Getting away from $X$ (and other Gaussian-based procedures)

- Pays big dividends

But,

- Classical approaches aren’t very flexible
- Limited availability of SE’s and other properties

Enter M estimates
M-estimators

"MLE-like"

- Found by adapting regression "rreg"

- Natural computation for the variance estimate

- Direct assessment of properties

- Useful for a single batch, regression, ANOVA, etc.

Motivation: The sample mean
The sample mean as least-squares

Data: \( x_1, \ldots, x_n \) \( \quad x_i = \theta_1 + \epsilon_i \)

Find the best \( \theta \) (MLE for \( \theta \))
\[
\sum_{i=1}^{n} (x_i - \theta)^2 \quad \text{is minimized.}
\]

For any "\( b \)"
\[
\sum (x_i - b)^2 = \sum (x_i - \theta)^2
\]
\[
= \sum (x_i - a + a - b)^2
\]
\[
= \sum (x_i - a)^2 + n(a - b)^2 + 2(a-b)\sum (x_i - a)
\]

So, if \( \sum (x_i - a) = 0 \), \( \quad (a = \bar{x}_n) \)
\[
\sum (x_i - b)^2 \leq \sum (x_i - a)^2
\]
CALCULUS

Mean (Gaussian)

\[ \min_a \sum (x_i - a)^2 \]

\[ \Rightarrow \text{solve} \quad \Rightarrow \sum (x_i - a) = 0 \]

Median (Bilateral exponential)

\[ \min_a \sum |x_i - a| \]

\[ \Rightarrow \text{solve} \quad \Rightarrow \sum \text{sgn}(x_i - a) = 0 \]

\[ a = \begin{cases} X(\alpha_i) & \text{if } \alpha_i \text{ odd} \\ \text{any value between} & \begin{cases} X(\beta_i), X(\beta_i + 1) & \text{if } \alpha_i \text{ even} \end{cases} \end{cases} \]
General M-estimators

Find $T$ s.t.

$$
\sqrt{n} \sum_{i=1}^{n} \varphi \left\{ \frac{X_i - T}{c/S} \right\} \text{ is a min}
$$

Some (resistant) estimate of scale

Tuning constant

mean: $g(u) = u^2$, $\psi(u) = u$

median: $g(u) = |u|$, $\psi(u) = \text{sgn}(u)$

Equivalent

Find a $T$ that solves

$$
\frac{n}{c/S} \sum_{i=1}^{n} \psi \left( \frac{X_i - T}{c/S} \right) = 0
$$

Estimating equation

$\psi = g'$

$T$ is not necessarily unique
Requirements

- $y$ must be odd: $y(-u) = -y(u)$
  (unbiased)

- Smooth - for mathematical convenience
  and for resistant/robust/efficient

- "like "a" near 0"

- Turn down as data get far away
  (redescend)

- Smooth approach to zero
INTERPRETATION

- $\psi$ measures how much influence to give data points that are a number of scale units (CS) from $T$.

- $S$ measures scale. Want it robust/resistant.

  IQR: $Q_3 - Q_1$

  MAD: Median Absolute Residual

    \[ r_i = |x_i - T| \]

    \[ \text{MAD} = \text{median}(|r_1|, \ldots, |r_n|) \]

- $C$ measures how much we penalize

  \[ C \rightarrow \infty \quad \text{ALL } x^n \text{ are "close"} \]

  \[ C = 0 \quad "\quad " \quad "\text{far}" \]
WHAT IF GAUSSIAN?

IQR: 1.35σ

MAD: 1.35σ

Exercise: Show these
(Don't need to hand in)
\[ u(x) = \left( \frac{X_i - T}{c} \right) \]
• $\Psi$ is proportional to the influence curve

• Except for $\Psi(u) = u$
  
  $\Psi(u) = \text{sgn}(u)$

  There is no closed form.
  (for M-estimates!)

• Many non-M-estimates have "implicit" $\Psi$'s.
  (for example, Trimmed means)

• We will have a data-based version (see HW1).
The biweight bisquare weight

\[ \psi(u) = \begin{cases} \kappa u (1-u^2)^2, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases} \]

\[ u = \frac{x - \bar{x}}{c s} \]

...60.9... MAD
Computing M-estimates

Solve
\[ \sum \psi \left( \frac{x_i - t}{c_s} \right) = 0 \]

Let \( u^{(k)}_i = \frac{x_i - t_k}{c \cdot s_k} \), \( k \neq \text{step} \)

Solve
\[ \sum \omega^{(k)} \cdot \frac{\psi(u^{(k)}_i)}{\omega^{(k)}(u_i)} = 0 \]

Or
\[ \sum \omega^{(k)}(u_i) \omega(u_i) = 0 \]

L.S. idea
\[ \omega(u) = \frac{\psi(u)}{u} \]

Iteratively re-weighted least squares
Iteration

\[ \sum u_{i}^{(k+1)} w(u_{i}^{(k)}) = 0 \]

\[ \sum_{i=1}^{n} \frac{(x_{i} - T_{k+1}) w(u_{i}^{(k)})}{c_{i} s_{k+1}} = 0 \]

\[ T_{k+1} = \frac{\sum x_{i} w(u_{i}^{(k)})}{\sum w(u_{i}^{(k)})} \]

\[ S_{k+1} = \text{MAD using } T_{k+1}, \text{ etc.} \]

Iteratively re-weighted least squares!!
<table>
<thead>
<tr>
<th>Name</th>
<th>( \psi )</th>
<th>( \psi' )</th>
<th>( w ) prop'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Squares</td>
<td>( u )</td>
<td>( \frac{\psi'(u)}{u} )</td>
<td>1</td>
</tr>
<tr>
<td>Least Abs Val</td>
<td>( \text{SGN}(u) )</td>
<td>( \frac{\text{SGN}(u)}{u} )</td>
<td></td>
</tr>
<tr>
<td>Tukey's Biweight</td>
<td>( u(1-u^2)^2 )</td>
<td>( (1-u^2)^2 )</td>
<td>(</td>
</tr>
</tbody>
</table>
$$w(u) = \frac{\psi(u)}{u}$$

$$u = \frac{x - T}{\varepsilon s}$$

Plot of Relative Weights
<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\rho(t)$</th>
<th>$\psi(t)$</th>
<th>$w(t)$</th>
<th>Range of $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>$\frac{1}{2} t^2$</td>
<td>$t$</td>
<td>1</td>
<td>$</td>
</tr>
<tr>
<td>LAR</td>
<td>$</td>
<td>t</td>
<td>$</td>
<td>$\text{sgn}(t)$</td>
</tr>
<tr>
<td>Huber$^a$</td>
<td>$\frac{1}{2} t^2$</td>
<td>$t$</td>
<td>1</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$k</td>
<td>t</td>
<td>- \frac{1}{2} k^2$</td>
<td>$k \text{sgn}(t)$</td>
</tr>
<tr>
<td>Andrews$^a$</td>
<td>$\frac{A^2}{2A^2} [1 - \cos(t/A)]$</td>
<td>$A \sin(t/A)$</td>
<td>$\frac{A}{t} \sin(t/A)$</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$</td>
</tr>
<tr>
<td>Biweight$^a$ (bisquare)</td>
<td>$\frac{B^2}{6} [1 - (1 - (t/B)^2)^3]$</td>
<td>$t[1 - (t/B)^2]^2$</td>
<td>$[1 - (t/B)^2]^2$</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$</td>
</tr>
</tbody>
</table>

$^a$The illustrative examples of $\rho$-functions and $\psi$-functions in Figure 8-2 use $k = 1$ for the Huber, $A = 1/\pi$ for the Andrews, and $B = 1$ for the biweight.
Figure 8-2. The $\rho$-function and $\psi$-function for five $M$-estimators. (a) OLS $\rho$-function. (b) OLS $\psi$-function. (c) LAR $\rho$-function. (d) LAR $\psi$-function. (e) Huber $\rho$-function. (f) Huber $\psi$-function. (g) Andrews $\rho$-function. (h) Andrews $\psi$-function. (i) Biweight $\rho$-function. (j) Biweight $\psi$-function. Table 8-2 gives scaling.
FEATURES OF ESTIMATORS

Breakdown Bound
Largest possible data fraction that can be moved (to \( \pm \infty \)) w/o sending the estimator to \( \pm \infty \)

\[
\text{mean} = 6^{-1/2} \quad \text{median} = 0.5 - \frac{2}{n} \\
\text{biweight} = 0.5 \left( \text{due to scale measure} \right)
\]

Gross error sensitivity
Height of the influence curve (\( \infty \) to \( \psi \))

Rejection Point
Where does \( \psi(u) = 0 \)
\[\therefore \psi(u) = 0 \quad |u| > ?\]

mean + median = \( \infty \)

biweight = c.S.
biweight(mdat, 6, 10)

```
function(y, c, reps, init = median(y))
{
    coef <- init
    resid <- y - coef
    for(iter in 1:reps) {
        scale <- median(abs(resid))/0.6745
        u <- abs(resid)/(scale * c)
        w <- ((1 + u) * (1 - u))^-2
        w[u > 1] <- 0
        w <- w/sum(w)
        coef <- t(w) %*% y
        resid <- y - coef
        next
    }
    list(resids = resid, w = w, est = coef, c = c)
}
```

"y" is the data vector
"c" is the tuning constant
"reps" is the number of iterations you want
"init" is the starting value for the measure of center (defaulted at the median)

Example:
You have made mountain data an S variable by: mdat<-scan('mountain')
Then, mtsmry<-biweight(mdat, 6, 10)
DOCUMENTATION FOR "BIWEIGHT"

The program "biweight" is simple to use and programmed simply. Here is the printout of the function:

```r
function(y, c, reps, init = median(y))
{
  coef <- init
  resid <- y - coef
  for(iter in 1:reps) {
    scale <- median(abs(resid))/0.6745
    u <- abs(resid)/(scale * c)
    w <- ((1 + u) * (1 - u))^2
    w[u > 1] <- 0
    w <- w/sum(w)
    coef <- t(w) %*% y
    resid <- y - coef
    next
  }
  list(resids = resid, w = w, est = coef, c = c)
}
```

"y" is the data vector
"c" is the tuning constant
"reps" is the number of iterations you want
"init" is the starting value for the measure of center (defaulted at the median).

Example:
You have made mountain data an S variable by: mdat$scan('mountain')

Then, mtsmry@biweight(mdat, 6, 10)
stores in "mtsmry" the biweight an various details. You can type mtsmry to see the stuff or use the estimate which will be named "mtsmry$est".

If you don't need to store things, don't assign the result and you'll see the printout on your screen.

If you want to start at the last iteration do:
biweight(mdat, 6, 15, init = mtsmry$est).
Mountain Data

\[ e = 6/1.675 \]

<table>
<thead>
<tr>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

\[ e = 9/1.675 \]

<table>
<thead>
<tr>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>f(x)</td>
</tr>
</tbody>
</table>

WEIGHTS

- MAD = 2.6
- MADmed = 1
- X = \text{median of } x

\[
\text{E} = -1.6
\]

\[
\text{MAD} = 0
\]
Histogram of Mountain Data

Tuning Constant vs Estimator by Biweight Function

Dung-tsa Chen
"LOGICAL" VARIANCE (SCALE)

\[ T: \sum_i \frac{\psi(x_i - T)}{c s_i} = 0 \]

\[ S: \sum_i \left\{ \frac{(x_i - T)}{c s_i^{k+1}} \psi(x_i - T) - \frac{1}{s_i^{k+1}} \right\} = 0 \]

\[ \psi = u \Rightarrow \text{Sample Variance} \quad s^2 \]

\[ \psi = \pm 1 \quad \text{Beware} \]

\( f \) is like a \(-\log\)-likelihood

\[ \frac{d}{ds} \left\{ + \log(s) + \psi \left( \frac{x - T}{c s} \right) \right\} \]

\[ = \left\{ \frac{1}{s} - \frac{x - T}{c s^2} \psi \left( \frac{x - T}{c s} \right) \right\} \]

Likelihood: at \( \psi = \frac{1}{x - T} \)
M-ESTIMATES for Regression

\[ \min \beta \sum_i f \left( \frac{y_i - x_i \beta}{c_s} \right) \]

\[ y_i = x_i \beta + \epsilon_i \]

\[ \frac{d}{d\beta} \Rightarrow \sum_i \psi \left( \frac{y_i - x_i \beta}{c_s} \right) x_i = 0 \]

No protection here.

Outlier and high leverage.
Bounded Influence

\[ Y_i = \beta_0 + \beta_1 X_i \]

OLS

\[ \hat{\beta}_1 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \]

\[ = \sum \hat{\beta}_1 \tilde{w}_i \]

\[ \hat{\beta}_1 = \frac{Y_i - \bar{Y}}{X_i - \bar{X}} \quad \text{and} \quad \tilde{w}_i = \frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2} \]

Bounded Influence

M-estimates

\[ \hat{\beta}_1^n = \sum \hat{\beta}_1 \cdot w(u_i) \cdot \tilde{w}_i \cdot \tilde{w}_i^*(x) \]

\[ u_i = \frac{Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i}{s} \]

\[ 0.5 \cdot \omega = (1 - h_{ii})^{1/2} \]