# 9. Robust regression

## Least squares regression

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 = \text{argmin}_\theta \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{argmin}_\theta \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Why least squares regression?

- Historic (used since 1800)
- The least squares estimator $\hat{\theta} = (X^T X)^{-1} X^T y$ has a closed form solution, and is simple to compute
- If $y = X\theta + \epsilon$ and $\epsilon \sim N_n(0, \sigma^2 I)$:
  - Least squares estimator = MLE
  - Least squares estimator has smallest variance among all unbiased estimators (Gauss-Markov)



## Problems with LS regression

- When the statistical errors are not Normally distributed, the level of confidence intervals and tests is about right, but the power can be low (power = $P(\text{reject } H_0 | H_a \text{ is true})$).
- It is sensitive to outliers, since large residuals that are squared carry a lot of weight

## Robust regression

- Robust regression can (partly) resolve these problems. We will look at the following methods:
  - $L_1$ regression (=Least Absolute Deviations (LAD) regr.)
  - Huber regression
  - Mallows regression
  - Schweppe regression
  - Least Median of Squares (LMS) regression

4 / 12

## $L_1$ regression

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} |y_i - x_i^T \theta|$$

- Older than LS: Boscovich (1760), Laplace (1789)
- Did not become popular, since the solution cannot be written in closed form (no problem anymore with modern computers; can be solved efficiently with interior point methods)
- In location model $y_i = \alpha + \epsilon_i$, $L_1$ regression gives median of the data
- Is more robust against outliers in the $y$-direction, but still very sensitive to outliers in the $x$-direction
- Is inefficient when the errors are normally distributed; needs about $50\%$ more observations for same precision

5 / 12

## Huber regression

$$\hat{\theta} = \text{argmin}_{\theta} \sum_{i=1}^{n} \rho_c(y_i - x_i^T\theta),$$

where

$$\rho_c(u) = \begin{cases} u^2/2 & \text{if } |u| \leq c \\ c(|u| - c/2) & \text{if } |u| \geq c \end{cases}$$

- Compromise between $L_1$ and $L_2$ regression:
  - $c = \infty \Rightarrow L_2$ regression (=least squares)
  - $c = 0 \Rightarrow L_1$ regression (use $\rho_c(u) = |u|$)
- Idea: penalize small residuals quadratically, and large residuals linearly
- Computation: solve $\sum_{i=1}^{n} \psi_c(y_i - x_i^T\theta)x_i = 0$, where $\psi_c(u) = \rho_c'(u) = \text{sign}(u)\min(|u|, c)$.
- The changepoint $c$ should be chosen suitably w.r.t residuals. Computation with iterated weighted least squares.

6 / 12

## $L_1$/Huber estimators

- One cannot write down the exact distribution of the estimators $\Rightarrow$ use asymptotic arguments or bootstrap
- Outliers in the $y$-direction have limited influence, but outliers in the $x$-direction don't.
  Solution: Mallows/Schweppe

7 / 12

## Mallows/Schweppe regression

$$\sum_{i=1}^{n} \eta\left(x_i, \frac{y_i - x_i^T\hat{\theta}}{\hat{\sigma}}\right) x_i = 0$$

- Mallows:

$$\eta(x, r) = \min\left(1, \frac{a}{\|Ax\|}\right) \psi_c(r)$$

- Schweppe:

$$\eta(x, r) = \frac{1}{\|Ax\|} \psi_c(\|Ax\| r)$$

- $\|Ax\|$ is a measure of leverage of $x$, for example $\|Ax\|^2 = \text{const} \cdot x^T(X^TX)^{-1}x$, but then robust version
- $\psi_c = \rho'(c)$ from Huber regression

8 / 12

## Breakdown point

The breakdown point of an estimator = the proportion of incorrect observations (i.e. arbitrarily large observations) an estimator can handle before giving an arbitrarily large result

- Breakdown point of average: 0
- Breakdown point of median: 1/2
- Breakdown point of Least Squares regression: 0
- Breakdown point of $L_1$ and Huber: 0 (in $x$-direction)
- Breakdown point Mallows/Schweppe: $\leq 1/p$

9 / 12

## LMS regression

$$\hat{\theta} = \text{argmin}_{\theta}\text{median}((y_i - x_i^T\theta)^2)$$

- See picture on slide
- Hampel (1975), Rousseeuw (1984)
- Breakdown point is approximately 0.5
- Difficult to compute because of many local minima
- Inefficient when statistical errors are normally distributed (convergence rate $n^{-1/3}$). This can be improved by replacing the median by an $\alpha$-truncated mean that leaves out the $\alpha n$ observations with the largest residuals (least trimmed squares).



## MM-estimation

- First find highly robust M-estimate of $\sigma$ (first M).
- Then keep $\hat{\sigma}$ fixed and find a close by M-estimate of $\theta$, for example using a Newton step (second M).

**Some closing thoughts (see Faraway Ch 13)**

- Robust estimators protects against long-tailed errors, but not against problems with model choice and variance structure. These latter problems can be more serious than non-normal errors.
- Inference for $\hat{\theta}$ is more difficult. One can use bootstrap.
- Robust methods can be used in addition to least squares. There is cause to worry if the two estimators differ a lot.