their 20s who purchase beer on Fridays after work are also likely to buy a pack of diapers
- put beer and diapers near each other to increase sales for both
- put one (but not both) of these products on sale on Friday evenings

NEW YORK KNICKS
Vs
CLEVELAND CAVALIERS

EXAMPLE

Basketball Positions

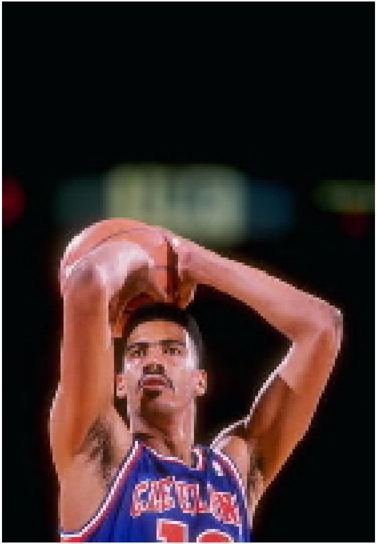


Basketball Game (Jan 6, 1995)

The New York Knicks (103) Versus
the Cleveland Cavaliers (93)

Mark Price (Cavaliers' Player)

John Williams (Cavaliers' Player)



Basketball and Data Mining

- Computer softwares analyze the movements of players to help coaches orchestrate plays and strategies.

Basketball and Data Mining

- Computer softwares analyze the movements of players to help coaches orchestrate plays and strategies.
- NBA is exploring a data mining application that can be used in conjunction with image recordings of basketball games.

The January 6, 1995 Basketball Game (Continues)

- The Cavaliers lost the game (103 - 93) and had an overall shooting percentage of **49.30%**

The January 6, 1995 Basketball Game (Continues)

- The Cavaliers lost the game (103 - 93) and had an overall shooting percentage of **49.30%**
- When **Mark Price** played the Point Guard position, **John Williams** attempted four jump shots and made each one!

The January 6, 1995 Basketball Game (Continues)

- The Cavaliers lost the game (103 - 93) and had an overall shooting percentage of **49.30%**
- When **Mark Price** played the Point Guard position, **John Williams** attempted four jump shots and made each one!
- It is interesting because it differs considerably from the average shooting percentage of 49.30%.

PARALLEL BETWEEN STATISTICS AND DATA MINING

Data Size

Statistics Data Mining

Data Size

Statistics	Data Mining
n = tens, hundreds, thousands (?)	n = thousands, millions

Data Size

Statistics	Data Mining
n = tens, hundreds, thousands (?)	n = thousands, millions
p = a handful, rarely more than a few tens	p = hundreds, thousands

Logical Data Sequence

Statistics Data Mining

Logical Data Sequence

Statistics	Data Mining
Data collected to answer a given question	Data collected electronically for future possible use

Logical Data Sequence

Statistics	Data Mining
Data collected to answer a given question	Data collected electronically for future possible use
Questions come first, data come second	Data come first, questions come second

Data Type

Statistics Data Mining

Statistics	Data Mining
First hand data	Second hand data

Data Type

Statistics	Data Mining
First hand data	Second hand data
Data collected to fit/test a model	Data collected ellectronically for future "mining"

Data Type

Statistics	Data Mining
First hand data	Second hand data
Data collected to fit/test a model	Data collected electronically for future "mining"
Case-control studies Sampling surveys Designed experiments etc	Supermarket sales Internet traffic Stock market transactions etc

Data Processing

Statistics

Data Mining

Data Processing

Statistics	Data Mining
Hand-on procedures	Highly automated procedures

Data Processing

Statistics	Data Mining
Hand-on procedures	Highly automated procedures
Data analyzed by people with the aid of computers	Data processed by computer algorithms with the aid of people

Typical Tasks

Statistics Data Mining

Typical Tasks

Statistics	Data Mining
Model fitting/testing	Patterns seeking and identification
Confidence and prediction intervals	Grouping
Sample size / power calculations	Ranking/short listing

Research Goals

Statistics

Data Mining

Statistics	Data Mining
Develop better statistical procedures	Develop better/faster algorithm for data mining
Study statistical properties of methods	Study empirical performance of mining algorithms
Asymptotic distributions of statistical procedures	Construct scalable data mining systems
Asymptotic approximations	

Statistics Data Mining

Dissemination of Research Results

Statistics	Data Mining
Mostly Journals	Mostly Conference Procedures

SUPERVISED AND UNSUPERVISED

LEARNING

SUPERVISED LEARNING

- LEARNING WITH A "TEACHER"

SUPERVISED LEARNING

- LEARNING WITH A “TEACHER”
- OBSERVATION OF THE “OUTPUT” VARIABLE (RESPONSE) ARE AVAILABLE

- LEARNING WITH A "TEACHER"
- OBSERVATION OF THE "OUTPUT" VARIABLE (RESPONSE) ARE AVAILABLE
- TRAINING DATA

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} & y_1 \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} & y_2 \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3d} & y_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{Nd} & y_N \end{pmatrix}$$

SUPERVISED LEARNING

- LEARNING WITH A "TEACHER"
- OBSERVATION OF THE "OUTPUT" VARIABLE (RESPONSE) ARE AVAILABLE
- TRAINING DATA

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} & y_1 \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} & y_2 \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3d} & y_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{Nd} & y_N \end{pmatrix}$$

- CONDITIONAL DISTRIBUTION OF THE "OUTPUT VARIABLE" GIVEN THE "INPUT VARIABLES"

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

CONTINUOUS OUTPUT

♠ LINEAR AND NONLINEAR REGRESSION

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

♠ LINEAR AND NONLINEAR REGRESSION

♠ PREDICTION AND FORECASTING

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

- ♠ LINEAR AND NONLINEAR REGRESSION
- ♠ PREDICTION AND FORECASTING

CATEGORICAL OUTPUT

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

♠ LINEAR AND NONLINEAR REGRESSION

♠ PREDICTION AND FORECASTING

CATEGORICAL OUTPUT

♠ CLASSIFICATION

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

- ♠ LINEAR AND NONLINEAR REGRESSION
- ♠ PREDICTION AND FORECASTING

CATEGORICAL OUTPUT

- ♠ CLASSIFICATION
- ♠ RANKING

SUPERVISED LEARNING (CONTINUED)

CONTINUOUS OUTPUT

- ♠ LINEAR AND NONLINEAR REGRESSION
- ♠ PREDICTION AND FORECASTING

CATEGORICAL OUTPUT

- ♠ CLASSIFICATION
- ♠ RANKING
- ♠ SHORT LISTING

UNSUPERVISED LEARNING

- LEARNING WITHOUT A "TEACHER"

UNSUPERVISED LEARNING

- LEARNING WITHOUT A “TEACHER”
- RESPONSE VARIABLE NOT GIVEN

UNSUPERVISED LEARNING

- LEARNING WITHOUT A “TEACHER”
- RESPONSE VARIABLE NOT GIVEN
- TRAINING DATA

$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1d} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2d} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3d} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & X_{N3} & \cdots & X_{Nd} \end{pmatrix}$$

UNSUPERVISED LEARNING

- SEARCH FOR FEATURES OF THE JOINT DISTRIBUTION OF THE GIVEN VARIABLES

♣ GROUPING ITEMS (CLUSTERING)

♣ DATA COMPRESSION (PCA, ICM)

♣ SEARCH FOR PATTERNS

THE **THREE** STEPS

IN

DATA MINING

BOSTON HOUSING DATA

14 SOCIOECONOMIC VARIABLES FOR 506 TOWNS IN THE BOSTON AREA

- ▲ Crime rate
- ▲ Prop. of non-retail business
- ▲ Charles River dummy variable
- ▲ Average number of rooms
- ▲ Distances to employment centres
- ▲ Property-tax rate
- ▲ $1000(Bk - 0.63)^2$ where (Bk = proportion of blacks)
- ▲ Prop. of residential land
- ▲ Nitric oxides concentration
- ▲ Proportion of owner occupied homes
- ▲ Access to radial highways
- ▲ Pupil-teacher ratio by town
- ▲ % Lower status population
- ▲ Median value of owner occupied homes

STEP 1:

DEFINING THE DATA MINING GOAL

ILLUSTRATION USING THE BOSTON HOUSING DATA

An incomplete list of possible goals includes:

- To predict the median house price using other variables

ILLUSTRATION USING THE BOSTON HOUSING DATA

An incomplete list of possible goals includes:

- To predict the median house price using other variables
- To predict the crime rate using other variables

ILLUSTRATION USING THE BOSTON HOUSING DATA

An incomplete list of possible goals includes:

- To predict the median house price using other variables
- To predict the crime rate using other variables
- Find linear/non-linear relations among the recorded variables

ILLUSTRATION USING THE BOSTON HOUSING DATA

An incomplete list of possible goals includes:

- To predict the median house price using other variables
- To predict the crime rate using other variables
- Find linear/non-linear relations among the recorded variables
- Find clusters of similar towns

ILLUSTRATION USING THE BOSTON HOUSING DATA

An incomplete list of possible goals includes:

- To predict the median house price using other variables
- To predict the crime rate using other variables
- Find linear/non-linear relations among the recorded variables
- Find clusters of similar towns
- Find clusters of similar variables

STEP 2:

ASSESSING (SCORING) DATA MINING RESULTS

SCORING

- Choice of evaluation criterion to assess how well a certain procedure realizes the mining goal

SCORING

- Choice of evaluation criterion to assess how well a certain procedure realizes the mining goal
- Related to choice of a loss function in Statistics

SCORING

- Choice of evaluation criterion to assess how well a certain procedure realizes the mining goal
- Related to choice of a loss function in Statistics
- non-robust scoring demands success in all the cases

SCORING

- Choice of evaluation criterion to assess how well a certain procedure realizes the mining goal
- Related to choice of a loss function in Statistics
- non-robust scoring demands success in all the cases
- robust scoring allows for partial success

SCORING

- Choice of evaluation criterion to assess how well a certain procedure realizes the mining goal
- Related to choice of a loss function in Statistics
- non-robust scoring demands success in all the cases
- robust scoring allows for partial success
- Example below will illustrate this point

STEP 3:

NUMERICAL IMPLEMENTATION

GETTING IT DONE

- Devise numerical procedures for efficient implementation of the mining task

GETTING IT DONE

- Devise numerical procedures for efficient implementation of the mining task
- Related to "Statistical Computing"

GETTING IT DONE

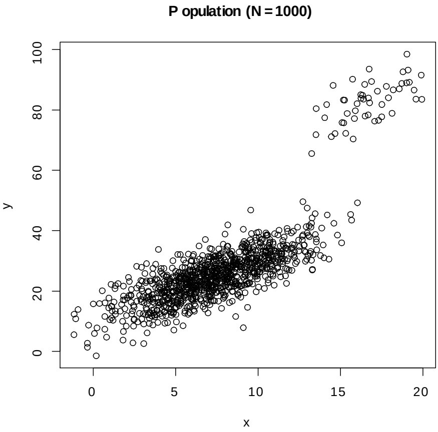
- Devise numerical procedures for efficient implementation of the mining task
- Related to “Statistical Computing”
- Emphasis placed here on "scalability”, regarding the number of case and variables.

PART IV

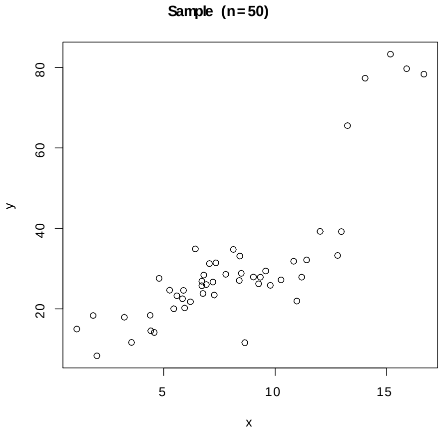
PART IV

ONE LAST EXAMPLE...

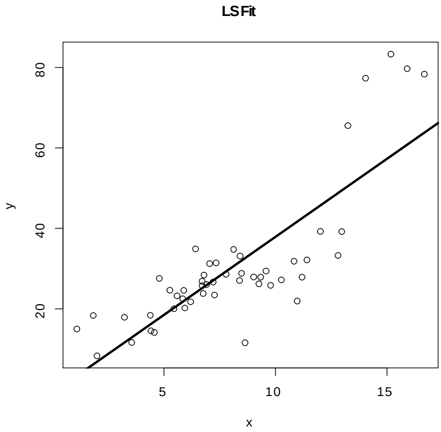
Target Population



Training Sample



Least Squares Fit



A Very Simple Robust Alternative to LS

- Sorted Residuals

$$
\begin{aligned}
r_i^2\left(b_0, b_1\right) &= \left(y_i - b_0 - b_1 x_i\right)^2 \\
r_{(1)}^2\left(b_0, b_1\right) &\leq r_{(2)}^2\left(b_0, b_1\right) \leq \cdots \leq r_{(50)}^2\left(b_0, b_1\right)
\end{aligned}
$$

A Very Simple Robust Alternative to LS

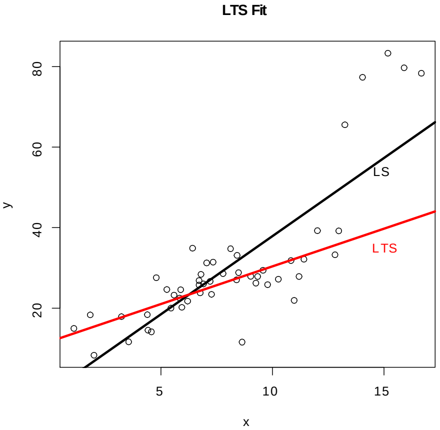
- Sorted Residuals

$$r_i^2(b_0, b_1) = (y_i - b_0 - b_1 x_i)^2$$

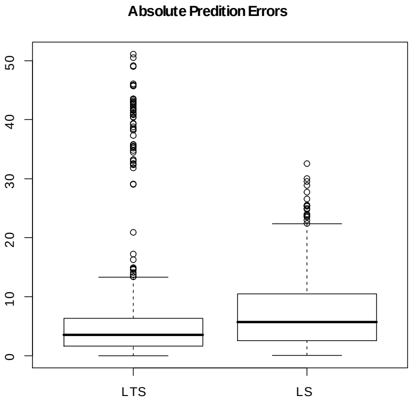
$$r_{(1)}^2(b_0, b_1) \leq r_{(2)}^2(b_0, b_1) \leq \cdots \leq r_{(50)}^2(b_0, b_1)$$

- Least Trimmed Squares (LTS)

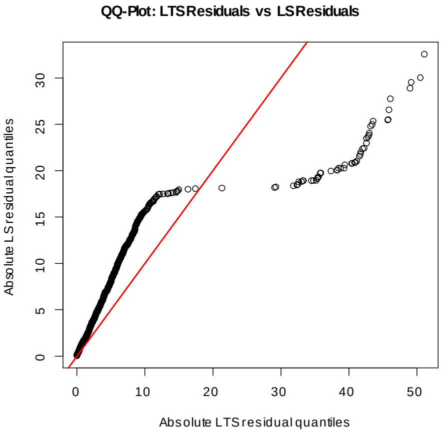
$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_1, b_2} \sum_{i=1}^{30} r_{(i)}^2(b_0, b_1)$$


LTS Fit
y
80
60
40
20
5
10
15
x
LS
LTS

Absolute Prediction Errors



LTS-Quantiles Vs LS-Quantiles



Summary for the Absolute Prediction Errors

	LTS	LS
Min	0.001	0.046
First Quartile	1.634	2.539
Median	3.537	5.703
Third Quartile	5.739	7.100
Max	51.10	32.55