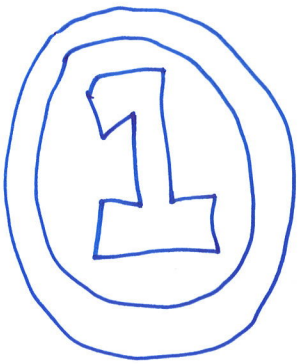
Robust Regression (when errors are non-normally heavy-tailed, extreme)

M-Estimation

- Chooses the $\hat{\beta}$ values to minimize

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \hat{\beta}) \quad (*)$$



where $\rho(\cdot)$ is a function. Common choices:

① $\rho(x) = x^2$ (least squares regression)

② $\rho(x) = |x|$ least absolute deviation regression (LAD)
or L_1 regression ($|x| = \sum_{i=1}^{n} |x_i|$ is the L_1-norm)

③ Huber's Method:

$$\rho(x) = \begin{cases} x^2/2 & \text{for } |x| \le c \text{ (some threshold)} \\ c|x| - c^2/2 & \text{o/w} \end{cases}$$

Here, c is a robust estimator of σ, such as the median of the $|\hat{\epsilon}_i|$ values.

Relation of M-Estimation to Weighted Least Squares

Normal Equations for Least Squares (what you get when differentiating w.r.t. β and setting to 0)

$$X^T X \hat{\beta} = X^T Y$$

Normal Equations for Weighted Least Squares:

$$X^T \Sigma^{-1} X \hat{\beta} = X^T \Sigma^{-1} Y,$$ or in summations: $\sum_{i=1}^{n} w_i x_{ij} y_i = \sum_{i=1}^{n} w_i x_{ij} \sum_{j=1}^{p} x_{ij} \hat{\beta}_j$

... or with summations ...

2

$$\sum_{i=1}^{n} w_i x_{ij} \sum_{k=1}^{p} x_{ik} \hat{\beta}_k = \sum_{i=1}^{n} w_i x_{ij} y_i \quad , \quad j \in \{1, 2, \dots, p\} \qquad (**)$$

Differentiating the M-estimate criterion (*) w.r.t. $\hat{\beta}_j$ and setting to 0:

$$\frac{d}{d\hat{\beta}_j} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) = \sum_{i=1}^{n} \rho'(y_i - x_i^T \beta) \cdot (-x_{ij}) = 0$$

(in summation form) $\Rightarrow$

$$\sum_{i=1}^{n} \rho'\Big(\underbrace{y_i - \sum_{k=1}^{p} x_{ik} \hat{\beta}_k}_{i^{th} \text{ residual} \, = \, u_i}\Big) x_{ij} = 0 \quad , \quad j \in \{0, 1, \dots, p\}$$

Or

$$\sum_{i=1}^{n} \underbrace{\frac{\rho'(u_i)}{u_i}}_{w(u_i)} x_{ij} \left(y_i - \sum_{k=1}^{p} x_{ik} \hat{\beta}_k\right) = 0, \quad j \in \{1, 2, \dots, p\}$$

or

$$\sum_{i=1}^{n} w(u_i) x_{ij} \left(y_i - \sum_{k=1}^{p} x_{ik} \hat{\beta}_k\right) = 0, \quad j \in \{1, 2, \dots, p\}$$

weight function

SAME as (**)

1. $w(u) =$ constant $\Rightarrow$ ordinary least squares.

2. $w(u) = \frac{1}{|u|} \Rightarrow$ L.A.D. Note the weight decreases as the residual increases: more extreme observations have smaller weight.

3. $w = \begin{cases} 1 & \text{for } |u| \le c \\ c/|u| & \text{o'wise} \end{cases}$

is the HUBER method, which is a sort-of compromise between least-squares and L.A.D.

---

* There are many other potentially good choices for $\rho(\cdot)$.
* M-estimation can require significant computing time because it is carried out iteratively: the weights depend on the residuals: get weights, get $\hat{\beta}$'s, get residuals, get the weights, get $\hat{\beta}$'s, get residuals, get ....

Get standard errors using weighted-least-squares with $\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}^2 (X^T W X)^{-1}$, w/ a robust estimate of σ^2.