Iteratively Reweighted Least Squares (IRLS)

NOTE: converted to LaTeX via LLM.

Two use cases:

- Robust regression / M-estimation
- GLMs

Here we only care about robust regression.

Given $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$,

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} L(\beta) \\ &= \sum_{i=1}^{n} \rho(r_i(\beta)), \quad r_i(\beta) = \frac{y_i - x_i^T \beta}{\sigma_i}.\end{aligned}$$

where ρ is convex, even, and increasing in $|r|$.

Define score function $\psi(r) = \rho'(r)$ and weight function

$$w(r_i) = \frac{\psi(r_i)}{r_i\, \sigma_i^2}, \quad (with\ w(0) := \rho''(0)).$$

We can write robust weights $w(r) \equiv \frac{\psi(r)}{r}$ and precision weights $\lambda_i \equiv \frac{1}{\sigma_i^2}$ such that

$$w_{\mathrm{tot},i} = \lambda_i\, w(r_i),$$

which will be used in WLS.

Aside: Majorize-Minimize (MM)

We want

$$\hat{\theta} = \arg\min_{\theta} f(\theta).$$

MM iterates at $\theta^{(t)}$ using a surrogate $g(\theta|\theta^{(t)})$ such that

1. Majorize condition:
$$f(\theta) \le g(\theta|\theta^{(t)}) \quad \forall\theta, \qquad f(\theta^{(t)}) = g(\theta^{(t)}|\theta^{(t)}).$$
2. Minimization step:
$$\theta^{(t+1)} = \arg\min_{\theta} g(\theta|\theta^{(t)}).$$

This generalizes EM. Proof (sketch):

$$f(\theta^{(t+1)}) \le g(\theta^{(t+1)}|\theta^{(t)}) \le g(\theta^{(t)}|\theta^{(t)}) = f(\theta^{(t)}).$$

For IRLS we replace $\rho(r)$ with a quadratic tangent at the current point and solve WLS.

Key Idea

A (quasi-)Newton or MM step for L is equivalent to solving WLS with weights $w_i = w(r_i)$.

At iterate $\beta^{(t)}$, set

$$W^{(t)} = \operatorname{diag}(w(r_1^{(t)}), \ldots, w(r_n^{(t)})),$$

and solve

$$\beta^{(t+1)} = \arg\min_{\beta} \sum_{i=1}^{n} w_i^{(t)} (y_i - x_i^T \beta)^2 = (X^T W X)^{-1} X^T W y.$$

Justifying the weights

Quasi-Newton view:

$$\nabla L(\beta) = -X^T \operatorname{diag}\left(\tfrac{1}{\sigma_i} \psi(r_i)\right),$$
$$\nabla^2 L(\beta) = X^T \operatorname{diag}\left(\tfrac{1}{\sigma_i^2} \psi'(r_i)\right) X.$$

Since $\psi'(r)$ is messy, approximate with $\psi'(r) \approx \frac{\psi(r)}{r}$, so

$$H = X^T \operatorname{diag}(w_i) X.$$

MM view: Each $\rho(r)$ is concave in $u = r^2$, so linearizing in u gives

$$\rho(r) \le \tfrac{1}{2} w(r^{(t)}) r^2 + \text{const}.$$

Substituting $r_i = (y_i - x_i^T \beta)/\sigma_i$ produces a quadratic surrogate.

Canonical examples

Gaussian:

$$\begin{aligned} \rho(r) &= \tfrac{1}{2} r^2, \\ \psi(r) &= r, \\ w(r) &= 1, \\ w_{\text{tot},i} &= \tfrac{1}{\sigma_i^2}. \end{aligned}$$

$\Rightarrow$ standard GLS.

Huber:

$$\begin{aligned} w(r) &= 1, \quad |r| \le c, \\ w(r) &= \tfrac{c}{|r|}, \quad \text{else}, \\ w_{\text{tot},i} &= \frac{1}{\sigma_i^2} \times (\le 1). \end{aligned}$$

Cauchy:

$$
\begin{aligned}
\rho(r) &= \tfrac{c^2}{2} \log\left(1 + \frac{r^2}{c^2}\right), \\
\psi(r) &= \frac{c^2 r}{c^2 + r^2}, \\
w(r) &= \frac{1}{1 + r^2/c^2}, \\
w_{\text{tot},i} &= \frac{1}{\sigma_i^2 (1 + r^2/c^2)}.
\end{aligned}
$$

Note: $r_i = (y_i - x_i^T \beta)/\sigma_i$.

Practicalities for robust regression

1. Initialization

- OLS / GLS
- LTS / S-estimator
- IRLS is local, so initialization matters.

2. Convergence

- Track $\|\beta^{(t+1)} - \beta^{(t)}\|$
- Track $L(\beta) = \sum \rho(r_i)$
- For convex ρ (Huber, Cauchy), convergence is guaranteed for full-rank X.

3. Numerical stability

- Normal equations $(X^T W X)\beta = X^T W y$ can be ill-conditioned if weights are tiny or X is collinear.
- Use QR/Cholesky with damping.
- Ridge: $(X^T W X + \lambda I)$.