Lecture 7: The Just-Identified Case & Introduction to LAD

Econometrics 2 — *When $q = p$: Simplifications and Constraints*

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

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

◇ **Teal boxes = Student's Notes**

Recall from Lecture 6

In Lecture 6 we established the **IV estimator** framework:

$$\hat{\beta}_n = (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' Y_n$$

where Z_n is an $n \times q$ matrix of instruments, W is a $q \times q$ positive definite weighting matrix, and we require $q \geq p$ for identification. Key questions remained: *How does the estimator depend on W? When can we ignore W?*

1 Recap: The IV Estimator and Moment Conditions

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

The linear model with instrumental variables supports the system of moment conditions:

$$\mathbb{E}\left[\frac{1}{n} Z_n'(Y_n - X_n\beta)\right] = 0_{q\times 1} \quad (*)$$

Provided the relevant identification conditions hold, $(*)$ is satisfied if and only if $\beta = \beta_0$. Choosing to work with Mahalanobis-type squared distances and independently of the weighting matrix W, we arrived at the usual version of the Instrumental Variables Estimator (IVE):

$$\hat{\beta}_n \in \arg\min_{\beta\in\Theta} (Y_n - X_n\beta)' Z_n W Z_n' (Y_n - X_n\beta)$$

We previously questioned how the properties of $\hat{\beta}_n$ depend on the choice of W. By examining the case where $\Theta = \mathbb{R}^p$, the optimization is solved analytically:

$$\hat{\beta}_n = (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' Y_n$$

Since $Y_n = X_n\beta_0 + \varepsilon_n$, we substituted this to find:

$$\hat{\beta}_n = \beta_0 + (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' \varepsilon_n$$

This led to the conclusion that $\hat{\beta}_n$ is not generally unbiased.

◇ Student's Notes

The central question for today:
We established that in the **over-identified case** ($q > p$), the estimator $\hat{\beta}_n$ generally depends on W—different weighting matrices give different estimates.

But what about the just-identified case ($q = p$)?
This turns out to be special. When we have exactly as many instruments as parameters, something remarkable happens: the weighting matrix W *cancels out*, and all positive definite W give the same estimator!

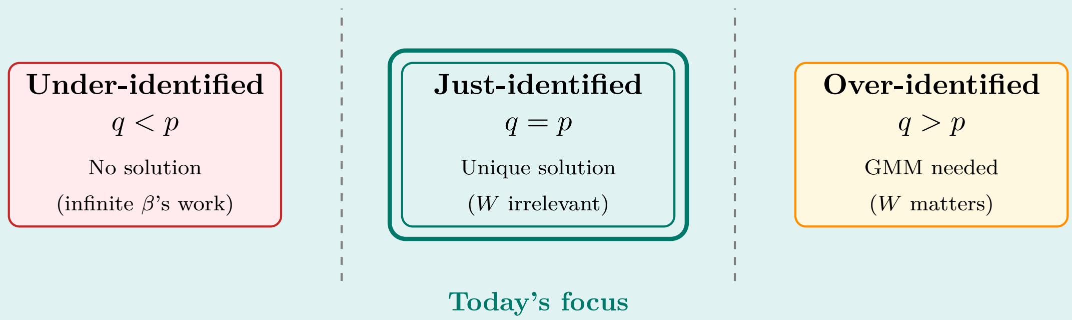

Figure 1: The three identification regimes. Today we focus on the just-identified case, where $q = p$ leads to major simplifications.

2 The Just-Identified Case ($q = p$)

▷ Handwritten Notes (what the professor said)

Let us investigate what happens to our estimator when $q = p$. This means the number of instrumental variables (q) perfectly coincides with the number of regressors (p).
Consider the matrix $X_n' Z_n$:

- Generally, it has dimensions $p \times q$.
- When $q = p$, it becomes a square $p \times p$ matrix.

By our general rank condition assumption, $\text{rank}(X_n' Z_n) = p$. Therefore, when $q = p$, the square matrix $X_n' Z_n$ is **invertible**. Consequently, its transpose $Z_n' X_n$ is also invertible.
Furthermore, the weighting matrix W is now $p \times p$. Since it is positive definite, it is also invertible.
These properties allow us to expand the inverse of the matrix product. Using the matrix property $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$:

$$\left((X_n' Z_n)W(Z_n' X_n)\right)^{-1} = (Z_n' X_n)^{-1} W^{-1} (X_n' Z_n)^{-1}$$

(Note: Each of these individual separated matrices is $p \times p$ and invertible).

◇ Student's Notes

Why invertibility changes everything:
When $q > p$, the matrix $X'Z$ is $p \times q$ (wide), and $Z'X$ is $q \times p$ (tall). Neither is square, so neither is invertible on its own. We *must* use the product $X'ZWZ'X$ (which is $p \times p$) and cannot simplify further.
When $q = p$, both $X'Z$ and $Z'X$ are $p \times p$ and (under our rank condition) invertible. This allows algebraic manipulation that was previously impossible.

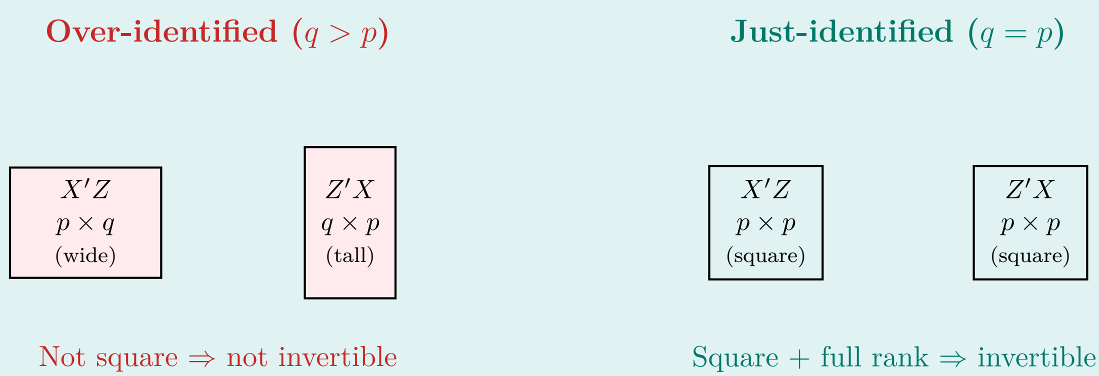

Figure 2: Matrix dimensions in the over-identified vs. just-identified cases. When $q > p$, $X'Z$ is wide (p rows, q columns) and $Z'X$ is tall (q rows, p columns)—neither is invertible. When $q = p$, both are square and invertible.

3 Simplification of the Estimator

▷ Handwritten Notes (what the professor said)

We can now substitute this expanded inverse back into the analytical formula for $\hat{\beta}_n$:

$$\hat{\beta}_n = \underbrace{(Z_n'X_n)^{-1}W^{-1}(X_n'Z_n)^{-1}}_{\text{Inverse part}}(X_n'Z_n)WZ_n'Y_n$$

Notice the immediate cancellations via identity matrices ($I_{p\times p}$):

- $(X_n'Z_n)^{-1}(X_n'Z_n) = I_{p\times p}$
- $W^{-1}W = I_{p\times p}$

This perfectly collapses the estimator down to:

$$\hat{\beta}_n = (Z_n' X_n)^{-1} Z_n' Y_n$$

Conclusion for $q = p$:
When $q = p$ and $\Theta = \mathbb{R}^p$, we find that $\hat{\beta}_n = (Z_n' X_n)^{-1} Z_n' Y_n$. Therefore, our estimator—and consequently the properties of the IVE—become **completely independent** of the choice of the weighting matrix W.

Key Result

The Just-Identified IV Estimator
When $q = p$ (number of instruments equals number of parameters) and $\Theta = \mathbb{R}^p$:

$$\boxed{\hat{\beta}_n^{IV} = (Z_n' X_n)^{-1} Z_n' Y_n}$$

Key property: This estimator is **independent of W**. Any positive definite weighting matrix gives exactly the same result.

Derivation summary:

$$\begin{aligned} \hat{\beta}_n &= (X'ZWZ'X)^{-1} X'ZWZ'Y \\ &= (Z'X)^{-1} W^{-1} (X'Z)^{-1} (X'Z) W Z'Y \\ &= (Z'X)^{-1} \underbrace{W^{-1} W}_{=I} Z'Y \\ &= (Z'X)^{-1} Z'Y \end{aligned}$$

⋄ Student's Notes

Intuition for why W disappears:
When $q = p$, we have exactly as many moment conditions as parameters. This means there is a *unique* β that sets all sample moments to exactly zero:

$$\frac{1}{n} Z'(Y - X\beta) = 0 \quad \Rightarrow \quad Z'Y = Z'X\beta \quad \Rightarrow \quad \beta = (Z'X)^{-1} Z'Y$$

Since we can achieve *exact* zero for the sample moments, we don't need to "trade off" between fitting different moments better or worse. The weighting matrix W determines how to make such trade-offs, but when no trade-off is needed, W becomes irrelevant.

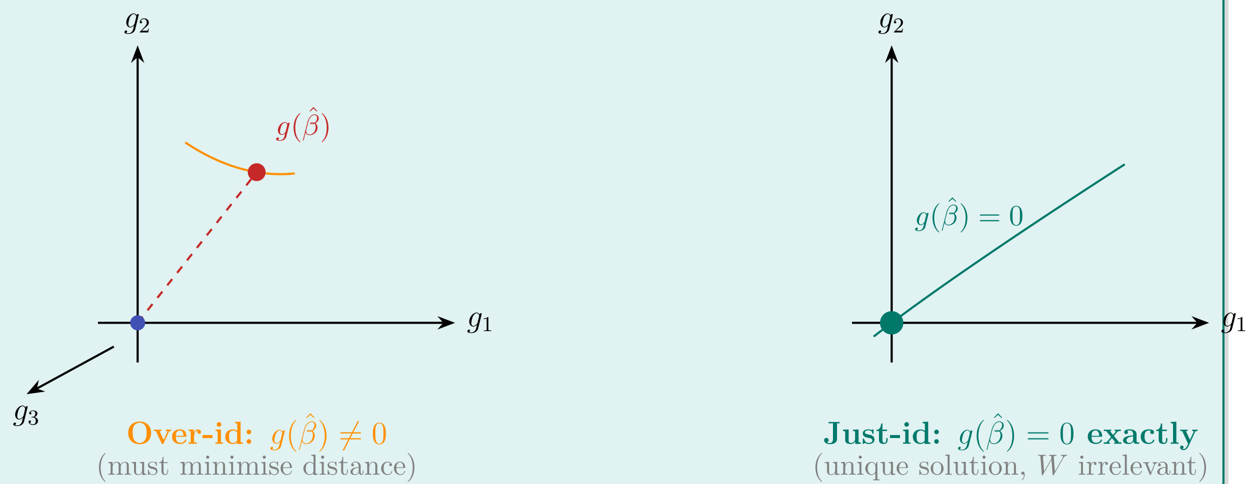

Figure 3: Geometric difference between over- and just-identified cases. In the over-identified case, we cannot achieve $g(\beta) = 0$ and must minimise distance (which depends on W). In the just-identified case, we achieve exact zero.

4 OLS as a Special Case of IV

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

What if, additionally, the instrumental variables simply coincide with the regressors themselves? (i.e., $Z_n = X_n$).
(Note: This relies on the core exogeneity assumption $\mathbb{E}(\varepsilon_n|X_n) = 0_{n\times 1}$).
Substituting $Z_n = X_n$ into our simplified just-identified estimator yields:

$$\hat{\beta}_n = (X_n' X_n)^{-1} X_n' Y_n$$

This is exactly the Ordinary Least Squares Estimator (OLSE). Therefore, for $\Theta = \mathbb{R}^p$, we see that the Least Squares estimator becomes a special case of the instrumental variables estimator.
Caveat: If we restrict the parameter space such that $\Theta \subset \mathbb{R}^p$ (i.e., we operate in some smaller, constrained space), we can no longer guarantee this exact analytical outcome, as we do not know what happens with the inverse at the boundaries.

Key Result

OLS is a Special Case of IV

When instruments equal regressors ($Z = X$) and $q = p$:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'Y = (X'X)^{-1}X'Y = \hat{\beta}^{OLS}$$

Interpretation: Under exogeneity ($\mathbb{E}[\varepsilon|X] = 0$), the regressors X serve as their own instruments. IV with $Z = X$ reduces to OLS.

◇ Student's Notes

The hierarchy of estimators:

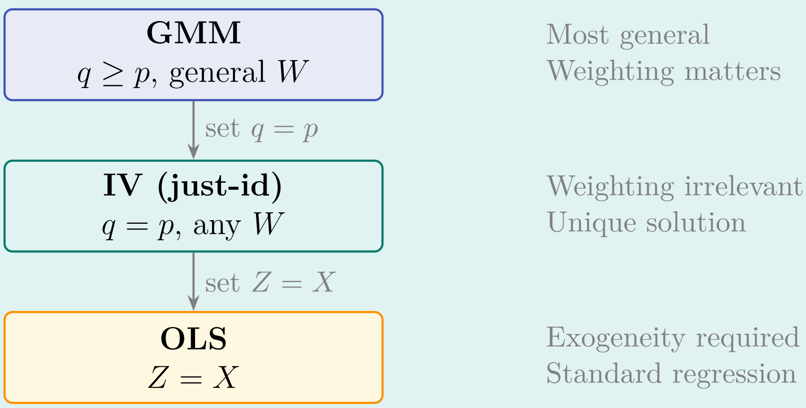

Figure 4: The nesting structure of estimators. OLS is a special case of just-identified IV, which is a special case of GMM.

When to use which:

Situation	q vs. p	Estimator	W matters?
Exogenous regressors	$q = p$	OLS	No
Endogenous, exact ID	$q = p$	IV	No
Endogenous, over-ID	$q > p$	GMM/2SLS	Yes
Endogenous, under-ID	$q < p$	—	Not identified

5 The Just-Identified Case in a Constrained Space

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

When $q = p$ but the parameter space is constrained ($\Theta \subset \mathbb{R}^p$), does the independence of the IV estimator from the choice of W still hold?
Let's define the *unconstrained* estimator as:

$$\hat{\beta}_n^{(un)} = (Z_n' X_n)^{-1} Z_n' Y_n$$

The residuals based on this unconstrained estimator are:

$$e_n = Y_n - X_n \hat{\beta}_n^{(un)} = \big(I_{n\times n} - X_n (Z_n' X_n)^{-1} Z_n'\big) Y_n$$

Now, let's examine the sample moment conditions evaluated at these residuals:

$$\frac{1}{n} Z_n' e_n = \frac{1}{n}\big(Z_n' - Z_n' X_n (Z_n' X_n)^{-1} Z_n'\big) Y_n$$

Since $Z_n' X_n (Z_n' X_n)^{-1} = I_{p\times p}$, this simplifies to:

$$\frac{1}{n} Z_n' e_n = \frac{1}{n}(Z_n' - Z_n') Y_n = O_{p\times 1}$$

(Note: The sample property $\frac{1}{n} Z_n' e_n = 0$ perfectly mirrors our theoretical moment condition $\frac{1}{n}\mathbb{E}(Z_n' \varepsilon_n) = 0_{p\times 1}$. Exercise: Does this exact zero hold if $q > p$?)

◇ **Student's Notes**

Key insight: The unconstrained IV estimator $\hat{\beta}^{(un)} = (Z'X)^{-1} Z'Y$ satisfies the sample moment conditions *exactly*:

$$Z'(Y - X\hat{\beta}^{(un)}) = 0$$

This is analogous to how OLS residuals are orthogonal to the regressors: $X'(Y - X\hat{\beta}^{OLS}) = 0$.

What happens when $\hat{\beta}^{(un)} \notin \Theta$?
If the unconstrained solution lies outside our parameter space, we cannot use it directly. We must find the *constrained* optimum, which will generally *not* set moments to zero.

▷ **Exercise**

Does $Z'e = 0$ hold when $q > p$?
Answer: No! When $q > p$:

- We have more moment conditions (q) than parameters (p)
- We cannot find β that sets all q moments to exactly zero

- The GMM estimator minimises a weighted sum of squared moments
- At the optimum, $Z'e \neq 0$ in general (only approximately zero)

This is precisely why the over-identified case requires a weighting matrix W—to determine how to "trade off" fitting different moments.

6 Reformulating the Objective Function

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

In the case we are examining, we have $\hat{\beta}_n \in \arg\min_{\beta\in\Theta} M_n(\beta)$. Let's substitute $Y_n - X_n\beta$ in the objective function using our unconstrained estimator:

$$\begin{aligned} Z_n'(Y_n - X_n\beta) &= Z_n'\big(Y_n - X_n\hat{\beta}_n^{(un)} + X_n\hat{\beta}_n^{(un)} - X_n\beta\big) \\ &= Z_n'\big(e_n + X_n(\hat{\beta}_n^{(un)} - \beta)\big) \\ &= Z_n'e_n + Z_n'X_n(\hat{\beta}_n^{(un)} - \beta) \end{aligned}$$

Since $Z_n'e_n = 0$, we are left with:

$$Z_n'(Y_n - X_n\beta) = Z_n'X_n(\hat{\beta}_n^{(un)} - \beta) \quad (*)$$

Consequently, because of $(*)$, our objective function becomes:

$$\begin{aligned} M_n(\beta) &= \frac{1}{n^2}(\hat{\beta}_n^{(un)} - \beta)'X_n'Z_nWZ_n'X_n(\hat{\beta}_n^{(un)} - \beta) \\ &= (\hat{\beta}_n^{(un)} - \beta)'\left(\frac{X_n'Z_n}{n}W\frac{Z_n'X_n}{n}\right)(\hat{\beta}_n^{(un)} - \beta) \end{aligned}$$

Let $A_n = \frac{X_n'Z_n}{n}W\frac{Z_n'X_n}{n}$. Since W is positive definite, we can use the Cholesky decomposition ($W = LL'$), which helps us prove that A_n is also a positive definite matrix. Therefore, for $q = p$, the estimator $\hat{\beta}_n$ can equivalently be expressed as minimizing a quadratic form of the Mahalanobis distance with respect to A_n over the set Θ:

$$\hat{\beta}_n \in \arg\min_{\beta\in\Theta}(\hat{\beta}_n^{(un)} - \beta)'A_n(\hat{\beta}_n^{(un)} - \beta) \quad (**)$$

Key Result

Equivalent Formulation of Constrained IV
When $q = p$ and $\Theta \subset \mathbb{R}^p$, the IV estimator solves:

$$\boxed{\hat{\beta}_n \in \arg\min_{\beta\in\Theta}(\hat{\beta}^{(un)} - \beta)'A_n(\hat{\beta}^{(un)} - \beta)}$$

where:

- $\hat{\beta}^{(un)} = (Z'X)^{-1}Z'Y$ is the unconstrained IV estimator
- $A_n = \frac{1}{n^2}X'ZWZ'X$ is positive definite (depends on W!)

Interpretation: Find the point in Θ closest to the unconstrained solution, using Mahalanobis distance with matrix A_n.

◇ **Student's Notes**

Geometric visualisation:

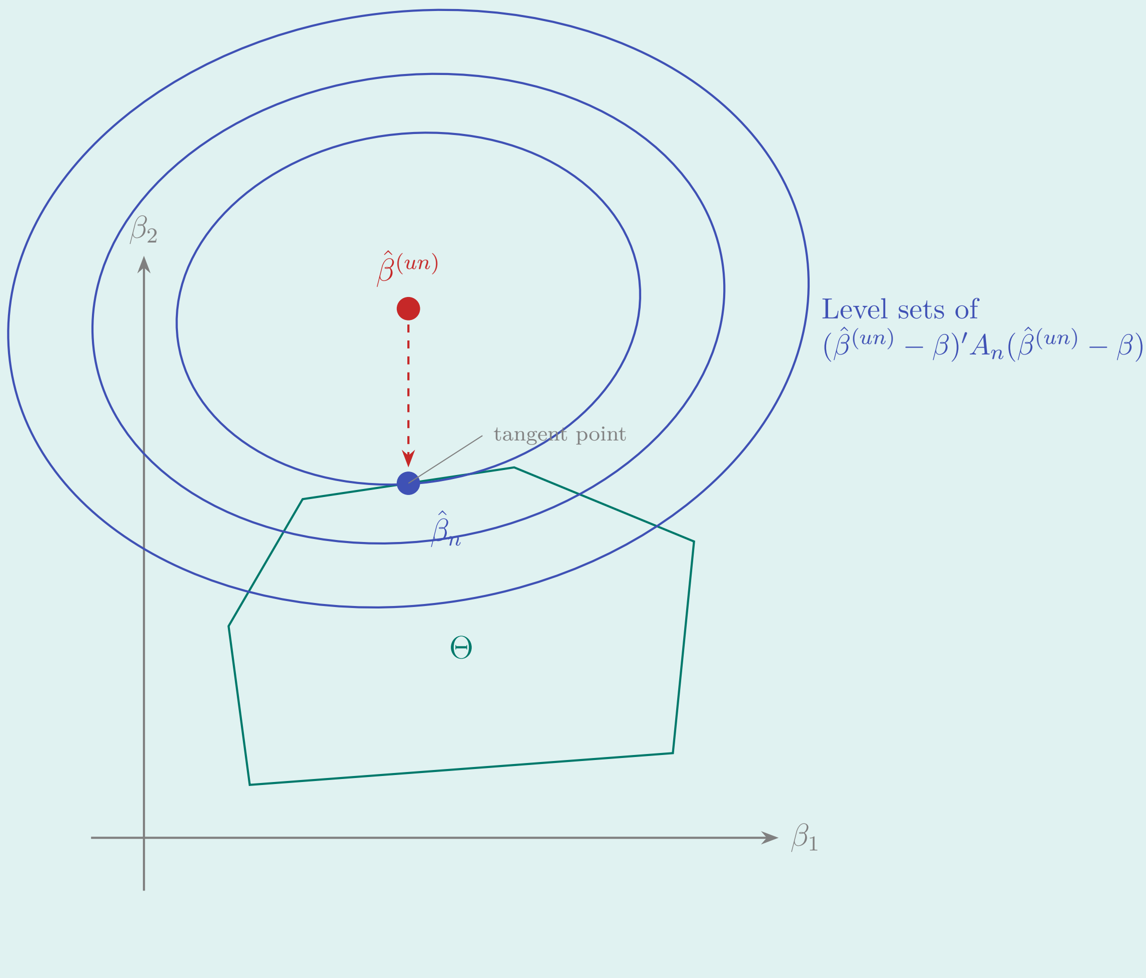

Figure 5: Constrained IV optimisation. The ellipses are level sets of the Mahalanobis distance centered at the unconstrained solution $\hat{\beta}^{(un)}$. The constrained estimator $\hat{\beta}_n$ is the point where the smallest ellipse is exactly tangent to the boundary of Θ—any smaller ellipse would not intersect Θ at all.

When does the unconstrained solution suffice?

If $\hat{\beta}^{(un)} \in \Theta$ (the unconstrained solution happens to satisfy the constraints), then $\hat{\beta}_n = \hat{\beta}^{(un)}$ regardless of A_n (and hence regardless of W).
Otherwise, the constrained estimator depends on the shape of the ellipses (determined by A_n), which in turn depends on W.

6.1 Dependence on W in the Constrained Case

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

From (∗∗), we observe that the objective function strictly depends on W (through A_n). Consequently, the estimator is generally **expected to depend on** W.
It will only be independent of W when the unconstrained minimum happens to fall inside our parameter space: $\hat{\beta}_n^{(un)} \in \Theta$. (For example, when $\Theta = \mathbb{R}^p$, which we already proved).
(Exercise: Show if we can form the Mahalanobis distance if $p > q$).

! Watch Out

Constrained IV: W can matter even when $q = p$!
The simple result "W doesn't matter when $q = p$" only holds for $\Theta = \mathbb{R}^p$.
With constraints ($\Theta \subset \mathbb{R}^p$):

- If $\hat{\beta}^{(un)} \in \Theta$: W still irrelevant (use unconstrained solution)
- If $\hat{\beta}^{(un)} \notin \Theta$: W matters! Different W give different constrained estimates

Practical example: Estimating a production function with non-negativity constraints on coefficients. If IV without constraints gives a negative coefficient, imposing $\beta_j \geq 0$ creates a constrained problem where W affects the result.

◇ **Student's Notes**

Summary table for the just-identified case:

Parameter space	$\hat{\beta}^{(un)} \in \Theta$?	W matters?
$\Theta = \mathbb{R}^p$	Always	No
$\Theta \subset \mathbb{R}^p$	Yes	No
$\Theta \subset \mathbb{R}^p$	No	**Yes**

7 Introduction to the LAD Estimator

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

LAD = Least Absolute Deviations

Recall that in the usual version of the Least Squares estimator, the objective function is:

$$M_n(\beta) = \frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_{(i)} - X_{(i)}\beta)^2$$

Let our matrices be defined as:

$$X_n = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

Where $X_{(i)} = (X_{i1}, X_{i2}, \ldots, X_{ip})$ is the i-th row of X_n, and $Y_{(i)}$ is the i-th element of $Y_n = (Y_1, \ldots, Y_n)'$.

Questions to ponder:

1. What would result if, as our objective function, we chose:

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

2. From what statistical model structure could such a choice be justified?

Definition: Least Absolute Deviations (LAD) Estimator

The **LAD estimator** minimises the sum of absolute residuals:

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

Compare with OLS, which minimises squared residuals:

$$\hat{\beta}_n^{OLS} \in \arg\min_{\beta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

◇ **Student's Notes**

OLS vs. LAD: The loss functions

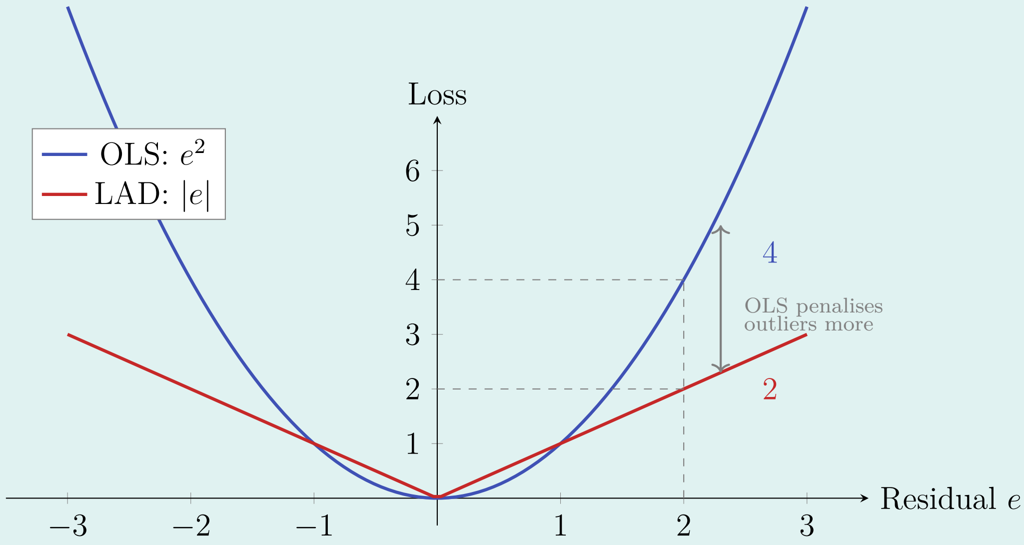

Figure 6: Squared loss (OLS) vs. absolute loss (LAD). For large residuals, squared loss grows much faster, making OLS more sensitive to outliers.

Key differences:

Property	OLS	LAD
Loss function	e^2 (quadratic)	$\|e\|$ (linear)
Sensitivity to outliers	High	Low (robust)
Differentiability	Smooth	Kink at 0
Optimisation	Closed-form	Linear programming
Target parameter	$\mathbb{E}[Y\|X]$ (mean)	Median$(Y\|X)$

★ **Intuition**

Why does LAD estimate the conditional median?
Consider the problem: find m that minimises $\mathbb{E}[|Y - m|]$.
Taking the derivative (using the fact that $\frac{d}{dm}|y - m| = -\text{sign}(y - m)$):

$$\frac{d}{dm}\mathbb{E}[|Y - m|] = -\mathbb{E}[\text{sign}(Y - m)] = P(Y < m) - P(Y > m)$$

Setting this to zero: $P(Y < m) = P(Y > m) = 0.5$, which defines $m = \text{Median}(Y)$.

For regression: LAD finds β such that $X'\beta$ equals the conditional median of Y given

X:

$$X'\beta^{LAD} = \text{Median}(Y \mid X)$$

In contrast, OLS targets the conditional mean:

$$X'\beta^{OLS} = \mathbb{E}[Y \mid X]$$

◇ Student's Notes

When to use LAD over OLS:

- **Heavy-tailed distributions:** When errors have heavier tails than normal (e.g., Cauchy, t with low df)
- **Outliers:** When data contains outliers that would unduly influence OLS
- **Median is of interest:** When the policy question concerns the median rather than the mean
- **Asymmetric distributions:** When mean $\neq$ median

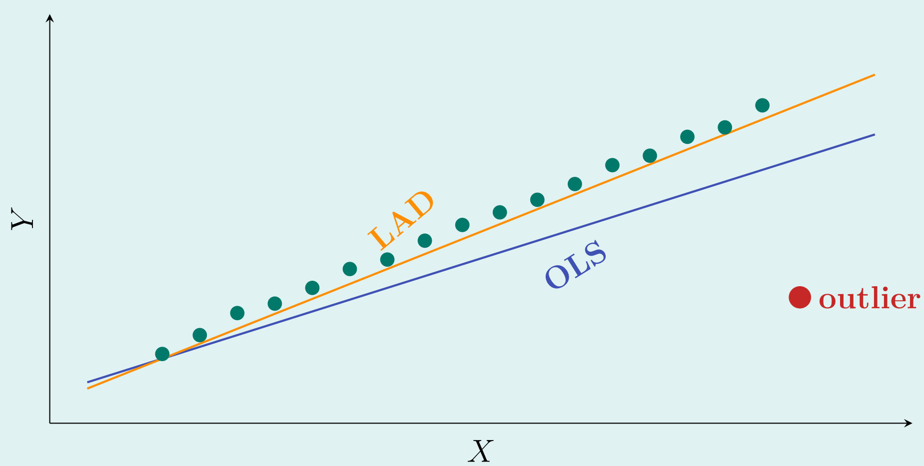

Figure 7: Effect of an outlier on OLS vs. LAD. The OLS line is pulled toward the outlier; the LAD line is much less affected because absolute deviations penalise large errors less severely than squared deviations.

◇ Student's Notes

Statistical justification: The Laplace distribution
Just as OLS can be derived from MLE under normal errors, LAD can be derived from

MLE under **Laplace (double-exponential)** errors.
The Laplace distribution has pdf:

$$f(\varepsilon) = \frac{1}{2b} \exp\left(-\frac{|\varepsilon|}{b}\right)$$

If $Y_i = X_i'\beta + \varepsilon_i$ with $\varepsilon_i \sim \text{Laplace}(0, b)$, the log-likelihood is:

$$\ell(\beta) = -n \log(2b) - \frac{1}{b} \sum_{i=1}^{n} |Y_i - X_i'\beta|$$

Maximising $\ell(\beta)$ is equivalent to minimising $\sum |Y_i - X_i'\beta|$, which is exactly the LAD criterion!

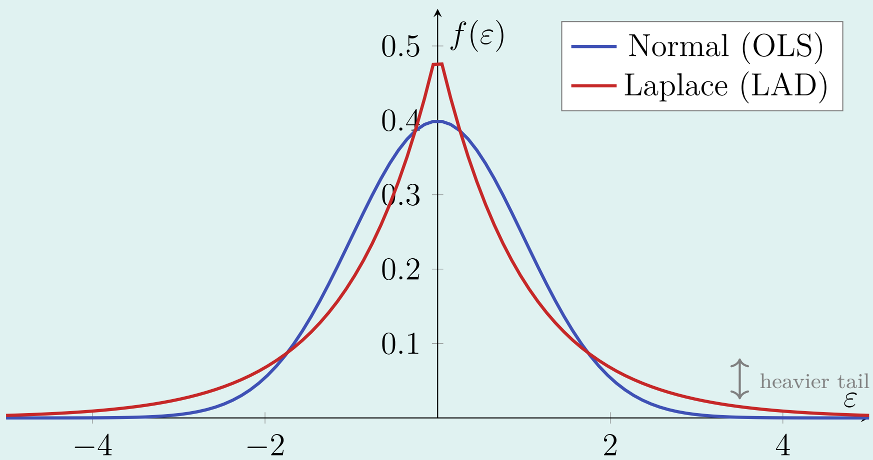

Figure 8: Normal vs. Laplace distributions. The Laplace has heavier tails (higher kurtosis) and a sharper peak at zero. LAD is the MLE under Laplace errors.

! Watch Out

Computational challenges with LAD
The absolute value function $|e|$ is **not differentiable** at $e = 0$. This means:

- Standard gradient-based methods (Newton–Raphson, BFGS) don't directly apply
- LAD is typically solved via **linear programming** (simplex algorithm)
- Software reformulates the problem using auxiliary variables

The LP reformulation:

Introduce $u_i^+, u_i^- \geq 0$ such that $Y_i - X_i'\beta = u_i^+ - u_i^-$. Then:

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

This is a linear program, solvable by standard LP algorithms.

Quick-Reference Summary

◇ **Student's Notes**

Lecture 7 narrative arc:

Topic	What was accomplished
Just-identified IV	When $q = p$, the estimator simplifies to $(Z'X)^{-1}Z'Y$, independent of W
OLS as special case	Setting $Z = X$ recovers OLS: $(X'X)^{-1}X'Y$
Constrained spaces	If $\hat{\beta}^{(un)} \notin \Theta$, then W matters even when $q = p$
Objective reformulation	Constrained IV minimises Mahalanobis distance from unconstrained solution
LAD introduction	Minimise $\sum |e_i|$ instead of $\sum e_i^2$; targets median, robust to outliers

Key formulas:

Estimator	Formula / Definition
Just-identified IV ($q = p$)	$\hat{\beta} = (Z'X)^{-1}Z'Y$
OLS (special case: $Z = X$)	$\hat{\beta} = (X'X)^{-1}X'Y$
Constrained IV	$\min_{\beta\in\Theta}(\hat{\beta}^{(un)} - \beta)'A_n(\hat{\beta}^{(un)} - \beta)$
LAD	$\min_\beta M_n(\beta) = \frac{1}{n}\sum_i |Y_i - X_i'\beta|$

Figures in this lecture:

Fig.	Content
1	Three identification cases (focus on just-identified)
2	Matrix dimensions: when $X'Z$ becomes invertible
3	Geometry: over-id vs. just-id moment conditions
4	Estimator hierarchy: GMM $\supset$ IV $\supset$ OLS
5	Constrained optimisation as projection onto Θ
6	Squared vs. absolute loss functions
7	Outlier robustness: OLS vs. LAD
8	Normal vs. Laplace error distributions