Robustness: Main ideas

Ruben Zamar
Dept. of Stats, UBC

September 9, 2015

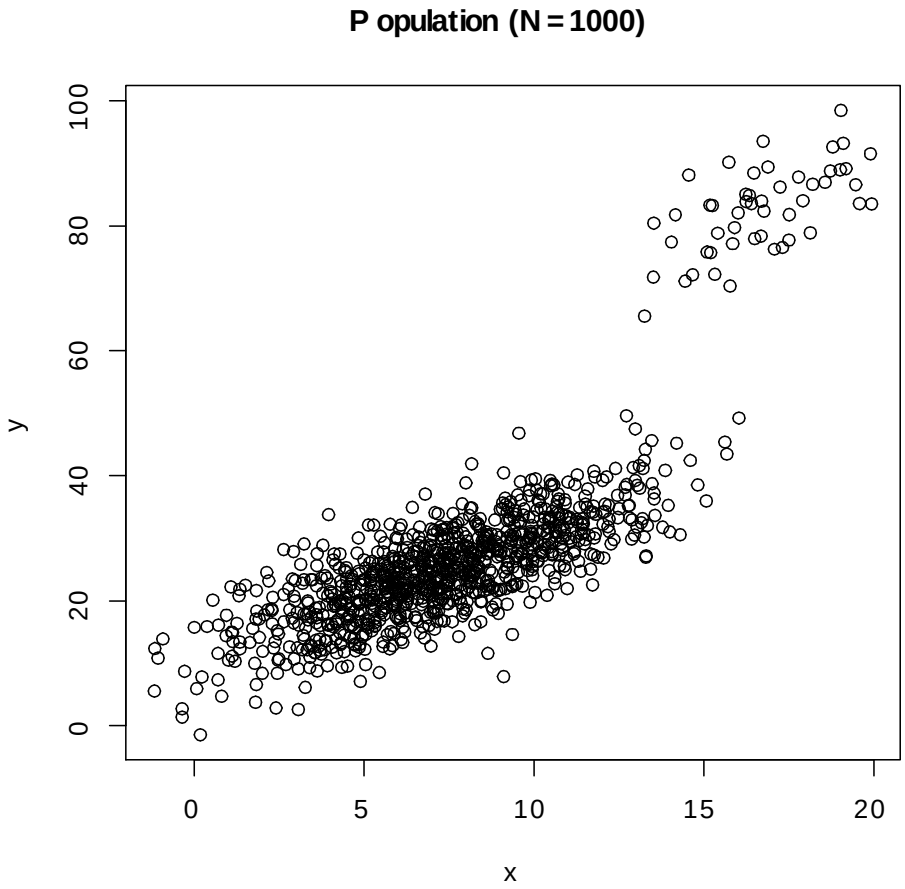
Fitting well all the data

versus

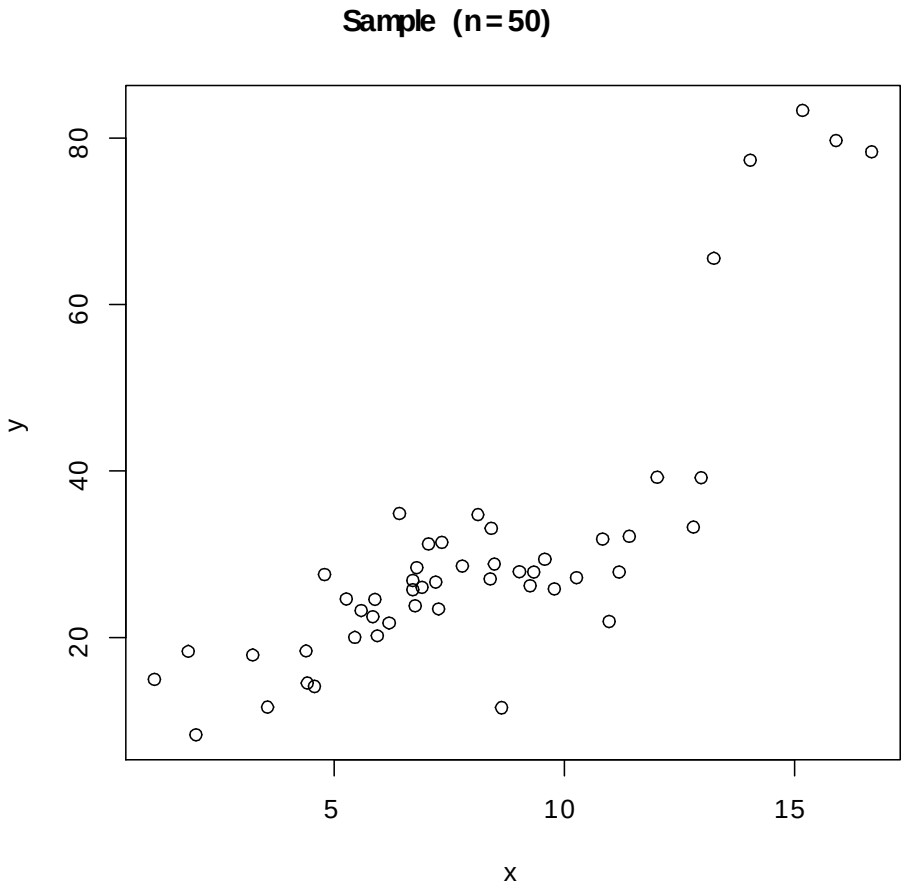
Fitting well most of the data

A NUMERICAL EXAMPLE

Target Population



Training Sample



A Prediction Model

- Linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, ..., 50$$

A Prediction Model

- Linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, 2, ..., 50$$

- Regression residuals

$$r_i\left(b_0, b_1\right) = y_i - b_0 - b_1 x_i$$

Least Squares (LS)

- Least Squares (LS)

$$\left(\widehat{\alpha}^{LS}, \widehat{\beta}^{LS}\right) = \arg\min_{b_1, b_2} \sum_{i=1}^{n=50} r_i^2 \left(b_0, b_1\right)$$

Least Squares (LS)

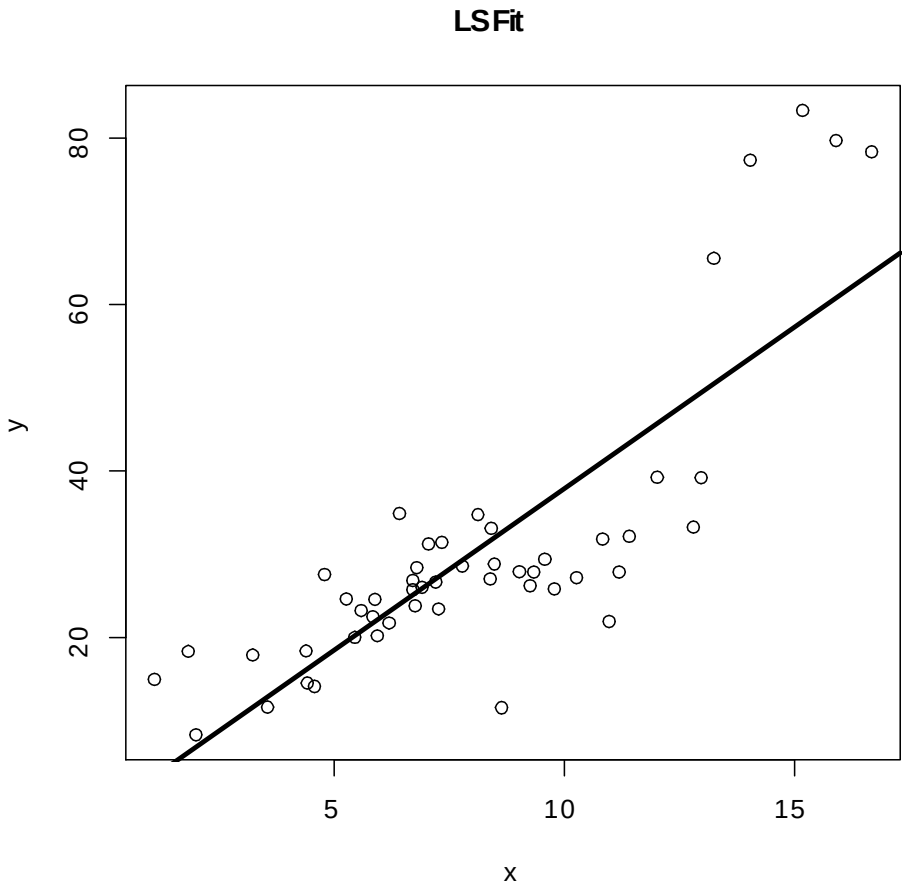
- Least Squares (LS)

$$\left(\widehat{\alpha}^{LS}, \widehat{\beta}^{LS}\right) = \arg\min_{b_1, b_2} \sum_{i=1}^{n=50} r_i^2 \left(b_0, b_1\right)$$

- For our example, using the function lm in R, we obtain:

$$\widehat{\alpha}^{LS} = -0.9237 \quad \text{and} \quad \widehat{\beta}^{LS} = 3.8808$$

Least Squares Fit



A Robust Alternative

- Sorted squared residuals

$$r^2_{(1)}(b_0, b_1) \leq r^2_{(2)}(b_0, b_1) \leq \cdots \leq r^2_{(n)}(b_0, b_1)$$

A Robust Alternative

- Sorted squared residuals

$$r^2_{(1)}(b_0, b_1) \quad \leq \quad r^2_{(2)}(b_0, b_1) \leq \cdots \leq r^2_{(n)}(b_0, b_1)$$

- Least Trimmed Squares (LTS)

$$\left(\widehat{\alpha}^{LTS}, \widehat{\beta}^{LTS}\right) = \arg\min_{b_1, b_2} \sum_{i=1}^{n(1-\alpha)} r^2_{(i)}(b_0, b_1)$$

A Robust Alternative (continued)

- For our example we set $\alpha = 0.4$, so

$$\left(\widehat{\alpha}^{LTS}, \widehat{\beta}^{LTS}\right) = \arg\min_{b_1, b_2} \sum_{i=1}^{30} r^2_{(i)}\left(b_0, b_1\right)$$

A Robust Alternative (continued)

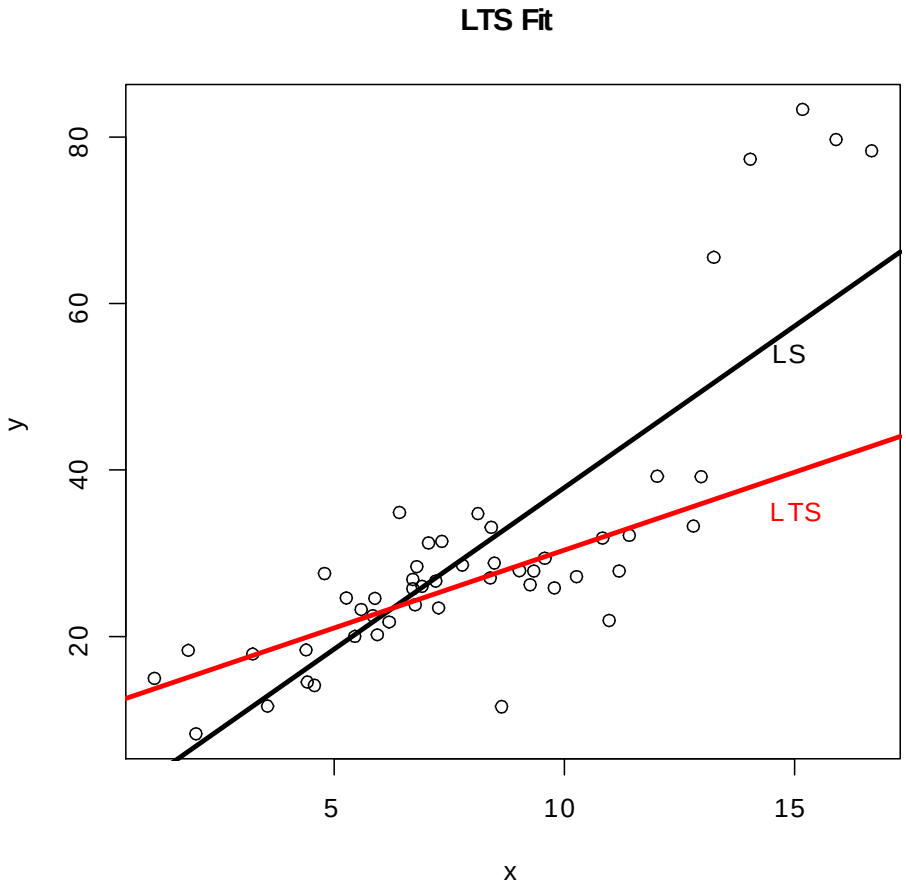
- For our example we set $\alpha = 0.4$, so

$$\left(\widehat{\alpha}^{LTS}, \widehat{\beta}^{LTS}\right) = \arg\min_{b_1, b_2} \sum_{i=1}^{30} r^2_{(i)}\left(b_0, b_1\right)$$

- Using the function ltsReg in the package robust in R, we obtain:

$$\widehat{\alpha}^{LTS} = 11.662 \quad \text{and} \quad \widehat{\beta}^{LTS} = 1.875$$

LTS and LS Fits



Out of Sample Prediction

- Predict the value of y_i for a new case with covariate value x_i

Out of Sample Prediction

- Predict the value of y_i for a new case with covariate value x_i
- Use the prediction model and the estimated parameters $\left(\widehat{\alpha}, \widehat{\beta}\right)$:

$$\hat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$$

Out of Sample Prediction (continued)

- Prediction error

$$\begin{aligned} e\left(x_i, \widehat{\alpha}, \widehat{\beta}\right) &= y_i - \hat{y}_i \\ &= y_i - \widehat{\alpha} - \widehat{\beta} x_i \end{aligned}$$

Out of Sample Prediction (continued)

- Prediction error

$$\begin{aligned} e\left(x_i, \widehat{\alpha}, \widehat{\beta}\right) &= y_i - \hat{y}_i \\ &= y_i - \widehat{\alpha} - \widehat{\beta} x_i \end{aligned}$$

- Absolute prediction error (APE)

$$APE = |y_i - \hat{y}_i| = \left| e\left(x_i, \widehat{\alpha}, \widehat{\beta}\right) \right|$$

Distribution of APE

- We consider the distribution of APE **for the** $N - n$ **out-of-sample cases**

Distribution of APE

- We consider the distribution of APE **for the** $N - n$ **out-of-sample cases**

- In our example

$$N - n = 950$$

Distribution of APE

- We consider the distribution of APE **for the** $N - n$ **out-of-sample cases**
- In our example

$$N - n = 950$$

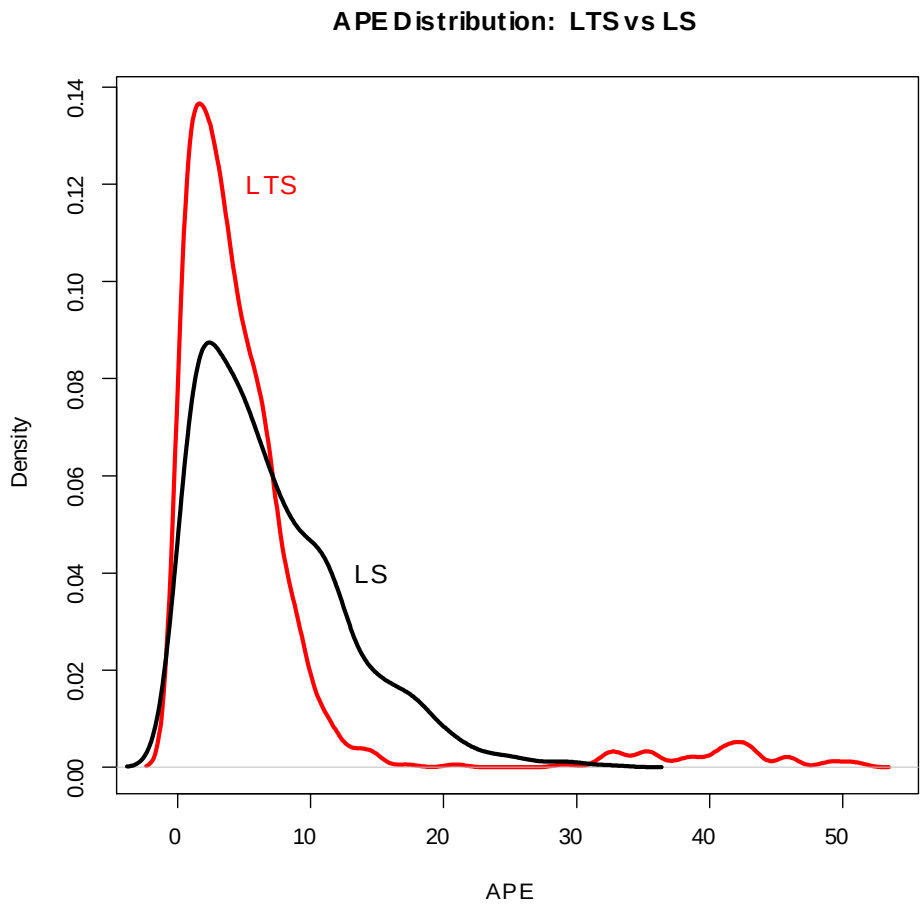
- Use numerical summaries and plots to compare APE's based on LS and LTS

APE Quantiles (LTS vs LS)

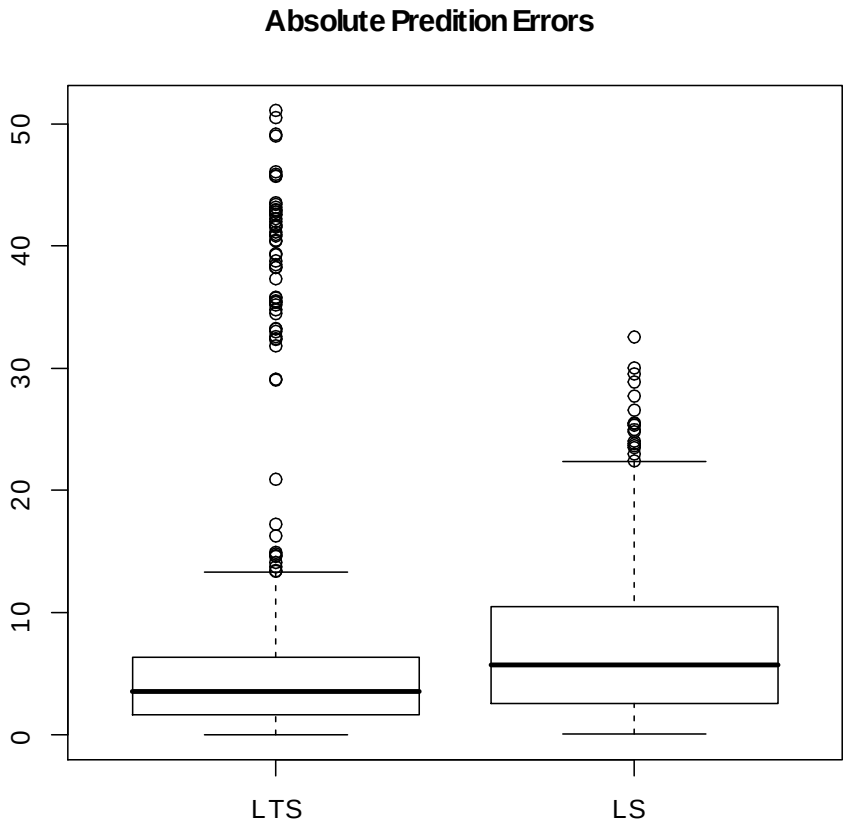
Comparison of α-Quantiles

δ	$q_{\mathbf{LTS}}(\delta)$	$q_{\mathbf{LS}}(\delta)$
0.01	**0.05**	**0.22**
0.05	**0.32**	**0.55**
0.25	**1.69**	**2.58**
0.50	**3.55**	**5.72**
0.75	**6.31**	**10.41**
0.95	**14.87**	**17.98**
0.99	**43.21**	**24.42**

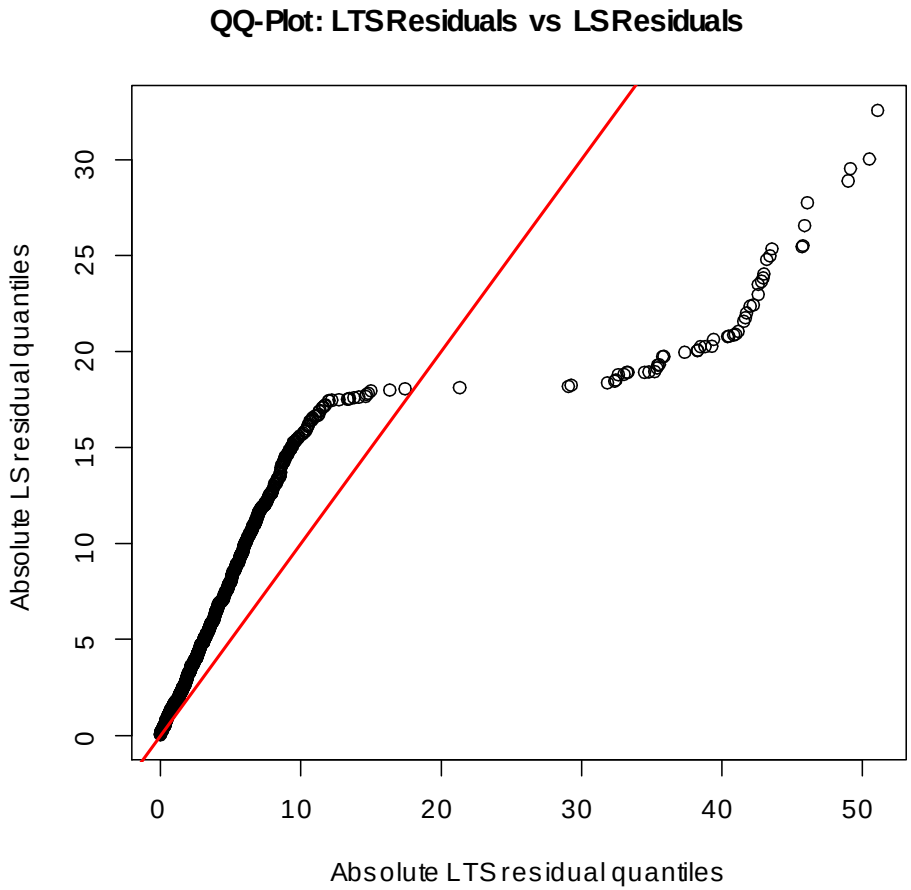
APE Densities (LTS vs LS)



APE Boxplots (LTS vs LS)



APE q-q Plots (LTS vs LS)



Classical methods results

versus

Robust methods results

J. W. Tukey (1979) said:

"... just which robust/resistant methods you use is not important – **what is important is that you use some...**"

John Tukey said:

> " ... It is perfectly proper to use both classical and robust/resistant methods routinely, and only **worry when they differ enough to matter.**"

John Tukey said:

" ... It is perfectly proper to use both classical and robust/resistant methods routinely, and only **worry when they differ enough to matter."**

"...when they differ, you should think hard."

A REAL DATA EXAMPLE

Stack Loss Data

- Data first published by Brownlee (1965)

Stack Loss Data

- Data first published by Brownlee (1965)
- Available in R (dataset name = stackloss)

Stack Loss Data

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

Stack Loss Data

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

Stack Loss Data

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

Stack Loss Data

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

The Variables

Input Variables

The Variables

Input Variables

Air Flow	The rate flow of cooling air

The Variables

Input Variables

Air Flow	The rate flow of cooling air
Water Temperature	Inlet cooling water temperature

The Variables

Input Variables

Air Flow	The rate flow of cooling air
Water Temperature	Inlet cooling water temperature
Acid Concentration	Concentration of acid

The Variables

Input Variables

Air Flow	The rate flow of cooling air
Water Temperature	Inlet cooling water temperature
Acid Concentration	Concentration of acid

Output Variable

The Variables

Input Variables

Air Flow	The rate flow of cooling air
Water Temperature	Inlet cooling water temperature
Acid Concentration	Concentration of acid

Output Variable

Stack Loss	Inverse measure for the overall efficiency of the plant

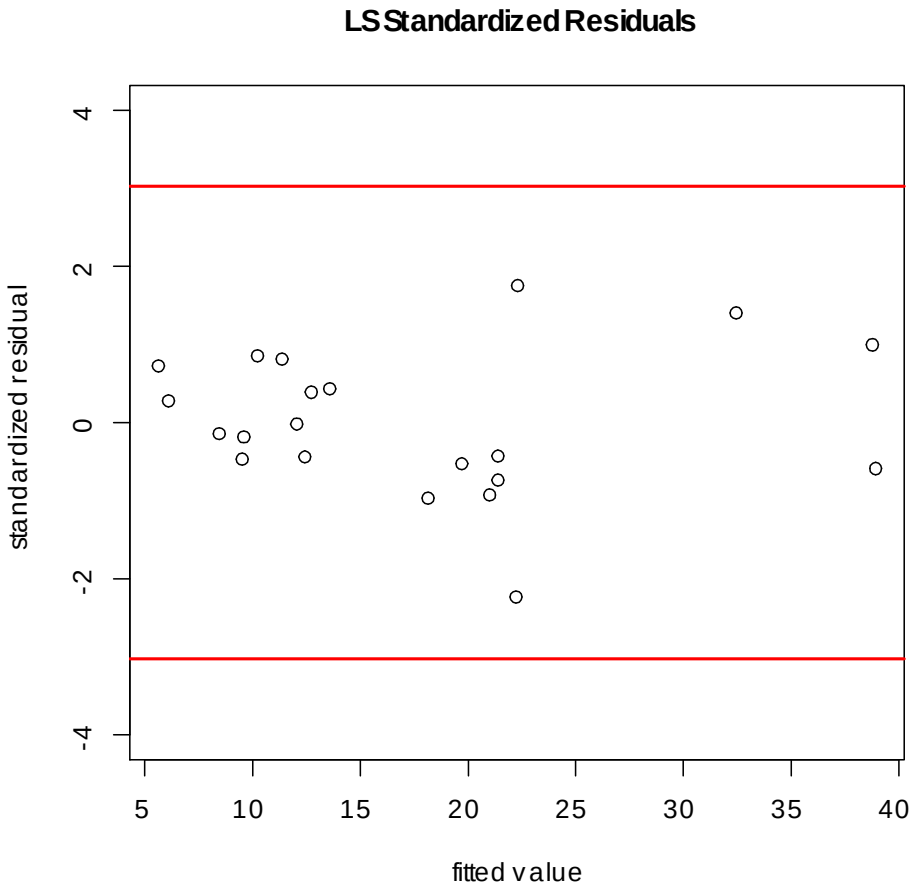
Classical and Robust Linear Models

Regression Coefficient Estimate	**LS**	**MM**
Intercept	-39.92	-37.65
Air Flow	0.716	0.798
Water Temperature	**1.300**	**0.577**
Acid Concentration	**-0.152**	**-0.067**
Residual SE	**3.243**	**1.837**

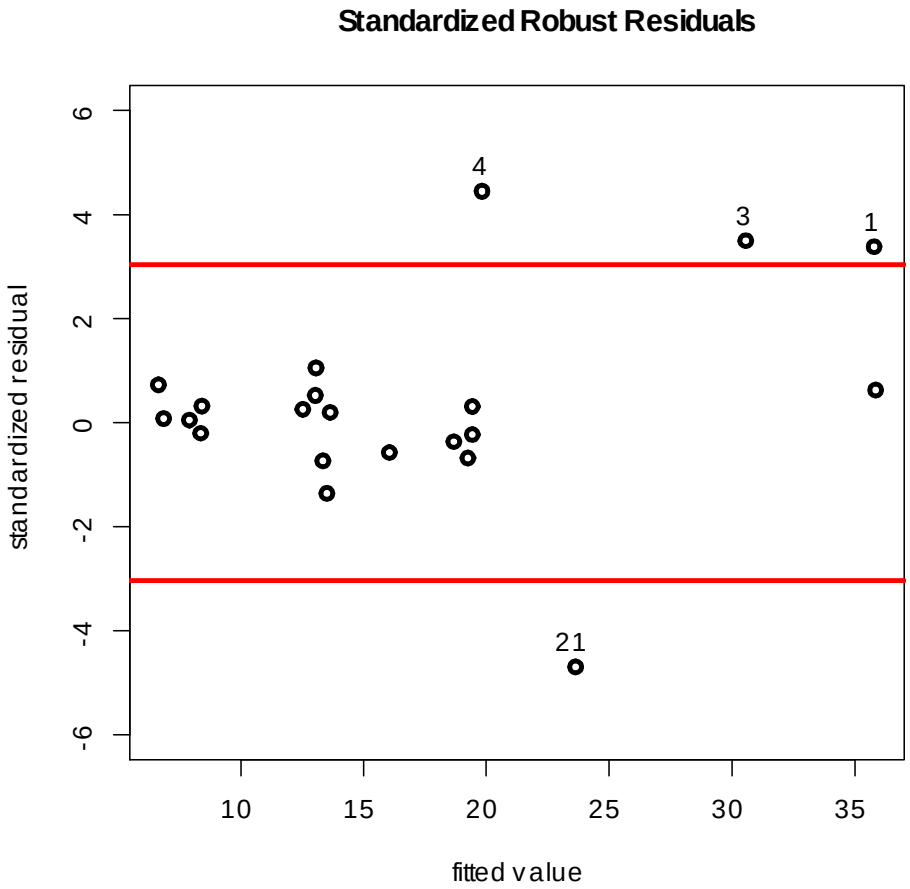
Big differences!

"We must think hard..."

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 correspond to cases 1, 3, 4 and 21** directly uncovered by the robust fit.

ANOTHER REAL DATA EXAMPLE

Boston Housing Dataset

- Socioeconomic-Demographic data on 506 urban districts in the Boston area, USA

Boston Housing Dataset

- Socioeconomic-Demographic data on 506 urban districts in the Boston area, USA
- Data downloaded from the R package "spdep" (dataset name = boston)

Boston Housing Dataset

- Socioeconomic-Demographic data on 506 urban districts in the Boston area, USA
- Data downloaded from the R package "spdep" (dataset name = boston)
- Number of Variables = 12

Boston Housing Dataset

- Socioeconomic-Demographic data on 506 urban districts in the Boston area, USA
- Data downloaded from the R package "spdep" (dataset name = boston)
- Number of Variables = 12
- Number of Cases = 506

Boston Housing Data (Continued)

Variable	Brief Description
CMEDV	median values of owner-occupied housing in USD 1000
CRIM	per capita crime
INDUS	proportions of non-retail business acres per town
NOX	nitric oxides concentration (parts per 10 million) per town
RM	average numbers of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	an index of accessibility to radial highways per town
TAX	full-value property-tax rate per USD 10,000 per town
PTRATIO	pupil-teacher ratios per town (constant for all Boston tracts)
B	1000*(Bk - 0.63)^2 where Bk is the proportion of blacks
LSTAT	percentage values of lower status population

Exploratory Data Analysis

- Compute estimates of multivariate location $\hat{\mu}$ and scatter matrix $\hat{\Sigma}$

Exploratory Data Analysis

- Compute estimates of multivariate location $\hat{\mu}$ and scatter matrix $\hat{\Sigma}$
- Compute (squared) Mahalanobis distances (MD) for each data point:

$$d_i^2 \quad = \quad (\mathbf{x}_i - \hat{\mu})' \hat{\Sigma}^{-1} (\mathbf{x}_i - \hat{\mu})$$

Exploratory Data Analysis

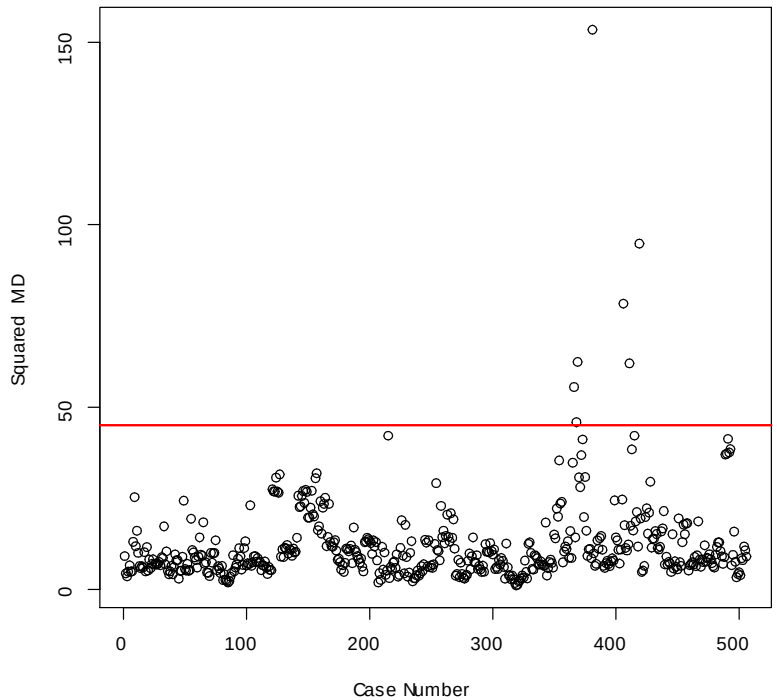
- Compute estimates of multivariate location $\hat{\boldsymbol{\mu}}$ and scatter matrix $\hat{\Sigma}$
- Compute (squared) Mahalanobis distances (MD) for each data point:

$$d_i^2 \quad = \quad (\mathbf{x}_i - \hat{\boldsymbol{\mu}})' \, \hat{\Sigma}^{-1} \, (\mathbf{x}_i - \hat{\boldsymbol{\mu}})$$

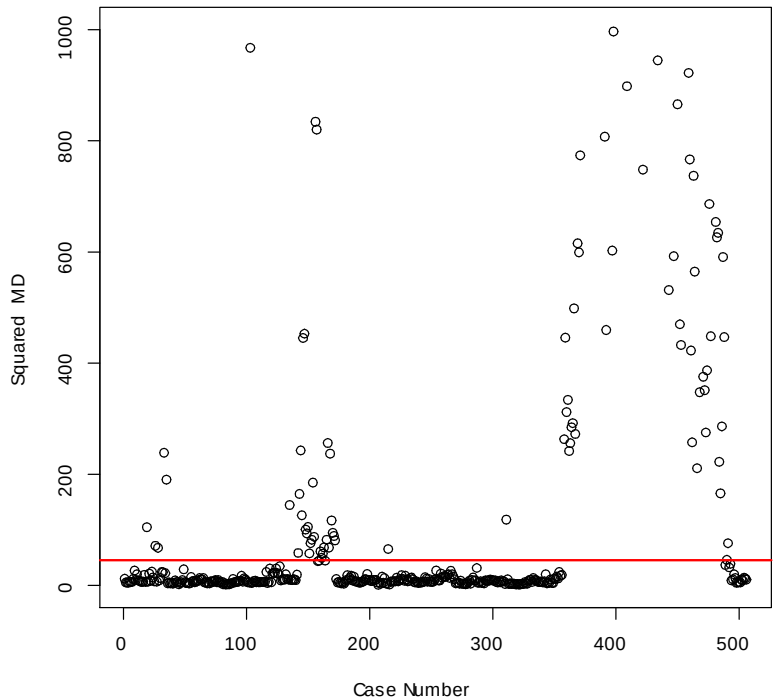
- Threshold: Compare the MD's with a quantile of a $\chi^2_{(p)}$

$$q \quad = \quad F^{-1}_{\chi^2_{(p)}} \, (0.99999) = 45.07615$$

Mahalanobis Distances: Using the Sample Mean and Covariance Matrix



Mahalanobis Distances: Using Robust Location-Scatter MVES-Estimates



Think Hard...

- Using a robust multivariate-location estimate (called MVE-S) we find 171 outliers
- Using classical sample mean and covariance matrix we find 7 outliers
- Big difference $\Rightarrow$ Must think hard!
- We will be back to this example

NOISE AND DATA QUALITY

Statistical Models

- Many data can be modeled as follows:

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

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

- Atypical noise comes from

OUTLIERS AND GROSS ERRORS

MEASUREMENTS OF UNEVEN QUALITY (mixture)

DATA CONTAMINATION (mixture)

MISSING DATA (declared or unsuspected)

DUPLICATIONS, ETC

Statistical Tasks (simplified view)

- **Filter noise (both typical and atypical noise)**

Statistical Tasks (simplified view)

- **Filter noise (both typical and atypical noise)**
- **Extract the signal (point estimation)**

Statistical Tasks (simplified view)

- **Filter noise (both typical and atypical noise)**
- **Extract the signal (point estimation)**
- **Measure the strength of the noise (statistical inference)**

Statistical Tasks (simplified view)

- **Filter noise (both typical and atypical noise)**
- **Extract the signal (point estimation)**
- **Measure the strength of the noise (statistical inference)**
- **Assess uncertainty in the estimates (statistical inference)**

Statistical Tasks (simplified view)

- **Filter noise (both typical and atypical noise)**
- **Extract the signal (point estimation)**
- **Measure the strength of the noise (statistical inference)**
- **Assess uncertainty in the estimates (statistical inference)**
- **Predict future data (prediction)**

Output from Statistical Analysis

- **Point Estimates** $\widehat{\theta}$

Output from Statistical Analysis

- **Point Estimates**

$$\widehat{\theta}$$

- **Confidence Regions**

$$Cov\left(\widehat{\theta}\right), \quad \text{Confidence Region for } \theta$$

Output from Statistical Analysis

- **Point Estimates**

$$\widehat{\theta}$$

- **Confidence Regions**

$$Cov\left(\widehat{\theta}\right), \quad \text{Confidence Region for } \theta$$

- **Prediction / Interpolation**

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

Only Typical Noise

- Typically

$$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$$

Only Typical Noise

- Typically

$$\widehat{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\theta}$$

- and

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

Only Typical Noise

- Typically

$$\widehat{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\theta}$$

- and

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

- Better when $C_{\widehat{\boldsymbol{\theta}}}$ is small $\Longrightarrow$ **use efficient procedures**

Only Typical Noise

- Typically

$$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$$

- and

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

- Better when $C_{\widehat{\boldsymbol{\theta}}}$ is small $\Longrightarrow$ **use efficient procedures**
- Much attention has been given to the problem of minimizing $C_{\widehat{\boldsymbol{\theta}}}$

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

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

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\widehat{\theta} \to \Delta, \qquad \Delta \neq \theta$$

- The difference between Δ and θ is called "contamination bias" (cb)

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\widehat{\theta} \to \Delta, \qquad \Delta \neq \theta$$

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

The Effect of Atypical Noise

- Atypical noise tends to produce asymptotic bias
- That is

$$\widehat{\theta} \to \Delta, \qquad \Delta \neq \theta$$

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