Robust statistics

Reference material

- Hampel,F.R., Ronchetti,E.M., Rousseeuw, P.J., Stahel, W.A. Robust Statistics: the Approach based on Influence Functions.Wiley Series in Probability and Mathematical Statistics.,1986.

Motivation

- Real world data often contains outliers or extreme values

- Most methods discussed so far on inferring models or parameter values from data can be adversely affected by outliers
  - Example: estimates of mean and variance from data
  - Estimation of linear regression parameters

- Robust statistics attempts to fit models that are largely unaffected by outliers, and fit based on “majority” of normal data.

- Robust fits enable better detection of outliers, as values that deviate from the fitted model

Assumptions

- We assume that the majority of the observations satisfy a **parametric** model and we want to estimate the parameters of this model.

  E.g. $x_i \sim N(\mu, \sigma^2)$
  $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$
  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$

- Moreover, we assume that some of the observations might not satisfy this model.

- We do NOT *model* the outlier generating process.

- We do NOT know the *proportion* of outliers in advance.

Example

The classical methods for estimating the parameters of the model may be affected by outliers.

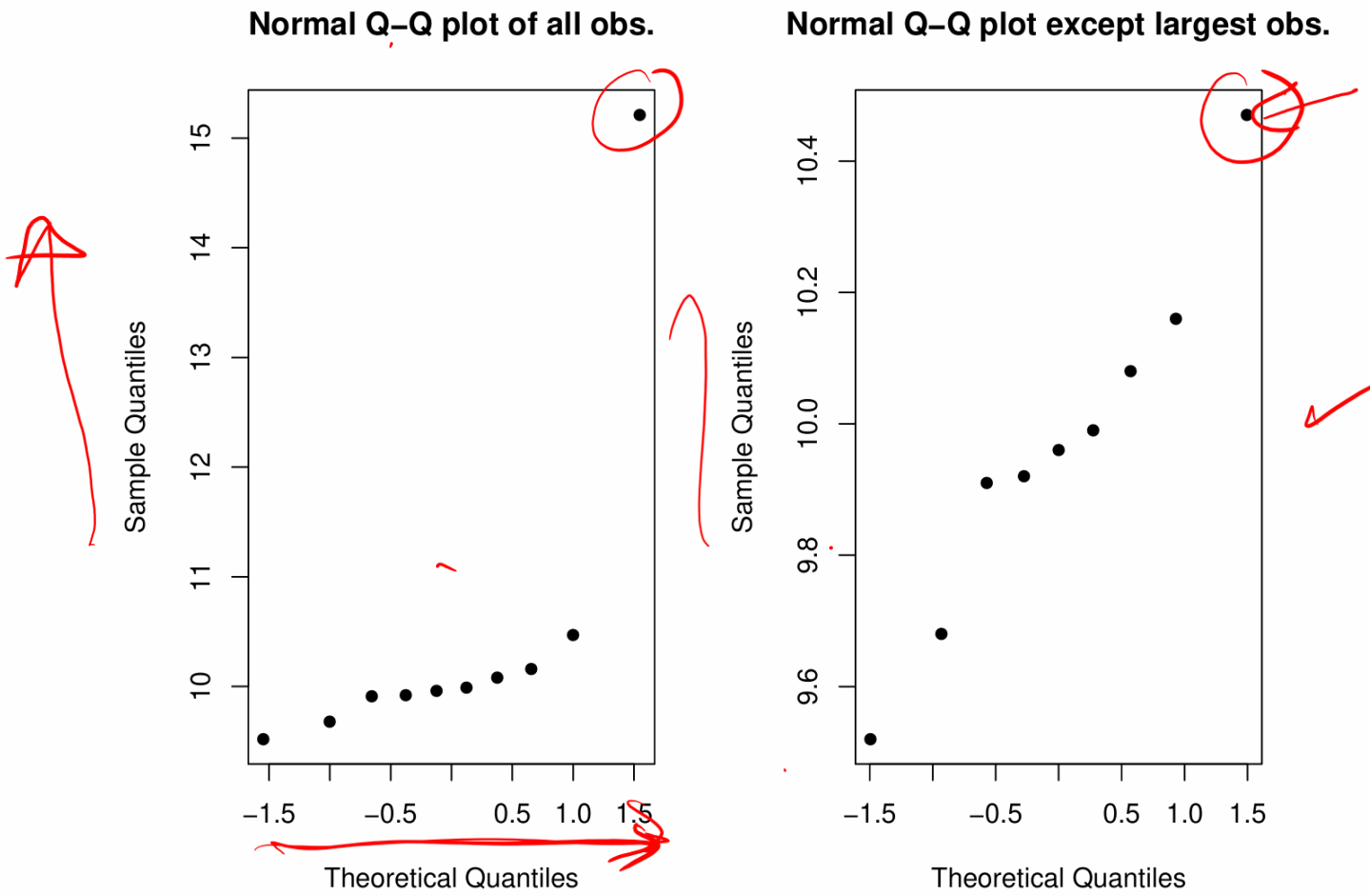
Example. Location-scale model: $x_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

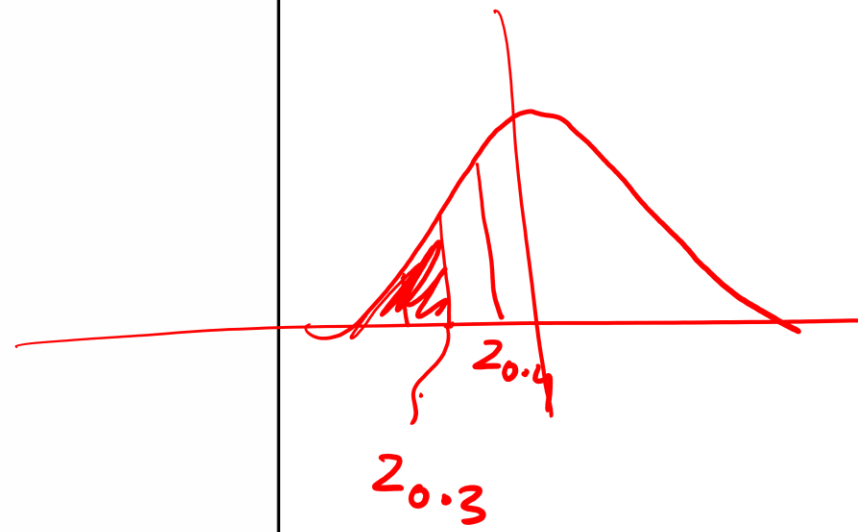
Data: $X_n = \{x_1, \ldots, x_{10}\}$ are the natural logarithms of the annual incomes (in US dollars) of 10 people.

9.52	9.68	10.16	9.96	10.08
9.99	10.47	9.91	9.92	15.21

Example

The income of person 10 is much larger than the other values.
Normality cannot be rejected for the remaining ('regular') observations:





Classical versus robust estimators

Location:

Classical estimator: arithmetic mean

$$\hat{\mu} = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Robust estimator: sample median

$$\hat{\mu} = \text{med}(X_n) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if} \quad n \text{ is odd} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if} \quad n \text{ is even} \end{cases}$$

with $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ the ordered observations.

Classical versus robust estimators

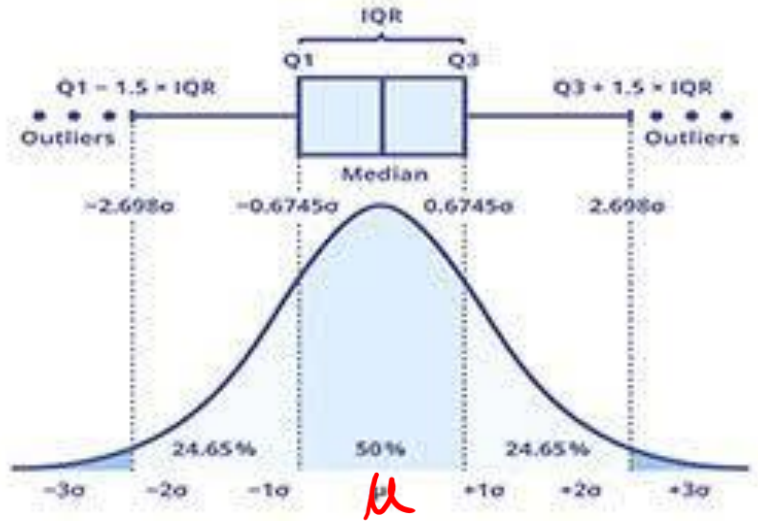
Scale:

Classical estimator: sample standard deviation

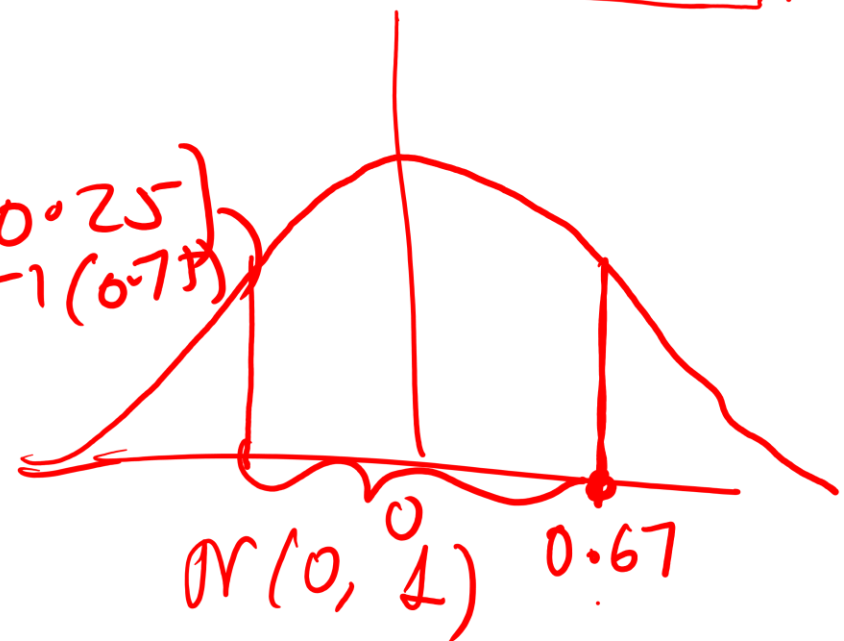
$$\hat{\sigma} = \text{Stdev}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

Robust estimator: interquartile range

$$\hat{\sigma} = \text{IQRN}(X_n) = \frac{1}{2\Phi^{-1}(0.75)}(x_{(n-[n/4]+1)} - x_{([n/4])})$$



For Gaussian $N(\mu, \sigma^2)$:
IQR can be shown to be $2\Phi^{-1}(0.75)\sigma$



$\Phi^{-1}(0.75) - \Phi^{-1}(0.25)$
$= \Phi^{-1}(0.75) - (-\Phi^{-1}(0.75))$
$= 2\Phi^{-1}(0.75)$

Classical versus robust estimators

For the data of the example we obtain:

	the 9 regular observations	all 10 observations
$\bar{x}_n$	9.97	10.49
med	9.96	9.98
Stdev_n	0.27	1.68
IQRN	0.13	0.17

1. The classical estimators are highly influenced by the outlier
2. The robust estimators are less influenced by the outlier
3. The robust estimate computed from all observations is comparable with the classical estimate applied to the non-outlying data.

Classical versus robust estimators

Robustness: being less influenced by outliers

Efficiency: being precise at uncontaminated data

Robust estimators aim to combine high robustness with high efficiency

Outliers

An outlier is an observation that deviates from the fit suggested by the majority of the observations.

The usual standardized values (z-scores, standardized residuals) are:

$$r_i = \frac{x_i - \bar{x}_n}{\text{Stdev}_n}$$

$r_i \sim N(0,1)$

Classical rule: if $|r_i| > 3$, then observation x_i is flagged as an outlier.

Here: $|r_{10}| = 2.8 \rightarrow$?

Outlier detection based on robust estimates:

$$r_i = \frac{x_i - \text{med}(X_n)}{\text{IQRN}(X_n)}$$

Here: $|r_{10}| = 31.0 \rightarrow$ very pronounced outlier!

MASKING is when actual outliers are not detected.

SWAMPING is when regular observations are flagged as outliers.

Characterizing robustness

- Breakdown value
- Sensitivity curve
- Influence function

Breakdown value

Breakdown value (breakdown point) of a location estimator

A data set with n observations is given. If the estimator stays in a fixed *bounded* set even if we replace any $m-1$ of the observations by any outliers, and this is no longer true for replacing any m observations by outliers, then we say that:

the breakdown value of the estimator at that data set is m/n

Notation:

$$\varepsilon_n^*(T_n, X_n) = m/n$$

Typically the breakdown value does not depend much on the data set. Often it is a fixed constant as long as the original data set satisfies some weak condition, such as the absence of ties.

Breakdown value

Example: $X_n = \{x_1, \ldots, x_n\}$ univariate data, $T_n(X_n) = \text{med}(X_n)$.

Assume n odd, then $T_n = x_{((n+1)/2)}$.

- Replace $\frac{n-1}{2}$ observations by any value, yielding a set X_n^*
  $\Rightarrow$ $T_n(X_n^*)$ always belongs to $[x_{(1)}, x_{(n)}]$, hence $T_n(X_n^*)$ is bounded.
- Replace $\frac{n+1}{2}$ observations by $+\infty$, then $T_n(X_n^*) = +\infty$.
- More precisely, if we replace $\frac{n+1}{2}$ observations by $x_{(n)} + a$,
  where a is any positive real number, then $T_n(X_n^*) = x_{(n)} + a$.
  Since we can choose a arbitrarily large, $T_n(X_n^*)$ cannot be bounded.

For n odd or even, the (finite-sample) breakdown value ε_n^* of T_n is

$$\varepsilon_n^*(T_n, X_n) = \frac{1}{n}\left[\frac{n+1}{2}\right] \approx 50\% \; .$$

Note that for $n \to \infty$ the finite-sample breakdown value tends to $\varepsilon^* = 50\%$
(which we call the asymptotic breakdown value).
For instance, the arithmetic mean satisfies $\varepsilon_n^*(T_n, X_n) = \frac{1}{n} \to \varepsilon^* = 0\%$.

Sensitivity curve

The **sensitivity curve** measures the effect of a single outlier on the estimator.

Assume we have $n-1$ fixed observations $X_{n-1} = \{x_1, x_2, \ldots, x_{n-1}\}$.
Now let us see what happens if we add an additional observation equal to x, where x can be any real number.

Sensitivity curve

$$\mathrm{SC}(x, T_n, X_{n-1}) = \frac{T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})}{1/n}$$

Example: for the arithmetic mean $T_n = \bar{X}_n$ we find $\mathrm{SC}(x, T_n, X_{n-1}) = x - \bar{x}_{n-1}$.

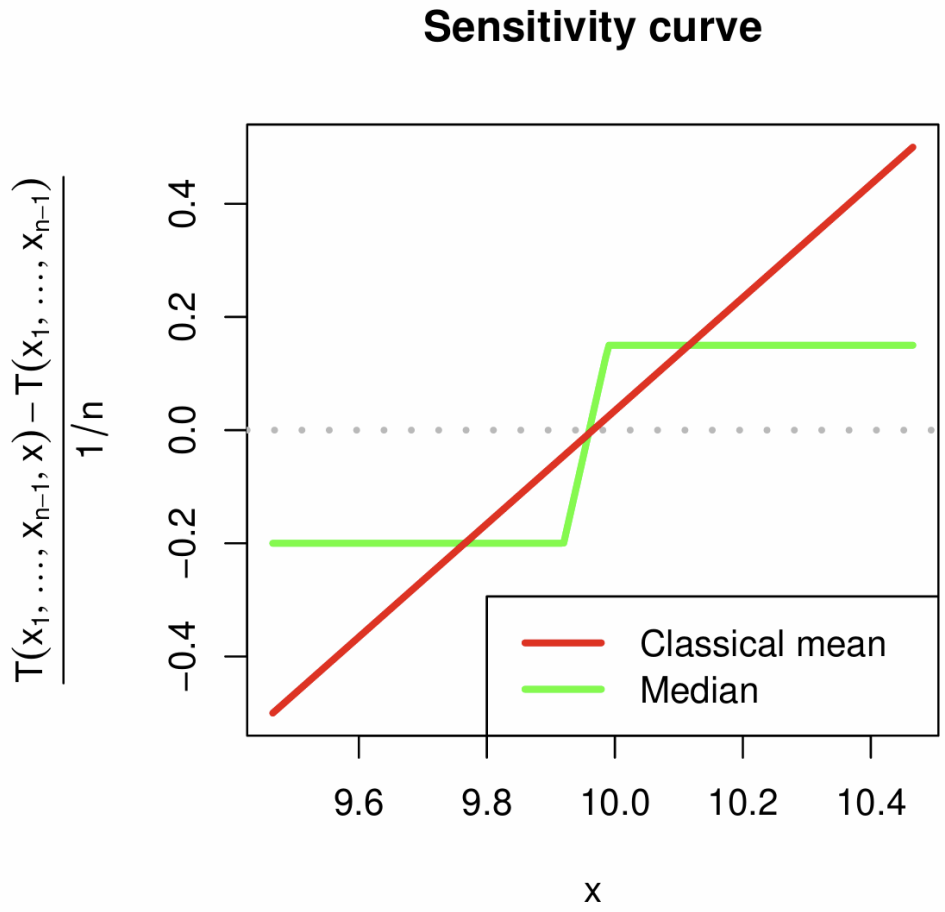
Note that the sensitivity curve depends strongly on the data set X_{n-1} .

Sensitivity curve: example

Annual income data: let X_9 consist of the 9 'regular' observations.

9.52 9.68 9.91 9.92 9.96 9.99 10.08 10.16 10.47

mean 9.97
med 9.96



Influence function

- The influence function is the asymptotic version of the sensitivity curve. It is computed for an estimator T at a certain distribution F, and does not depend on a specific data set.
- For this purpose, the estimator should be written as a function of a distribution F. For example, $T(F) = E_F[X]$ is the functional version of the sample mean, and $T(F) = F^{-1}(0.5)$ is the functional version of the sample median.
- The influence function measures how $T(F)$ changes when contamination is added in x. The contaminated distribution is written as

$$F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$$

for $\varepsilon > 0$, where Δ_x is the distribution that puts all its mass in x.

$\Delta_x(X) = 1$ if $X = x$
$= 0$ otherwise

Influence function

Influence function

$$\mathrm{IF}(x, T, F) = \lim_{\varepsilon \to 0} \frac{T(F_{\varepsilon,x}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon,x}) \mid_{\varepsilon=0}$$

Example: for the arithmetic mean $T(F) = E_F[X]$ at a distribution F with finite first moment:

$$\mathrm{IF}(x, T, F) = \frac{\partial}{\partial \varepsilon} E[(1-\varepsilon)F + \varepsilon \Delta_x] \mid_{\varepsilon=0}$$

$$= \frac{\partial}{\partial \varepsilon} [\varepsilon x + (1-\varepsilon) T(F)] \mid_{\varepsilon=0} = x - T(F)$$

At the standard normal distribution $F = \Phi$ we find $\mathrm{IF}(x, T, \Phi) = x$.

We prefer estimators that have a *bounded* influence function.

Other robust estimates of location

1. **Median**

2. **Trimmed mean**: ignore the m smallest and the m largest observations and just take the average of the observations in between:

$$\hat{\mu}_{TM} = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)}$$

with $m = [(n-1)\alpha]$ and $0 \leqslant \alpha < 0.5$.
For $\alpha = 0$ this is the mean, and for $\alpha \to 0.5$ this becomes the median.

3. **Winsorized mean**: replace the m smallest observations by $x_{(m+1)}$ and the m largest observations by $x_{(n-m)}$. Then take the average:

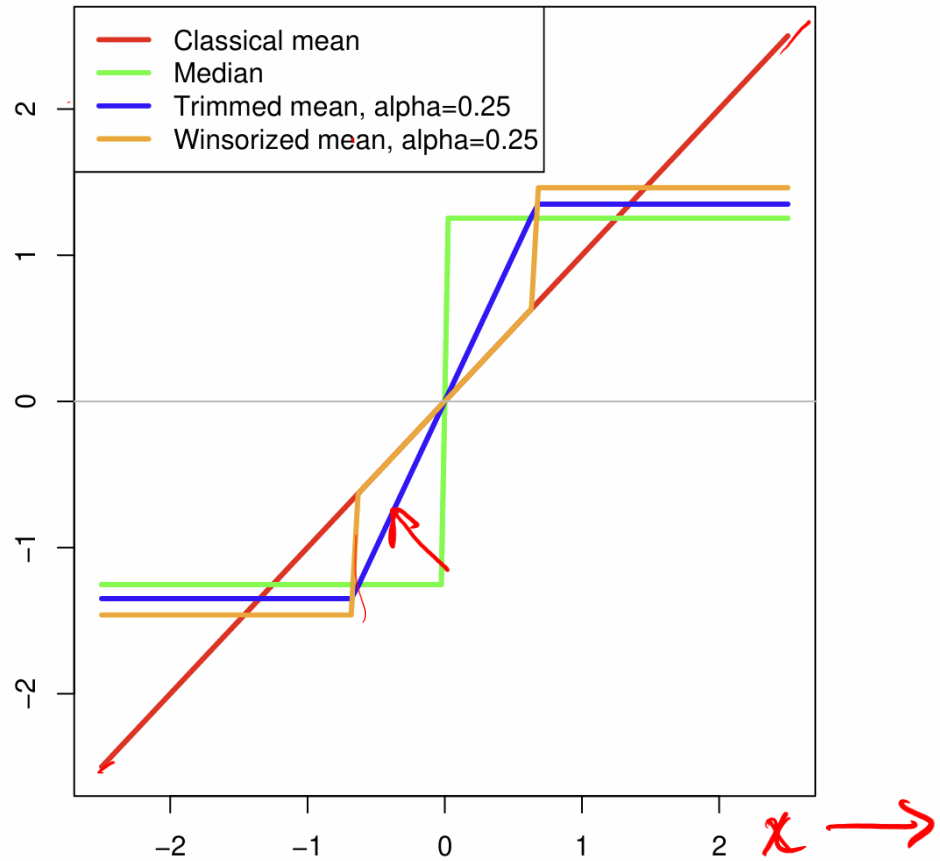
$$\hat{\mu}_{WM} = \frac{1}{n}\left(m x_{(m+1)} + \sum_{i=m+1}^{n-m} x_{(i)} + m x_{(n-m)} \right)$$

Robustness properties

Breakdown value: $\varepsilon_n^*(\text{med}) \to 0.5$; $\varepsilon_n^*(\hat{\mu}_{TM}) = \varepsilon_n^*(\hat{\mu}_{WM}) = (m+1)/n \to \alpha$.

Maxbias: For any ε, the median achieves the smallest maxbias among all location equivariant estimators.

Influence function at the normal model:



The pure scale model

The scale model assumes that the data are i.i.d. according to:

$$\boxed{F_\sigma(x) = F\left(\frac{x}{\sigma}\right)}$$

where $\sigma > 0$ is the unknown scale parameter. As before F is a continuous distribution with density f, but now

$$f_\sigma(x) = F'_\sigma(x) = \frac{1}{\sigma} f(\frac{x}{\sigma}) \; .$$

We say that a scale estimator S is Fisher-consistent at this model iff

$$S(F_\sigma) = \sigma \qquad \text{for all } \sigma > 0 \; .$$

ignore

Robust estimates of scale

Some explicit scale estimators:

1. **Standard deviation (Stdev)** Not robust.
2. **Interquartile range**

$$\mathrm{IQR}(X_n) = x_{(n-[n/4]+1)} - x_{([n/4])}$$

However, at $F_\sigma = N(0, \sigma^2)$ it holds that $IQR(F_\sigma) = 2\Phi^{-1}(0.75)\sigma \neq \sigma$.

Normalized IQR:

$$\mathrm{IQRN}(X_n) = \frac{1}{2\Phi^{-1}(0.75)}\,\mathrm{IQR}(X_n)\ .$$

The constant $1/2\Phi^{-1}(0.75) = 0.7413$ is a *consistency factor*.

When using software, it should be checked whether the consistency factor is included or not!

Explicit scale estimators

Estimators with 50% breakdown value:

3. **Median absolute deviation**

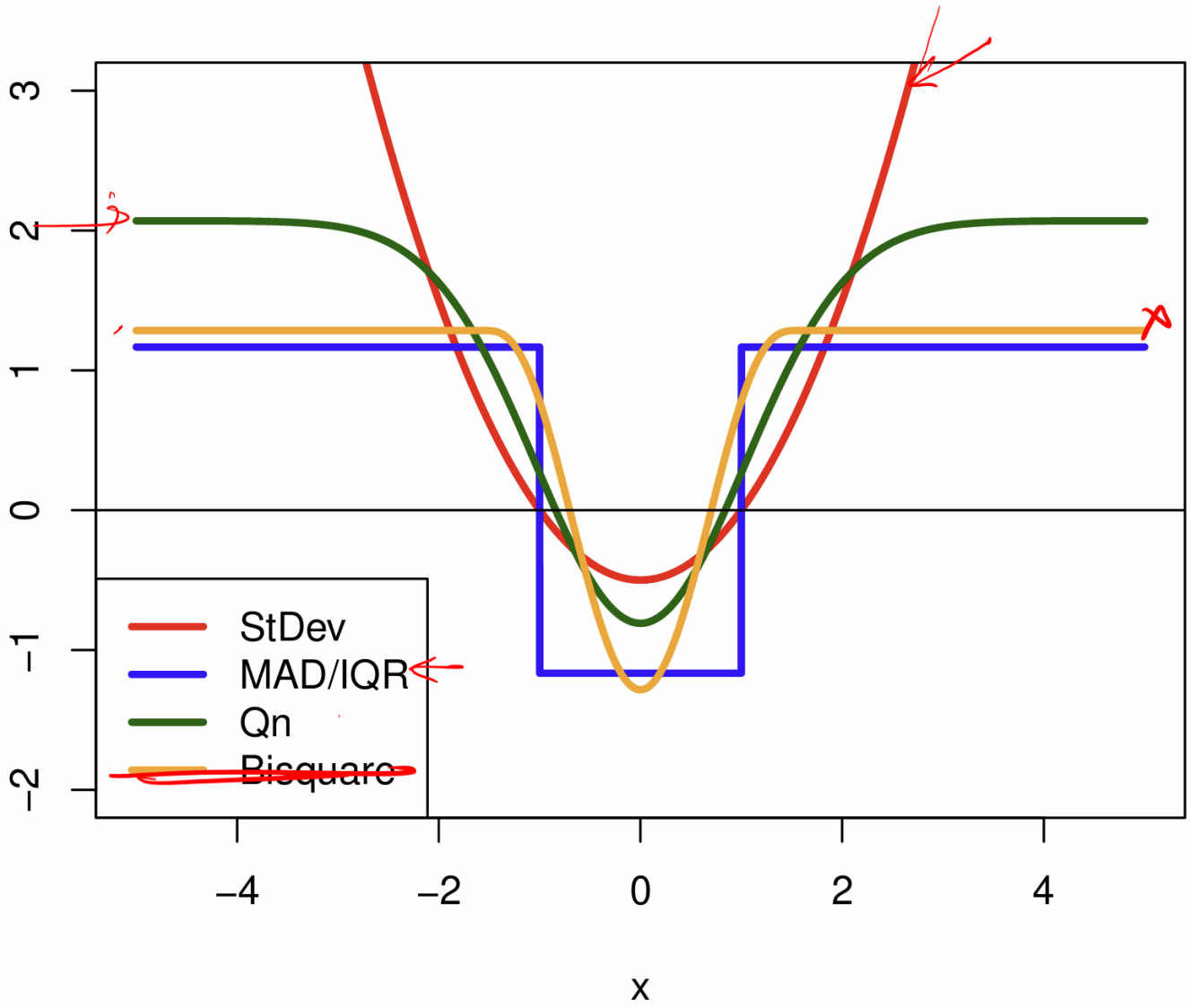
$$\mathrm{MAD}(X_n) = \underset{i}{\mathrm{med}}(|x_i - \mathrm{med}(X_n)|)$$

At any symmetric sample it holds that $\mathrm{IQR} = 2\,\mathrm{MAD}$.

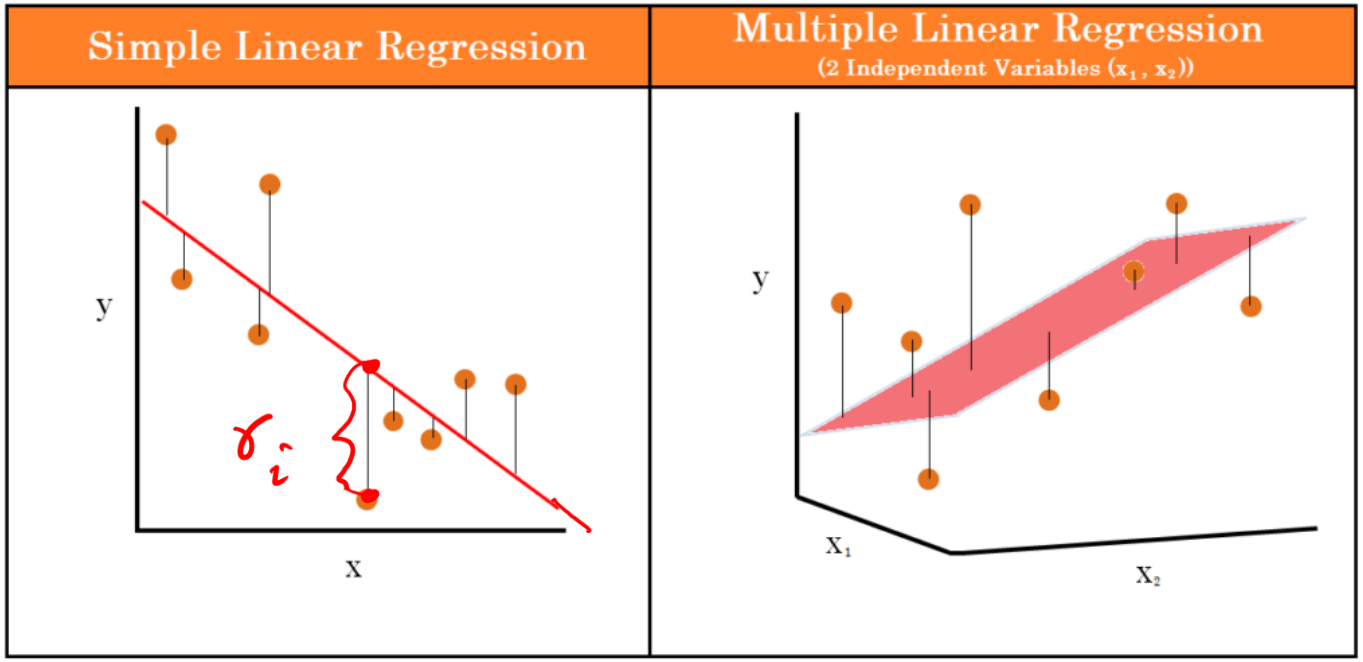
At the normal model we use the normalized version:

$$\mathrm{MADN}(X_n) = \frac{1}{\Phi^{-1}(0.75)}\,\mathrm{MAD}(X_n) = 1.4826\,\mathrm{MAD}(X_n)$$

Influence function of various scale estimators



Robust regression



Reading material

- Primary
  - https://en.wikipedia.org/wiki/Robust_regression
- Additional
  - T.1.1 - Robust Regression Methods | STAT 501

Ordinary least square regression

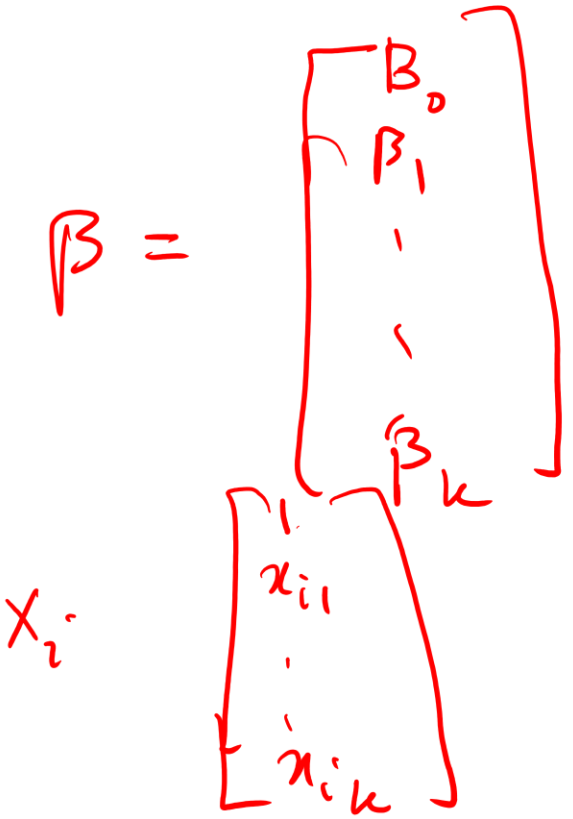
$$f(Y \mid x_1, \ldots, x_k) \sim N(\mu_x, \sigma^2), \quad where \;\; \mu_x = \beta_1 x_1 + \cdots + \beta_k x_k + \beta_0$$

Training data D denoted as

$$\{(x_{i1}, x_{i2}, \ldots x_{ik}, y_i) : i = 1 \ldots n\}$$

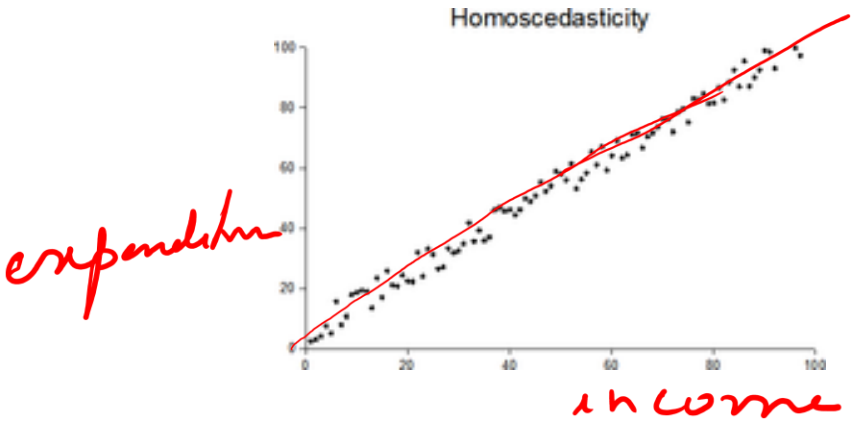
MLE training objective

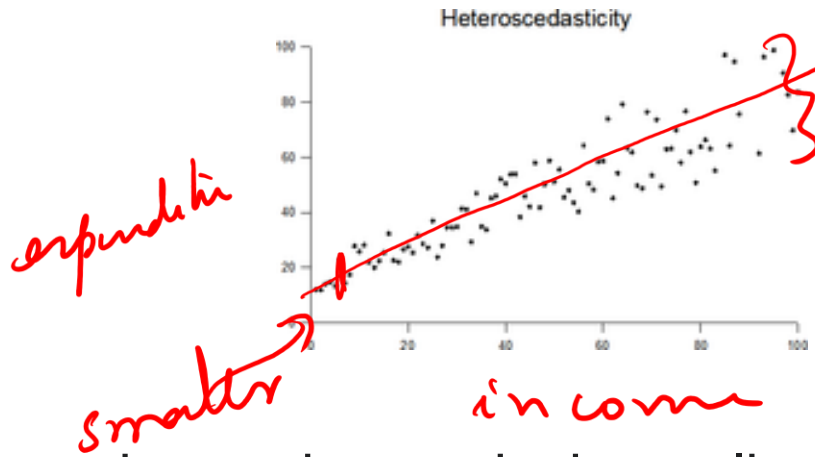
$$\min_{\{\beta\}} \sum_i \frac{(y_i - \beta x_i)^2}{\sigma^2}$$



Limitations

- Assumes same variance for all examples (**homoscedastic**). Real-life data often does not satisfy this assumption (**heteroscedasticity)**
  - For example, the variance of expenditure is often larger for individuals with higher income than for individuals with lower incomes.





- Cannot handle outliers.
  - The least squares predictions are dragged towards the outliers
  - The variance of the estimates is artificially inflated, causing outliers to be masked.
  - In many situations, including some areas of geostatistics and medical statistics, it is precisely the outliers that are of interest.

Robust regression: modify the loss function

normally $\sum_i (y_i - \beta x_i)^2$

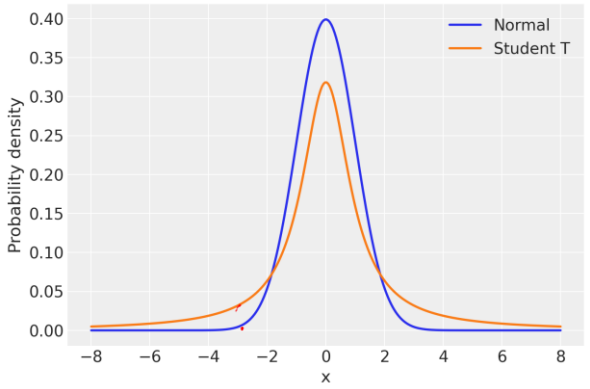
- Replace square loss by least absolute deviation
  - $\min_{\{\beta\}} \sum_i |y_i - \beta x_i|$
  - Solvable using a linear program.
- Least trimmed square: Ignore n-k largest residues during minimization

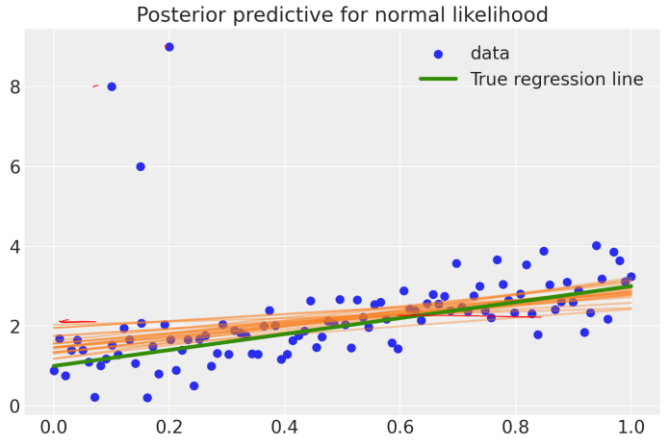
$$\min_{\beta} \sum \text{Bottom-}k \left(|y_1 - \beta x_1| - - - - |y_n - \beta x_n| \right)$$

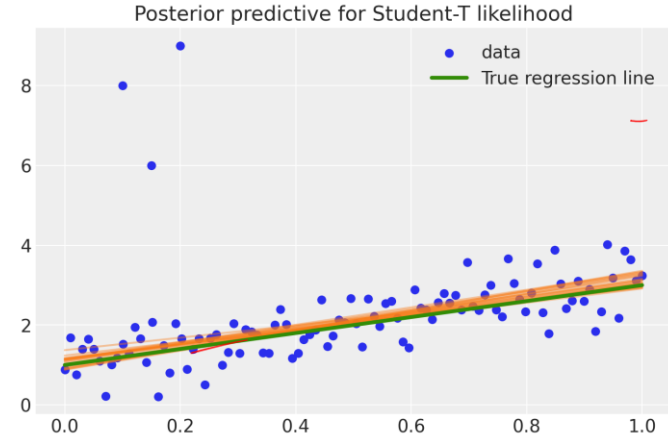
Robust regression: parametric alternative

Replace the normal distribution:

- Choose a heavy-tailed distribution. A *t*-distribution with 4–6 degrees of freedom has been reported to be a good choice in various practical situations.







https://colab.research.google.com/github/pymc-devs/pymc-examples/blob/main/examples/generalized_linear_models/GLM-robust.ipynb#scrollTo=285a756b

Robust regression: continued…

- Choose a mixture of normal and outlier distribution --- majority of observations are from a specified normal distribution, but a small proportion are from a normal distribution with much higher variance.

$$e_i \sim (1-\varepsilon)N(0,\sigma^2) + \varepsilon N(0, c\sigma^2).$$

Usefulness of robust regression

- If number of data points is large, the actual fitted model may not be different with robust methods, but robust estimation of the variance, can lead to better outlier detection.