

ECON220B Discussion Section 2

Geometric Properties of OLS

Lapo Bini

UC San Diego

June 12, 2026

PREDICTION PROBLEM

► Prediction Problem

We observe a vector of outcome $\mathbf{y} \in \mathbb{R}^n$ and a matrix of covariates $\mathbf{X} \in \mathbb{R}^{n \times d}$. We want to find a function of the data that gives the **best prediction** of $\mathbf{y}$. How?

- (i) We define the **prediction errors**: $\mathbf{e} := \mathbf{y} - f(\mathbf{X})$ in terms of a generic function $f(\cdot)$
- (i) We specify the **loss function**: $L(\mathbf{e}) := \mathbf{e}'\mathbf{e}$
- (i) We minimize the **expected loss** $\mu(\mathbf{X}) := \arg \min_{f(\mathbf{X})} \mathbb{E}[(\mathbf{y} - f(\mathbf{X}))'(\mathbf{y} - f(\mathbf{X}))]$

► Conditional Expectation

The solution to the prediction problem is the **conditional expectation** $\mu(\mathbf{X}) = \mathbb{E}[\mathbf{y} \mid \mathbf{X}]$, that is the **best mean square predictor** of $\mathbf{y}$ given $\mathbf{X}$.

CONDITIONAL EXPECTATION

► Projection Interpretation

The conditional expectation can be interpreted as a **projection** of $\mathbf{y}$ onto the space of all square-integrable functions of $\mathbf{X}$. The following decomposition always holds:

$$\mathbf{y} = \mathbb{E}[\mathbf{y} \mid \mathbf{X}] + \mathbf{u},$$

► Key Properties

$\mathbb{E}[\mathbf{y} \mid \mathbf{X}]$ is a **random variable** that is measurable with respect to the σ -field generated by $\mathbf{X}$.

$\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$ by construction, since $\mathbf{u}$ is what's left after using all the information in $\mathbf{X}$.

► Implications

(i) $\mathbb{E}[\mathbf{u}] = \mathbf{0}$

(ii) $\mathbb{E}[\mathbf{X}'\mathbf{u}] = \mathbf{0}$

(iii) $\mathbb{E}[g(\mathbf{X})'\mathbf{u}] = \mathbf{0}$ for any function $g(\cdot)$

POPULATION LINEAR REGRESSION

► Expanding the Framework

Let $\mathbf{Z} \in \mathbb{R}^{n \times k}$ be **additional covariates**, then $\mathbb{E}[\mathbf{u} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{0} \Rightarrow \mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$

We also impose the **linearity** of the conditional expectation:

$$\mathbb{E}[\mathbf{y} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}$$

► Partialling-Out Theorem

Define the residualized variables: $\mathbf{R} := \mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Z}]$ and $\mathbf{v} := \mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{Z}]$, under the linearity of the conditional expectation $\mathbb{E}[\mathbf{y} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}$, then:

$$\boldsymbol{\beta} = (\mathbb{E}[\mathbf{R}'\mathbf{R}])^{-1}\mathbb{E}[\mathbf{R}'\mathbf{v}]$$

Intuition: $\boldsymbol{\beta}$ measures the **partial correlation** between $\mathbf{X}$ and $\mathbf{y}$, after removing from both random variables all variation explained by $\mathbf{Z}$.

POPULATION LINEAR REGRESSION

Proof

Start from the model and take conditional expectations given $\mathbf{Z}$:

$$(i) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \mathbf{u}$$

$$(ii) \quad \mathbb{E}[\mathbf{y} \mid \mathbf{Z}] = \mathbb{E}[\mathbf{X} \mid \mathbf{Z}]\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}$$

Subtracting (ii) from (i) and using $\mathbb{E}[\mathbf{u} \mid \mathbf{Z}] = 0$:

$$\underbrace{\mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{Z}]}_{\mathbf{v}} = \underbrace{(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Z}])}_{\mathbf{R}}\boldsymbol{\beta} + \mathbf{u} \implies \mathbf{v} = \mathbf{R}\boldsymbol{\beta} + \mathbf{u}.$$

Premultiply by $\mathbf{R}'$ and take expectations:

$$\mathbb{E}[\mathbf{R}'\mathbf{v}] = \mathbb{E}[\mathbf{R}'\mathbf{R}]\boldsymbol{\beta} + \underbrace{\mathbb{E}[\mathbf{R}'\mathbf{u}]}_{=0} \implies \boldsymbol{\beta} = (\mathbb{E}[\mathbf{R}'\mathbf{R}])^{-1}\mathbb{E}[\mathbf{R}'\mathbf{v}].$$

The term $\mathbb{E}[\mathbf{R}'\mathbf{u}] = 0$ vanishes because given the assumption $\mathbb{E}[\mathbf{u} \mid \mathbf{X}, \mathbf{Z}] = 0$, by the law of iterated expectations $\mathbb{E}[\mathbf{R}'\mathbf{u}] = \mathbb{E}[\mathbf{R}'\mathbb{E}[\mathbf{u} \mid \mathbf{X}, \mathbf{Z}]] = 0$. ■

SAMPLE LINEAR REGRESSION

► Framework

We observe an outcome $\mathbf{y} \in \mathbb{R}^n$ and a matrix of **covariates** $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$, whose first column is a constant:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_1 & \cdots & \mathbf{x}_p \end{bmatrix}.$$

► Linear Regression

Goal: produce the best prediction of $\mathbf{y}$ under the sample average squared loss. We do so by using a **linear combination** of the columns of $\mathbf{X}$, with weights $\mathbf{b} \in \mathbb{R}^{p+1}$:

$$\mathbf{X}\mathbf{b} = 1 b_0 + \mathbf{x}_1 b_1 + \cdots + \mathbf{x}_p b_p.$$

The set of all such combinations is the **column span** of $\mathbf{X}$, $\mathcal{C}(\mathbf{X}) = \{\mathbf{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{p+1}\} \in \mathbb{R}^n$

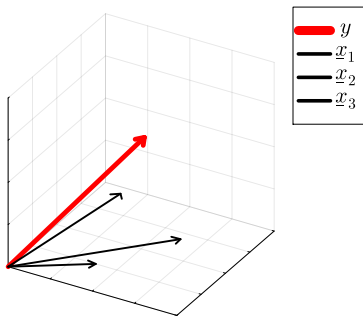
GEOMETRY OF THE PREDICTION

► Best Prediction

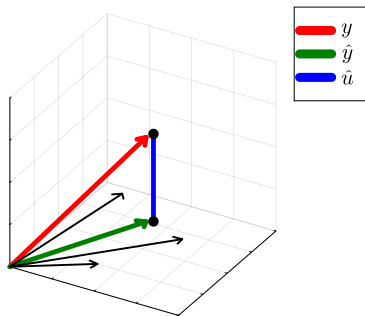
Assume $\mathbf{y}$ lies **outside** $\mathcal{C}(\mathbf{X})$. Our **best prediction** is the vector of fitted values $\hat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$ that is the closest to $\mathbf{y}$ (unique when $\mathbf{X}$ full column rank). Nice equivalence:

$$\hat{\mathbf{y}} := \arg \min_{\mathbf{c} \in \mathcal{C}(\mathbf{X})} \|\mathbf{y} - \mathbf{c}\|^2 \iff \hat{\boldsymbol{\beta}} := \arg \min_{\mathbf{b} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2.$$

Goal: best linear combination of them



Result: residual perpendicular to the plane



ORTHOGONALITY CONDITION

► Decomposition

We can split the observed $\mathbf{y}$ into the fitted value, our best prediction, plus the **residual**:

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

► Orthogonality

If $\hat{\mathbf{y}}$ is our best prediction, the **residual** must be **orthogonal** to the plane $\mathcal{C}(\mathbf{X})$, otherwise we could move a little inside the plane and get closer to $\mathbf{y}$. Then:

$$\hat{\mathbf{y}}'\hat{\mathbf{u}} = 0 \Rightarrow (\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{u}} = 0 \Rightarrow \hat{\boldsymbol{\beta}}'(\mathbf{X}'\hat{\mathbf{u}}) = 0 \Rightarrow \mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$$

(In the last step we rule out $\hat{\boldsymbol{\beta}} = \mathbf{0}$, otherwise $\hat{\boldsymbol{\beta}}'(\mathbf{X}'\hat{\mathbf{u}}) = \mathbf{0}$ for any residual.)

ORTHOGONALITY CONDITION

► Decomposition

Assume $\mathbf{X}$ has full column rank. If $\hat{\mathbf{y}}$ is the closest point in $\mathcal{C}(\mathbf{X})$ to $\mathbf{y}$, i.e.

$\hat{\mathbf{y}} = \arg \min_{\mathbf{c} \in \mathcal{C}(\mathbf{X})} \|\mathbf{y} - \mathbf{c}\|^2$, then the residual $\hat{\mathbf{u}} := \mathbf{y} - \hat{\mathbf{y}}$ satisfies $\mathbf{X}'\hat{\mathbf{u}} = 0$.

Proof

Any competitor in $\mathcal{C}(\mathbf{X})$ can be written as $\mathbf{y}^* = \hat{\mathbf{y}} + \mathbf{X}\boldsymbol{\delta}$, with $\boldsymbol{\delta} \in \mathbb{R}^{p+1}$. We argue **by contradiction**: suppose $\mathbf{X}'\hat{\mathbf{u}} \neq 0$ and build a competitor strictly closer to $\mathbf{y}$.

Measure the distance of such a competitor to $\mathbf{y}$:

$$\|\mathbf{y} - \mathbf{y}^*\|^2 = \|\hat{\mathbf{u}} - \mathbf{X}\boldsymbol{\delta}\|^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} + \underbrace{\boldsymbol{\delta}'\mathbf{X}'\mathbf{X}\boldsymbol{\delta} - 2\boldsymbol{\delta}'\mathbf{X}'\hat{\mathbf{u}}}_{=: Q(\boldsymbol{\delta})}.$$

$Q(\boldsymbol{\delta})$ is quadratic, with gradient $\nabla Q = 2\mathbf{X}'\mathbf{X}\boldsymbol{\delta} - 2\mathbf{X}'\hat{\mathbf{u}}$ and Hessian $2\mathbf{X}'\mathbf{X} \succ 0$, so it is **convex** with a unique minimizer

$$\boldsymbol{\delta}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{u}}, \quad Q(\boldsymbol{\delta}^*) = -\hat{\mathbf{u}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{u}} < 0.$$

Then $\|\mathbf{y} - \mathbf{y}^*\|^2 < \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, contradicting that $\hat{\mathbf{y}}$ is the closest point. Hence $\mathbf{X}'\hat{\mathbf{u}} = 0$. ■

FROM ORTHOGONALITY TO OLS

► Two Properties

Writing $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$ column by column, gives two nice properties: (i) the **residuals sum to zero**; (ii) they are **orthogonal to each column** of $\mathbf{X}$.

$$\mathbf{X}'\hat{\mathbf{u}} = \begin{bmatrix} 1'\hat{\mathbf{u}} \\ \underline{x}'_1\hat{\mathbf{u}} \\ \vdots \\ \underline{x}'_p\hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \sum_i \hat{u}_i \\ \sum_i x_{i1}\hat{u}_i \\ \vdots \\ \sum_i x_{ip}\hat{u}_i \end{bmatrix} = 0,$$

► The Old Estimator

Substituting $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ into $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$, we recover the familiar **OLS estimator**:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \implies \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

PROJECTION MATRICES

► Orthogonal Projection

We define the **orthogonal projection operator** $P_x := X(X'X)^{-1}X'$, which projects y onto $\mathcal{C}(X)$ to give the fitted values $\hat{y} = P_x y$. It keeps the columns of X and annihilates the residual:

$$P_x X = X, \quad P_x \hat{u} = 0.$$

► Annihilator Matrix

The **annihilator matrix** $M_x := I - P_x$ returns what is **left unexplained**, i.e. the residual $\hat{u} = M_x y$. It annihilates the columns of X and keeps the residual:

$$M_x X = 0, \quad M_x \hat{u} = \hat{u}.$$

► Properties

Both are **symmetric** and **idempotent**, and mutually orthogonal:

$$P_x' = P_x, \quad P_x^2 = P_x, \quad M_x' = M_x, \quad M_x^2 = M_x, \quad M_x P_x = 0.$$

FRISCH-WAUGH-LOVELL THEOREM

► Statement

Consider the full regression with two blocks of regressors: $y = X\beta + Z\delta + u$. The OLS coefficient $\hat{\beta}$ from the full regression is **identical** to the coefficient from a short regression on **residualized** variables:

$$\hat{\beta} = (X'M_Z X)^{-1} X' M_Z y = (\hat{R}' \hat{R})^{-1} \hat{R}' \hat{v},$$

where $\hat{R} = M_Z X$ is the residual from regressing X on Z , and $\hat{v} = M_Z y$ the residual from regressing y on Z .

► Reading

This is the **sample analogue** of the population partialling-out theorem: $\hat{\beta}$ captures the relation between y and X *after* netting Z out of both.

FRISCH-WAUGH-LOVELL THEOREM

Proof

Start from the fitted full regression and premultiply by M_z , using $M_z Z = 0$ and $Z' \hat{u} = 0 \Rightarrow M_z \hat{u} = \hat{u}$:

$$y = X\hat{\beta} + Z\hat{\delta} + \hat{u}$$
$$M_z y = M_z X\hat{\beta} + \underbrace{M_z Z}_{=0} \hat{\delta} + \underbrace{M_z \hat{u}}_{=\hat{u}}.$$

Premultiply by X' and use $X' \hat{u} = 0$:

$$X' M_z y = X' M_z X \hat{\beta} \implies \hat{\beta} = (X' M_z X)^{-1} X' M_z y.$$

Finally, since M_z is symmetric and idempotent ($M_z = M_z' M_z$):

$$X' M_z X = (M_z X)' (M_z X) = \hat{R}' \hat{R}, \quad X' M_z y = (M_z X)' (M_z y) = \hat{R}' \hat{v},$$

hence $\hat{\beta} = (\hat{R}' \hat{R})^{-1} \hat{R}' \hat{v}$. ■

GOODNESS OF FIT

► Diagnostic

Given our regression model, we want to study what **percentage** of the sample **variability** of y is **explained by the model**. Start from the argument of the sample variance:

$$S^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$$

$$y_i - \bar{y} = \hat{u}_i + (\hat{y}_i - \bar{y})$$

$$(y_i - \bar{y})^2 = \hat{u}_i^2 + (\hat{y}_i - \bar{y})^2 + 2 \hat{u}_i (\hat{y}_i - \bar{y})$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i \hat{u}_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \left(\sum_i \hat{u}_i \hat{y}_i - \bar{y} \sum_i \hat{u}_i \right)$$

The cross term vanishes since $\hat{y}'\hat{u} = 0$ and $\sum_i \hat{u}_i = 0$, leaving

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i \hat{u}_i^2}_{\text{RSS}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{ESS}} \implies R^2 := \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

$R^2 \in [0, 1]$ is the **fraction of the sample variability** of y explained by the model.

GOODNESS OF FIT

► Reading the R^2

The chart shows the data (x_i, y_i) , the fitted line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, the mean $\bar{y}$ (dashed), and the residuals $\hat{u}_i$ (gray)

