# Robust regression

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

January 17, 2023

# Outline

# Outlines

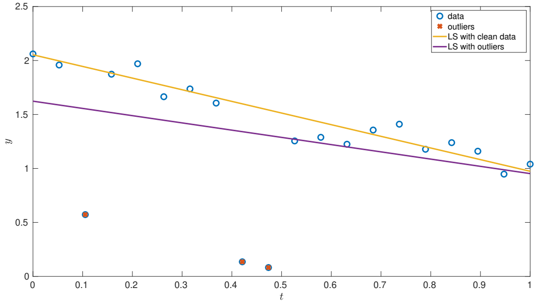
- outliers in measurements: outlying $x$ and outlying $y$
- outlier diagnostics
  - studentized residuals
  - hat matrix
  - Cook's distance
- robust regression
  - weighted least-squares (WLS)
  - M-estimator

setting: $y = X\beta + e$ where $X \in \mathbf{R}^{N \times n}$, $n$: the number of coefficients

# Outliers in measurements

# Outliers in measurements

data contains outliers: some samples are not generated from the same dgp



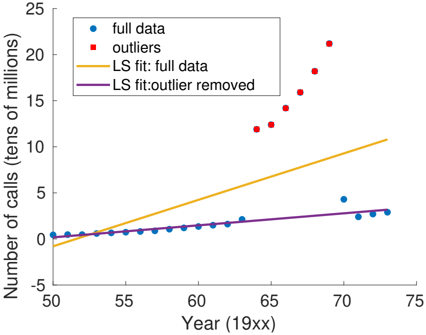

OLS estimate can be biased and pulled towards the outliers

robust regression is a method to overcome problems violating OLS assumptions

note: robust LS also refers to a method that is insensitive to parameter uncertainty

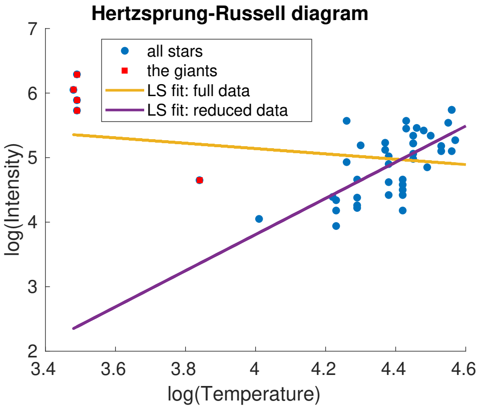
# Outlying in $y$-space

number of international phone calls from Belgium during 1950-1975




- contamination during 1964-1969: another recording system was used and gave total number of *minutes*
- full dataset LS fit: the slope is affected much (tilted upward) by the vertical outliers

# Outlying in $x$-space




- Hertzsprung-Russell diagram of the star cluster CYG OB1 (47 stars)
- $y$: log of star intensity
- $x$: log of temperature

- 43 stars lie on the main sequence, whereas the 4 remaining stars are called giants
- these giants are called (bad) leverage points, not errors but come from different populations

# Outlier diagnostics

# Residual analysis

we can use residual analysis to identify potentially unusual $y$ values

setting of dgp: $y = X\beta + e$, $e_i$'s are i.i.d. with variance $\sigma^2$

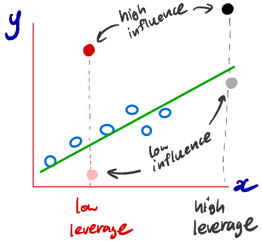
- (raw) residual: $r = y - \hat{y}$ (from a full regression)
- MSE $= s^2 = \hat{\sigma}^2 = \|r\|_2^2/(N-n)$ (unbiased estimate of $\sigma^2$)

1. **standardized residual**: $z_i = r_i/s$
2. **studentized (or jackknifed) residual**: $t_i = \frac{r_i}{\sqrt{\text{MSE}_{(-i)}(1-h_{ii})}}$

- $\text{MSE}_{(-i)}$ is mean square error based on the estimated model with $i$th data *removed*
- potential outliers on observations with $z_i$ or $t_i$ larger than 3 (threshold)
- another way to explain $t_i$ through *deleted residual*:

$$d_i = y_i - \hat{y}_i^{(-i)}, \quad t_i = \frac{d_i}{s(d_i)}, \quad s(d_i) : \text{standard deviation of all } d_j$$

# Hat matrix




the hat matrix is defined as

$$H = X(X^TX)^{-1}X^T$$

that maps $y$ to the prediction $\hat{y}$ (check $\hat{y} = X\hat{\beta} = Hy$)

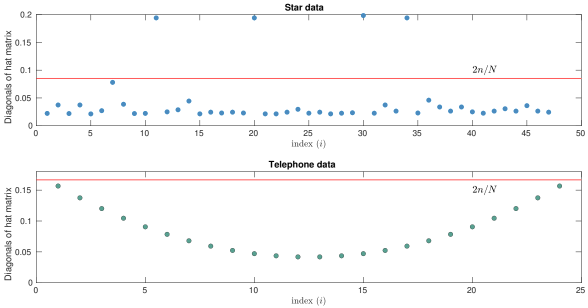
- $h_{ij}$: effect exerted by the $j$th observation on $\hat{y}_i$
- $h_{ii}$: equal to $\partial \hat{y}_i / \partial y_i$, giving the effect of the $i$th observation on its own prediction
- facts: using $H$ is symmetric and idempotent, it holds that

$$(1/N)\sum_{i=1}^N h_{ii} = n/N, \quad 0 \leq h_{ii} \leq 1, \quad i = 1, 2, \ldots, N$$

- the $i$th observation is called to have **high leverage** if

$$h_{ii} > \frac{2n}{N} \quad (h_{ii} \text{ is large by some threshold})$$

**example:** diagonals of hat matrix (star and telephone datasets); $X$ is standardized




- $h_{ii}$'s of the giant stars are larger than $2n/N$ (0.085)
- $h_{ii}$'s of the telephone data do not point out outlying observations
- the hat matrix, which is only based on $X$, can detect potential outliers in the $x$-direction
- whether the data point is influential or not also depends on $y_i$

# Cook's distance

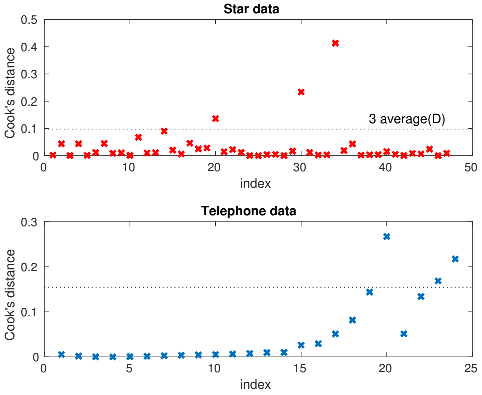
a distance measuring how far $\hat{\beta}$ moves when $(x_i, y_i)$ is removed

$$D_i = \frac{(\hat{y} - \hat{y}^{(-i)})^T(\hat{y} - \hat{y}^{(-i)})}{n\hat{\sigma}^2} = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^T(X^TX)(\hat{\beta} - \hat{\beta}^{(-i)})}{n\hat{\sigma}^2}$$

- $\hat{y}, \hat{y}^{(-i)}$ are the fitted responses with all data and with $i$th case removed, resp
- $\hat{\beta}^{(-i)}$ is the LS estimate on the dataset without $i$th observation
- $D_i$ measures the effect of the $i$th observation on the fitted vector and coefficients
- $D_i$ is algebraically equivalent to $D_i = \frac{r_i^2}{n\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$
- $D_i$ is made up of two components: 1) how well the model fits the response and 2) how far $x_i$ is from the rest of $x_i$'s
- any $i$th observation that has relatively high $D_i$ deserves a closer look (often, a threshold can be $4/N$ or $3 \cdot \bar{D}$)

# Example

Cook's distance calculated on star and telephone datasets




- telephone indices of high $D_i$ are 20, 23, 24 (years 1969, 1972, 1973)
- star indices of high $D_i$ are 20,30,34 (three of the giant stars)

# Robust methods

# Weighted least-squares

given $W$ a positive definite matrix that can be factorized as $W = L^T L$

a weighted least-squares (WLS) problem is

$$\underset{x}{\text{minimize}} \;\; (X\beta - y)^T W (X\beta - y)$$

- equivalent formulation: $\text{minimize}_x \;\; \|L(X\beta - y)\|^2$
- can be solved from the modified normal equation

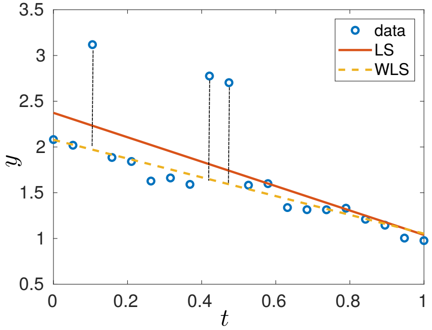
$$X^T W X \beta = X^T W y$$

- the solution is $\hat{\beta}_{\text{wls}} = (X^T W X)^{-1} X^T W y$ (if $X$ is full rank)
- $X\beta_{\text{wls}}$ is the *orthogonal projection* on $\mathcal{R}(X)$ w.r.t the new inner product

$$\langle x, y \rangle_W = \langle Wx, y \rangle$$

# Interpretation of WLS

when $m$-measurements contain some outliers (samples 3,9,10)

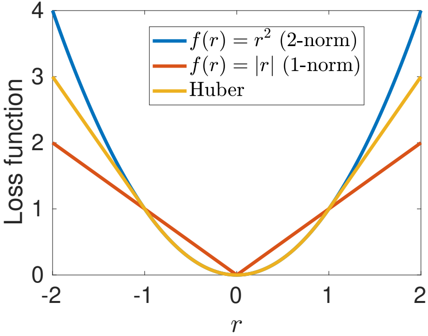



using $W = \mathbf{diag}(w_1, w_2, \ldots, w_m)$ gives WLS objective: $\sum_{i=1}^{m} w_i(y_i - x_i^T\beta)^2$

- use relatively **low** $w_3, w_9, w_{10}$ to penalize **less** on those samples
- the linear model tends not to adapt to outliers – making WLS a more robust method than LS

# Loss functions for robust regression

Huber and $\ell_1$ losses do not highly penalize large residuals




$$\text{Huber}(r) = \begin{cases} (1/2)r^2, & |r| \leq M \\ M(|r| - M/2), & |r| > M \end{cases}$$

- large residuals resulted from outliers should be less taken into account
- compared to square loss, Huber and $\ell_1$ have less penalty for large $r$
- for $r = y - X\beta$ (linear models), using all losses results in convex problems

# Regression M-estimators

minimize a penalty function of estimation residual: $r = y - X\beta$

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{N} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right), \qquad \hat{\sigma} \text{ is a preliminary scale estimate}$$

- for $\rho(r) = r^2$, we obtain LS estimate $\hat{\beta}_{\mathrm{ls}}$
- for $\rho(r) = |r|$, we obtain the least absolute deviation (LAD)

$$\hat{\beta}_{\mathrm{lad}} = \underset{\beta}{\mathrm{argmin}}\, \|y - X\beta\|_1$$

this allows us to compute the scale estimate

$$\hat{\sigma} = \frac{1}{0.675} \text{median}_i |r_i(\hat{\beta}_{\mathrm{lad}})|$$

and use it as initial scale parameter for other M-estimators

# Median absolute deviation

for samples $x = (x_1, x_2, \ldots, x_N)$,

$$\text{MAD} = \text{median}(|x_i - \text{median}(x)|)$$

MAD is the median of **absolute deviations** from data's median

- MAD is a robust estimate of scale parameter, which tells statistical dispersion
- MAD can be an estimator for the standard deviation

$$\hat{\sigma} = k \cdot \text{MAD}, \quad \text{where } \ k = 1/\Phi^{-1}(3/4) \approx 1.4826 \ \text{ for normal distribution}$$

this follows from that $\pm$MAD covers between 1/4 and 3/4 of the normal CDF

$$\frac{1}{2} = P(|X - \mu| \leq \text{MAD}) = P\left(|Z| \leq \frac{\text{MAD}}{\sigma}\right) \Rightarrow \Phi\left(\frac{\text{MAD}}{\sigma}\right) - \Phi\left(-\frac{\text{MAD}}{\sigma}\right) = \frac{1}{2}$$

and using $\Phi(-x) = 1 - \Phi(x)$ to show that $\Phi\left(\frac{\text{MAD}}{\sigma}\right) = 3/4$

# Iterative reweighted LS

1. start with calculating OLS and the corresponding residual $r = (r_1, \ldots, r_N)$
2. compute LS fit leverage values $h = (h_1, \ldots, h_N)$ (higher $h_i$, larger effect on LS fit)
3. compute the adjusted residuals: $\tilde{r}_i = \frac{r_i}{\sqrt{1-h_i}}$
4. standardize the residuals: $u_i = \frac{\tilde{r}_i}{Ks} = \frac{r_i}{Ks\sqrt{1-h_i}}$, $\quad s$ is an estimate of SD of error

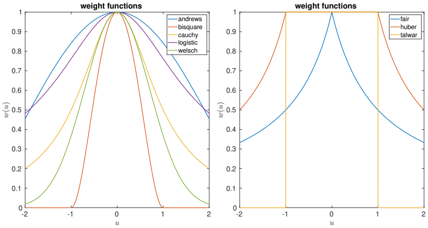
   $K$ : tuning constant, $\quad s = \text{MAD}/0.6745$ (making $s$ unbiased for normal dist)

5. compute the robust weights $w_i$ as a function of $u_i$ (**larger $u_i$**, **smaller $w_i$**)

$$\text{bisquare} : w_i = \begin{cases} (1-u_i^2)^2, & |u_i| < 1, \\ 0, & |u_i| \geq 1 \end{cases}, \quad W = \mathbf{diag}(w_1, w_2, \ldots, w_N)$$

see other weight functions in `robustfit` (MATLAB guide)

# Iterative reweighted LS



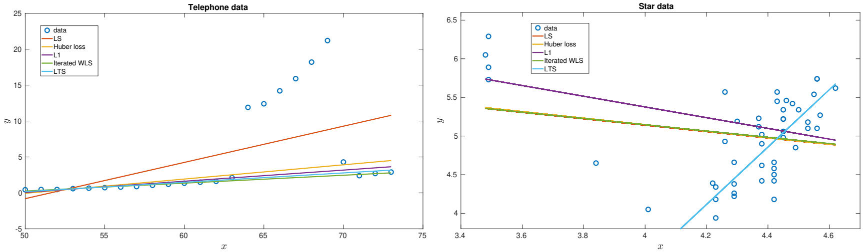

bisquare and talware use zero weight on the observation with large residual

6. compute the robust estimate: $\hat{\beta} = (X^T W X)^{-1} X^T W y$ (WLS estimate)
7. estimate the WLS error

$$e = \sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2$$

8. iteration stops if the fit converges or the maximum number of iterations is reached; otherwise, go to the second step

# Result using robust methods




- telephone dataset has outliers in $y$-direction; robust methods can ameliorate the effects of outliers quite well
- star dataset has bad leverage points; most robust methods (WLS, huber, L1) do not perform well on this dataset
- LTS (least trimmed square estimator) finds a subset of observations with small residuals and estimate $\hat{\beta}$ based on that subset of data (more detail in P. J. Rousseeuw book)

# Softwares and summary

# Summary

- robust regression is a broad term referring to extended methods that ameliorate issues in OLS (here, we focus on outlier problems)
- outliers cause OLS estimate to have large residuals
- a remedy for outlier issues can be done by using loss functions that *less* penalize large residuals (huber, 1-norm or MAE, (iterative) weighted LS)
- iterative WLS adjusts the weight function according to fitted residual in each step
- other methods/algorithms exist: least median square (LMS), least trimmed squares estimator (LTS), random sample consensus (RANSAC)

# Softwares

**MATLAB:** Statistical and machine learning toolbox

- `fitlm` with robust options
- `robustfit`: robust regression fit

**Python modules:**

- `statmodels`: robust linear model
- `scikit-learn`: robust linear model

# References


1. P. J. Rousseeuw, and A.M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, 1987
2. P. J. Rousseeuw, *Lecture note on Robust Statistics Part 3: Regression analysis*, LARS-IASC School, May 2019
3. MATLAB instruction, *Reduce Outlier Effects Using Robust Regression*, `https://www.mathworks.com/help/stats/robust-regression-reduce-outlier-effects.html`