Lecture 8: The LAD Estimator

Econometrics 2 — *Medians, Absolute Deviations, and the Linear Model*

Instructor: Prof. S. Arvanitis | **Digitisation of notes by:** T. Kourtalis

Date: March 17, 2026

▷ **Amber boxes = Handwritten Notes (professor's words)**

◇ **Teal boxes = Student's Notes**

Recall from Lecture 7

In Lecture 7 we introduced the LAD (Least Absolute Deviations) estimator as an alternative to OLS that minimises $\sum_i |Y_i - X_i'\beta|$ instead of $\sum_i (Y_i - X_i'\beta)^2$. We noted that LAD is robust to outliers and targets the *conditional median* rather than the conditional mean. This lecture formalises the theoretical justification for LAD, proves that minimising absolute deviations yields the median, and connects this to the linear model framework.

1 Linear Model and the LAD Estimator

▷ **Handwritten Notes** (what the professor said)

The Least Absolute Deviations Estimator (LADE) results from the minimization of:

$$\frac{1}{n}\sum_{i=1}^{n} |Y_{(i)} - X_{(i)}\beta|$$

Question: What modification of the linear model could support such an objective function?

The modification will be related to the **median** of the distribution of $\varepsilon_{(i)}$, where:

$$\varepsilon_n = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

◇ Student's Notes

Comparing OLS and LAD objectives:

Estimator	Objective function	Targets
OLS	$\min_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\beta)^2$	$\mathbb{E}[Y\|X]$ (mean)
LAD	$\min_\beta \frac{1}{n}\sum_{i=1}^n \|Y_i - X_i'\beta\|$	$\mathrm{Med}(Y\|X)$ (median)

The key question: Just as OLS is justified by assuming $\mathbb{E}[\varepsilon|X] = 0$, what assumption on the error distribution justifies LAD?
Answer preview: We will assume $\mathrm{Med}(\varepsilon|X) = 0$ instead of $\mathbb{E}[\varepsilon|X] = 0$.

2 Defining the Median

▷ **Handwritten Notes** (what the professor said)

Let the random variable $\varepsilon \sim \mathbb{P}$ (where $\mathbb{P}$ is a probability distribution on $\mathbb{R}$). The cumulative distribution function (CDF) of $\mathbb{P}$ is $F : \mathbb{R} \to \mathbb{R}$, defined as:

$$F(x) := \mathbb{P}(\varepsilon \leq x)$$

The median of $\mathbb{P}$ (or equivalently of ε) is any real number ζ such that:

$$F(\zeta) = \mathbb{P}(\varepsilon \leq \zeta) = \frac{1}{2} \iff \mathbb{P}(\varepsilon > \zeta) = \frac{1}{2}$$

Definition: Median

The **median** of a random variable ε with CDF F is any value $\zeta \in \mathbb{R}$ satisfying:

$$F(\zeta) = \mathbb{P}(\varepsilon \leq \zeta) = \frac{1}{2}$$

Equivalently (when $\mathbb{P}(\varepsilon = \zeta) = 0$):

$$\mathbb{P}(\varepsilon < \zeta) = \mathbb{P}(\varepsilon > \zeta) = \frac{1}{2}$$

The median splits the probability mass into two equal halves.

◇ Student's Notes

Important distinction: mean vs. median

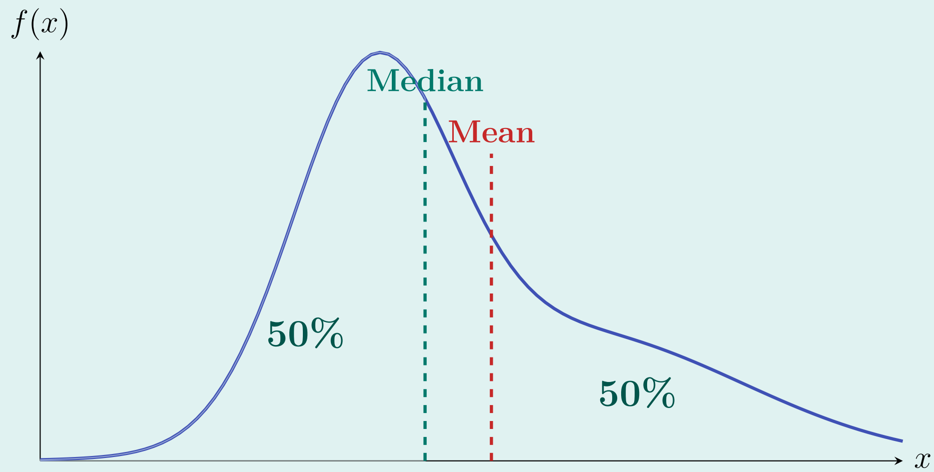


Figure 1: Median vs. mean in a right-skewed distribution. The median splits the area under the curve in half; the mean is pulled toward the tail. For symmetric distributions, they coincide.

Key properties:

Property	Mean	Median
Existence	Requires $\mathbb{E}[\|\varepsilon\|] < \infty$	Always exists
Uniqueness	Unique if exists	May not be unique
Sensitivity to outliers	High	Low (robust)
Symmetric distributions	$\mathbb{E}[\varepsilon] = \text{Med}(\varepsilon)$	(same)
Minimises	$\mathbb{E}[(\varepsilon - c)^2]$	$\mathbb{E}[\|\varepsilon - c\|]$

2.1 Example: Discrete Distribution

▷ **Handwritten Notes** (what the professor said)

Suppose ε takes values:

$$\varepsilon = \begin{cases} -1, & \text{with probability } 1/2 \\ 1, & \text{with probability } 1/2 \end{cases}$$

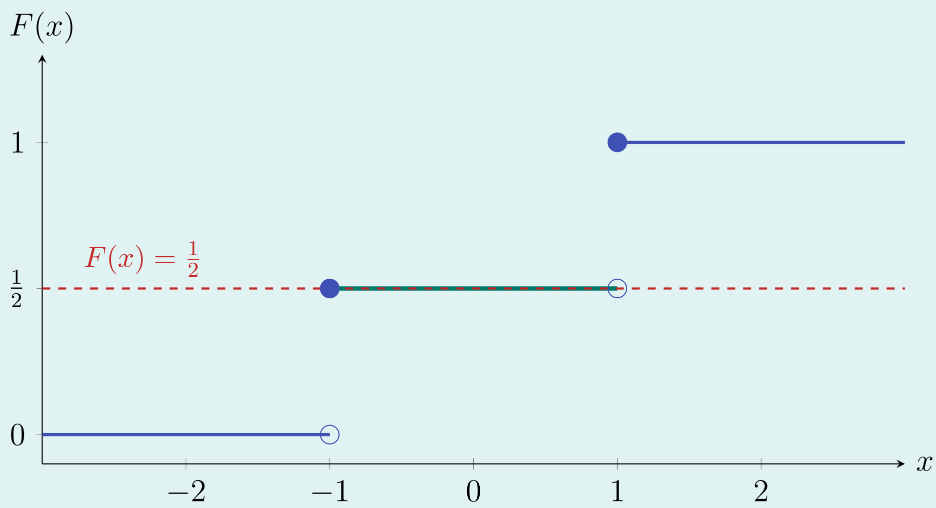
The CDF is:

$$F(x) = \begin{cases} 0, & x < -1 \\ 1/2, & -1 \le x < 1 \\ 1, & 1 \le x \end{cases}$$

In this specific case, *any* value $\zeta \in [-1, 1]$ is a median of this distribution.

◇ Student's Notes

Visualising the non-unique median case:



The flat region at $F(x) = \frac{1}{2}$ means **any** $\boldsymbol{\zeta \in [-1, 1]}$ is a median

Figure 2: CDF for the discrete distribution. The teal segment shows the flat region where the CDF equals $1/2$*.*

When does non-uniqueness happen?

- Discrete distributions with probability mass at exactly $1/2$
- Distributions with flat regions in the CDF
- Generally, when F is not strictly increasing around the median

2.2 Sufficient Condition for Uniqueness

▷ **Handwritten Notes** (what the professor said)

A sufficient condition for the median $\mathrm{Med}(\varepsilon)$ to be uniquely defined is for $\mathbb{P}$ to have a probability density function, say f, and that $f(\mathrm{Med}(\varepsilon)) > 0$.

Key Result

Uniqueness of the Median
If ε has a probability density function f such that:

1. f is continuous
2. $f(\zeta) > 0$ where ζ is a median

Then the median is **unique**.

Intuition: If the density is strictly positive at the median, the CDF is strictly increasing there, so there is only one point where $F(x) = 1/2$.

◇ **Student's Notes**

Examples of distributions and their medians:

Distribution	Median	Mean	Unique?
Normal(μ, σ^2)	μ	μ	Yes
Uniform$[a, b]$	$(a+b)/2$	$(a+b)/2$	Yes
Exponential(λ)	$\frac{\ln 2}{\lambda}$	$\frac{1}{\lambda}$	Yes
Cauchy$(0, 1)$	0	undefined	Yes
Discrete (above)	any $\in [-1, 1]$	0	No

Note: The Cauchy distribution has a well-defined median but its mean does not exist (the tails are too heavy). This shows the robustness of the median.

3 Minimising Expected Absolute Deviation

▷ **Handwritten Notes** (what the professor said)

Let us assume that the expectation $\mathbb{E}(\varepsilon)$ exists (which holds $\iff$ $\mathbb{E}(|\varepsilon|) < +\infty$).
This implies that the function:

$$\psi(c) := \mathbb{E}(|\varepsilon - c|)$$

is well-defined (i.e., $\psi(c) \in \mathbb{R}\ \forall c \in \mathbb{R}$). Therefore, we can ask whether there is some constant c at which ψ is minimized.
We know that:

$$|\varepsilon - c| = \begin{cases} \varepsilon - c, & \varepsilon \geq c \\ c - \varepsilon, & \varepsilon < c \end{cases}$$

We will ignore the specific point $c = \varepsilon$ for which we do not have differentiability. The derivative with respect to c is:

$$\frac{\partial |\varepsilon - c|}{\partial c} = \begin{cases} -1, & \varepsilon > c \\ 1, & \varepsilon < c \end{cases}$$

◇ **Student's Notes**

Setup for the optimisation:
We want to find:

$$c^* = \arg\min_{c \in \mathbb{R}} \psi(c) \quad \text{where} \quad \psi(c) = \mathbb{E}[|\varepsilon - c|]$$

Strategy:

1. Write the derivative of ψ using the chain rule and expectation
2. Set the derivative to zero (first-order condition)
3. Show that the solution is the median

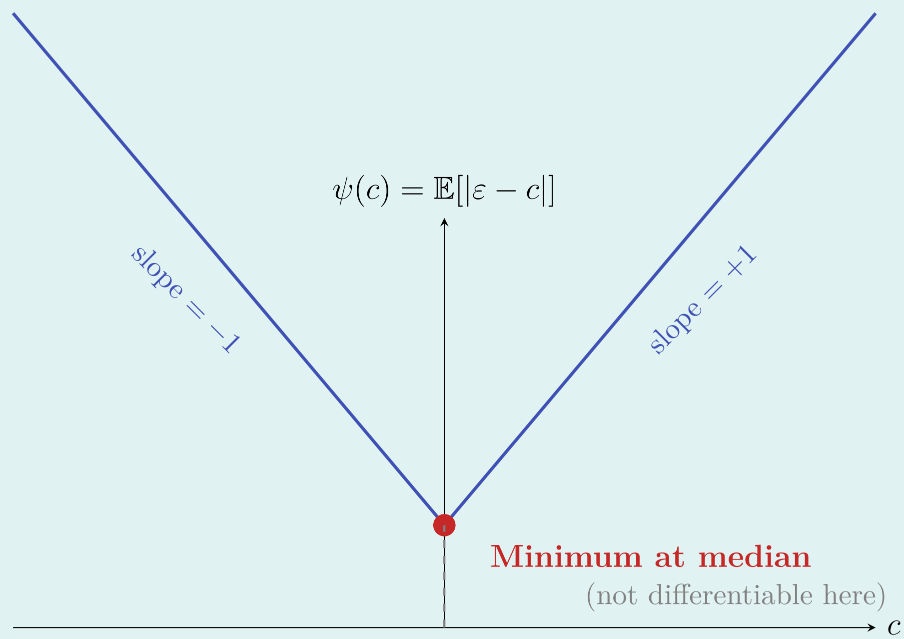


Figure 3: The expected absolute deviation $\psi(c)$ as a function of c. It is V-shaped, minimised at the median, and has a "kink" (non-differentiable) at that point.

4 First-Order Conditions (FOC)

▷ **Handwritten Notes** (what the professor said)

Assuming $\mathbb{P}$ has a continuous probability function, we can ignore the single point of non-differentiability. By the Dominated Convergence Theorem, we can pass the derivative inside the expectation:

$$\begin{aligned}\frac{\partial \mathbb{E}(|\varepsilon - c|)}{\partial c} &= \mathbb{E}\left(\frac{\partial |\varepsilon - c|}{\partial c}\right) \\ &= \mathbb{E}\begin{cases}-1, & \varepsilon > c \\ 1, & \varepsilon < c\end{cases} \\ &= -1 \cdot \mathbb{P}(\varepsilon > c) + 1 \cdot \mathbb{P}(\varepsilon < c)\end{aligned}$$

Therefore, the First-Order Condition to find the critical point c^* that minimizes the function is:

$$\left.\frac{\partial \mathbb{E}(|\varepsilon - c|)}{\partial c}\right|_{c=c^*} = 0$$
$$\iff -\mathbb{P}(\varepsilon > c^*) + \mathbb{P}(\varepsilon < c^*) = 0$$

$$\iff \mathbb{P}(\varepsilon < c^*) = \mathbb{P}(\varepsilon > c^*)$$

Because we assume $\mathbb{P}$ has a continuous density function (meaning $\mathbb{P}(\varepsilon = c^*) = 0$), we can write:

$$\mathbb{P}(\varepsilon \leq c^*) = \mathbb{P}(\varepsilon \geq c^*)$$

Using the complement rule $\mathbb{P}(\varepsilon \geq c^*) = 1 - \mathbb{P}(\varepsilon \leq c^*)$:

$$\mathbb{P}(\varepsilon \leq c^*) = 1 - \mathbb{P}(\varepsilon \leq c^*) \quad \iff \quad 2\mathbb{P}(\varepsilon \leq c^*) = 1 \quad \iff \quad \mathbb{P}(\varepsilon \leq c^*) = \frac{1}{2}$$

This proves that the constant c^* that minimizes the expected absolute deviation is exactly the **median** of the distribution.

Key Result

The Median Minimises Expected Absolute Deviation
For any random variable ε with median ζ:

$$\boxed{\zeta = \arg\min_{c \in \mathbb{R}} \mathbb{E}[|\varepsilon - c|]}$$

In words: The median is the value that minimises the expected absolute deviation from the random variable.

Compare with the mean: The mean $\mu = \mathbb{E}[\varepsilon]$ minimises the expected *squared* deviation:

$$\mu = \arg\min_{c \in \mathbb{R}} \mathbb{E}[(\varepsilon - c)^2]$$

■ Proof

Formal proof using subdifferentials
Since $|\varepsilon - c|$ is not differentiable at $c = \varepsilon$, we use the subdifferential. The subdifferential of $|x|$ at $x = 0$ is the interval $[-1, 1]$.
For $c \neq \varepsilon$ a.s., we have:

$$\frac{\partial}{\partial c}\mathbb{E}[|\varepsilon - c|] = -\mathbb{E}[\text{sign}(\varepsilon - c)] = -\big(\mathbb{P}(\varepsilon > c) - \mathbb{P}(\varepsilon < c)\big)$$

Setting this to zero:

$$\mathbb{P}(\varepsilon > c) = \mathbb{P}(\varepsilon < c)$$

For a continuous distribution:

$$\mathbb{P}(\varepsilon \leq c) = 1 - \mathbb{P}(\varepsilon \leq c) \quad \Rightarrow \quad \mathbb{P}(\varepsilon \leq c) = \frac{1}{2}$$

Therefore $c = \text{Med}(\varepsilon)$. □

◇ **Student's Notes**

Geometric intuition:

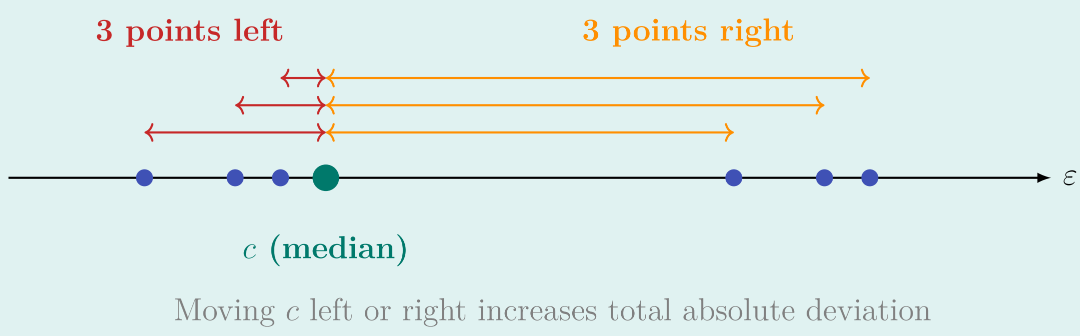


Figure 4: The median balances the number of observations on each side. Moving away from it increases the sum of absolute deviations because you move closer to fewer points than you move away from.

5 Returning to the Linear Model

▷ **Handwritten Notes** (what the professor said)

Let us return to the linear model where we have our correct specification:

$$Y_n = X_n\beta_0 + \varepsilon_n \iff Y_{(i)} = X_{(i)}\beta_0 + \varepsilon_{(i)} \quad \forall i = 1, \dots, n$$

We will make the following assumptions:

1. We assume the distribution of $\varepsilon_{(i)}$ given X_n is independent of i (Homogeneity).
2. We assume that the conditional median is zero:
$$\text{Med}(\varepsilon_{(i)} \,|\, X_n) = 0$$
3. We assume that the conditional distribution of $\varepsilon_{(i)}|X_n$ has a density function, say f^*, such that $f^*(0) > 0$.

Based on what we proved previously, these assumptions imply that:

$$\text{Med}(Y_{(i)} \,|\, X_n) = X_{(i)}\beta_0 \quad \forall i = 1, \dots, n$$

Consequently, searching for β_0 is equivalent to searching for the conditional median $\text{Med}(Y_{(i)}|X_n)$ for all $i = 1, \dots, n$.

Key Result

LAD Model Assumptions
Model: $Y_i = X_i'\beta_0 + \varepsilon_i$ for $i = 1, \dots, n$, with:

$$
\begin{array}{lll}
\textbf{(LAD1)} & \text{Med}(\varepsilon_i|X_n) = 0 & \text{(median zero)} \\
\textbf{(LAD2)} & \varepsilon_i|X_n \sim F \ \ \forall i & \text{(homogeneity)} \\
\textbf{(LAD3)} & f(0|X_n) > 0 & \text{(density} > 0 \text{ at zero)}
\end{array}
$$

$\Rightarrow$ $\text{Med}(Y_i|X_n) = X_i'\beta_0$
OLS comparison: LAD1 becomes $\mathbb{E}[\varepsilon_i|X_n] = 0$.

◇ Student's Notes

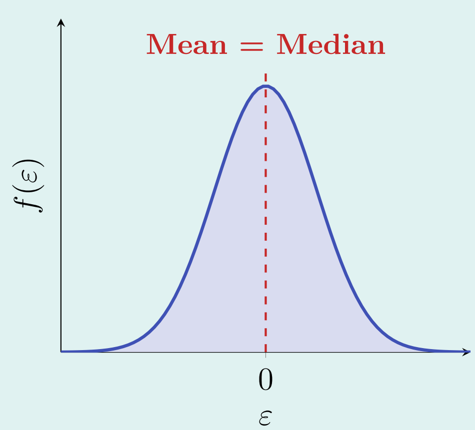
Why these assumptions?

Assumption	Role
$\text{Med}(\varepsilon_i\|X_n) = 0$	Makes $X'\beta_0$ the conditional median of Y
Homogeneity	All observations contribute equally; simplifies asymptotics
$f(0\|X_n) > 0$	Guarantees uniqueness of median; needed for consistency

When would we prefer LAD over OLS?

- **Heavy-tailed errors:** When $\mathbb{E}[\varepsilon^2] = \infty$, OLS is inefficient but LAD still works
- **Asymmetric errors:** When $\mathbb{E}[\varepsilon|X] \neq 0$ but $\text{Med}(\varepsilon|X) = 0$
- **Outlier-prone data:** Wages, asset returns, etc.

OLS: $\mathbb{E}[\varepsilon] = 0$



LAD: Med$(\varepsilon) = 0$

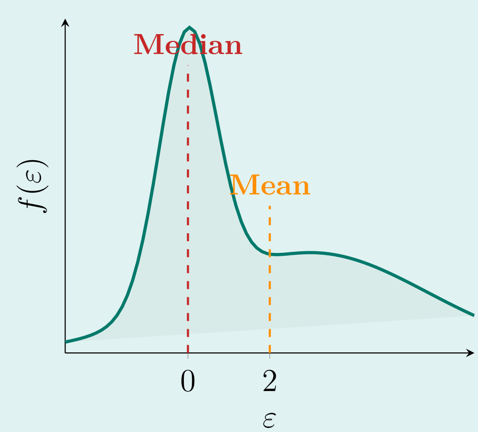


Figure 5: OLS assumes symmetric errors (left) where mean equals median. LAD allows skewed distributions (right) where the mean is pulled toward the tail but the median remains at zero.

6 Analysing the Objective Function

▷ Handwritten Notes (what the professor said)

Let us examine the expected value of our proposed absolute deviation objective function:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_{(i)} - X_{(i)}\beta| \;\middle|\; X_n\right)$$

Using the linearity of expectations:

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(|Y_{(i)} - X_{(i)}\beta| \,|\, X_n\big)$$

Substituting $Y_{(i)} = X_{(i)}\beta_0 + \varepsilon_{(i)}$:

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(|\varepsilon_{(i)} - X_{(i)}(\beta - \beta_0)| \,|\, X_n\big) \quad (*)$$

◇ Student's Notes

Key step: We have expressed the population objective function in terms of:

- The error $\varepsilon_{(i)}$, which has median zero
- The "shift" $X_{(i)}(\beta - \beta_0)$

This is analogous to the OLS derivation where we had $\mathbb{E}[(Y_i - X_i'\beta)^2|X] = \mathbb{E}[(\varepsilon_i - X_i'(\beta - \beta_0))^2|X]$.

Next: We show that $(*)$ is minimised at $\beta = \beta_0$ using the fact that the median minimises expected absolute deviation.

7 Minimisation and the True Parameter

▷ **Handwritten Notes** (what the professor said)

Let us define a function ψ inside the expectation:

$$\psi(c) = \mathbb{E}\big(|\varepsilon_{(i)} - c| \,|\, X_n\big)$$

Based on our assumptions (specifically that the conditional median of ε is zero), we know that $\psi(c)$ is uniquely minimized at $c = 0$.
Looking at equation $(*)$, the term acting as "c" is $X_{(i)}(\beta - \beta_0)$. Therefore, each individual term in the sum is minimized at any β^* for which:

$$X_{(i)}(\beta^* - \beta_0) = 0$$

Furthermore, when the full rank condition holds ($\text{rank}(X_n) = p$), then the *entire sum* is minimized uniquely at β_0.
Let's define the theoretical expected objective function as $M_n^*(\beta)$:

$$M_n^*(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big(|\varepsilon_{(i)} - X_{(i)}(\beta - \beta_0)| \,|\, X_n\big)$$

We have shown that:

$$\min_\beta M_n^*(\beta) = \frac{1}{n}\sum_{i=1}^{n} \min_\beta \mathbb{E}(\dots)$$

(Mathematical property note: generally $\min(A + B) \geq \min(A) + \min(B)$*, but here the minimum is achieved simultaneously for all* i *at* β_0*).*

■ **Proof**

Showing that $M_n^*(\beta)$ is uniquely minimised at β_0
From $(*)$:

$$M_n^*(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[|\varepsilon_i - X_i'(\beta - \beta_0)| \,|\, X_n\big]$$

For each i, define:

$$\psi_i(\beta) = \mathbb{E}\big[|\varepsilon_i - X_i'(\beta - \beta_0)| \,|\, X_n\big]$$

Since $\text{Med}(\varepsilon_i|X_n) = 0$ and $f(0|X_n) > 0$:

$$\arg\min_c \mathbb{E}[|\varepsilon_i - c| \,|\, X_n] = 0$$

Setting $c = X_i'(\beta - \beta_0)$, we need $X_i'(\beta - \beta_0) = 0$ for all i, i.e.:

$$X_n(\beta - \beta_0) = 0_{n\times 1}$$

Since $\text{rank}(X_n) = p$, this implies $\beta = \beta_0$ uniquely. Therefore:

$$\beta_0 = \arg\min_{\beta \in \Theta} M_n^*(\beta)$$

□

◇ Student's Notes

The identification argument in pictures:

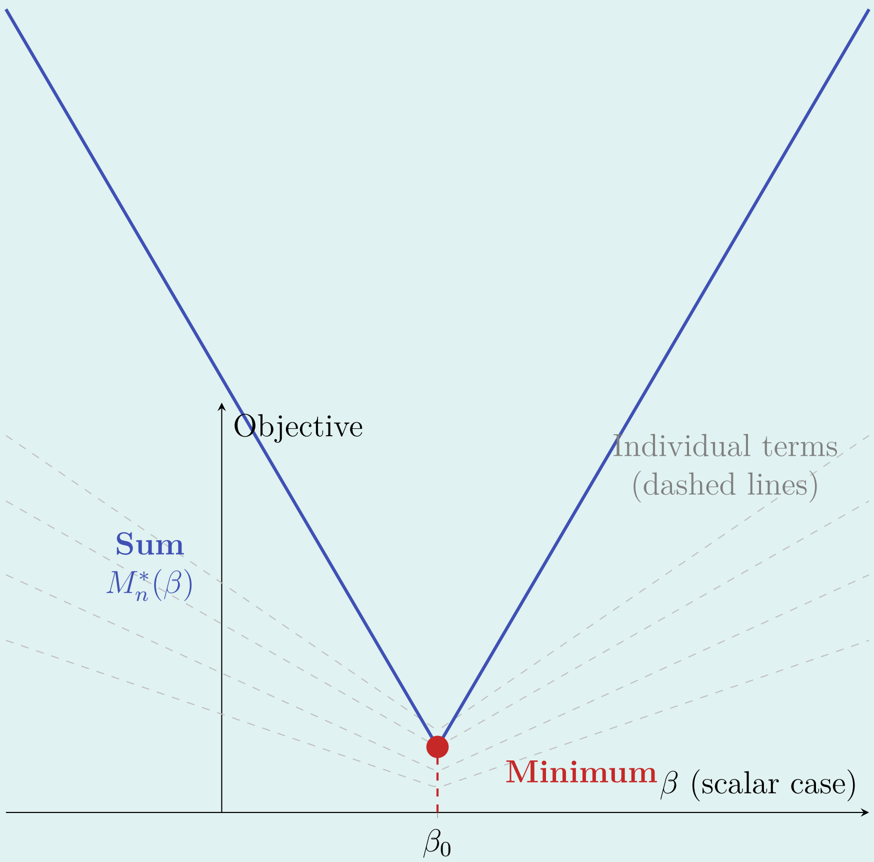


Figure 6: Each individual term $\mathbb{E}[|\varepsilon_i - X_i'(\beta - \beta_0)| \mid X_n]$ is minimised at $\beta = \beta_0$ (dashed). Their sum $M_n^*(\beta)$ (solid) is also minimised there. Under full rank, this is the unique global minimum.

Contrast with OLS:
For OLS, each term $\mathbb{E}[(Y_i - X_i'\beta)^2|X]$ is a *quadratic* in β, smooth everywhere. For LAD, each term is *piecewise linear* (V-shaped), with a kink at the optimum. This non-smoothness creates computational challenges.

8 The Sample LAD Estimator

▷ **Handwritten Notes** (what the professor said)

The theoretical function $M_n^*(\beta)$ is not observable because it depends on the unknown true parameter β_0.
Therefore, it is approximated by its sample analog:

$$M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} |Y_{(i)} - X_{(i)}\beta|$$

The M-estimator $\hat{\beta}_n$ defined by:

$$\hat{\beta}_n \in \arg\min_{\beta\in\Theta} M_n(\beta)$$

is called the **Least Absolute Deviations (LAD) Estimator**.

Definition: LAD Estimator

The **Least Absolute Deviations (LAD) Estimator** is:

$$\boxed{\hat{\beta}_n^{LAD} \in \arg\min_{\beta\in\Theta} M_n(\beta) = \arg\min_{\beta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} |Y_i - X_i'\beta|}$$

Equivalent formulation: Minimise $\sum_{i=1}^{n} |Y_i - X_i'\beta|$ (the constant $1/n$ doesn't affect the argmin).

Key Result

Properties of the LAD Estimator

(1) Existence: Guaranteed when Θ is compact (or coercivity holds)

(2) Uniqueness: Not guaranteed in general (objective is not strictly convex); sufficient condition: X has full column rank and no three observations are collinear

(3) Consistency: Under LAD1–LAD3 and mild regularity conditions, $\hat{\beta}_n^{LAD} \xrightarrow{p} \beta_0$

(4) Asymptotic normality: $\sqrt{n}(\hat{\beta}_n^{LAD} - \beta_0) \xrightarrow{d} N(0, V_{LAD})$, where V_{LAD} depends on $f(0|X)$

(5) Robustness: Less sensitive to outliers than OLS

◇ Student's Notes

Computational methods for LAD:
Since $M_n(\beta) = \sum_i |e_i|$ is not differentiable, standard optimisation methods don't apply directly.

Main approaches:

1. **Linear Programming (LP):**
   Introduce $u_i^+, u_i^- \geq 0$ with $e_i = u_i^+ - u_i^-$ and $|e_i| = u_i^+ + u_i^-$:

$$\min_{\beta, u^+, u^-} \sum_i (u_i^+ + u_i^-) \quad \text{s.t.} \quad Y_i - X_i'\beta = u_i^+ - u_i^-, \;\; u_i^\pm \geq 0$$

2. **Iteratively Reweighted Least Squares (IRLS):** Approximate $|e_i| \approx \sqrt{e_i^2 + \delta}$ and iterate.
3. **Subgradient methods:** Use $\partial |e_i| = \text{sign}(e_i)$ in a descent algorithm.

Software: R: `quantreg::rq()`. Python: `statsmodels`.

! Watch Out

Non-uniqueness of LAD estimates
Unlike OLS (where uniqueness is guaranteed by strict convexity when X has full rank), LAD can have **multiple solutions**.

When does this happen?

- When $n = p$ exactly: any hyperplane through p data points minimises $\sum |e_i|$
- When data lie in special configurations (collinearity)

In practice: Software returns *one* solution (typically a vertex of the LP feasible region).

Quick-Reference Summary

◇ Student's Notes

Lecture 8 narrative arc:

Topic	What was accomplished
Median definition	Defined as ζ with $\mathbb{P}(\varepsilon \leq \zeta) = 1/2$; discussed uniqueness
Key theorem	Proved: median minimises $\mathbb{E}[\lvert\varepsilon - c\rvert]$
LAD model	Linear model with $\text{Med}(\varepsilon \mid X) = 0$
Population objective	Showed $M_n^*(\beta)$ uniquely minimised at β_0
LAD estimator	$\hat{\beta}^{LAD} = \arg\min M_n(\beta)$

Key formulas:

Concept	Formula
Median	Any ζ with $F(\zeta) = 1/2$
Median minimises	$\text{Med}(\varepsilon) = \arg\min_c \mathbb{E}[\lvert\varepsilon - c\rvert]$
Population objective	$M_n^*(\beta) = \frac{1}{n}\sum_i \mathbb{E}[\lvert\varepsilon_i - X_i'(\beta - \beta_0)\rvert \mid X_n]$
Sample objective	$M_n(\beta) = \frac{1}{n}\sum_i \lvert Y_i - X_i'\beta\rvert$
LAD estimator	$\hat{\beta}^{LAD} = \arg\min_\beta M_n(\beta)$

Figures in this lecture:

Fig.	Content
1	Mean vs. median in skewed distributions
2	CDF showing non-unique median (discrete case)
3	V-shaped objective $\psi(c) = \mathbb{E}[\lvert\varepsilon - c\rvert]$
4	Geometric intuition: median balances deviations
5	Symmetric vs. skewed errors (OLS vs. LAD)
6	Sum of absolute deviation functions $M_n^*(\beta)$

OLS vs. LAD comparison:

Feature	OLS	LAD
Loss function	$(Y - X'\beta)^2$	$\lvert Y - X'\beta \rvert$
Assumption	$\mathbb{E}[\varepsilon \lvert X] = 0$	$\text{Med}(\varepsilon \lvert X) = 0$
Target	Conditional mean	Conditional median
Outlier sensitivity	High	Low
Closed form	Yes	No
Computation	Matrix algebra	Linear programming
Uniqueness	Guaranteed	Not guaranteed