12: Ordinary Least Square

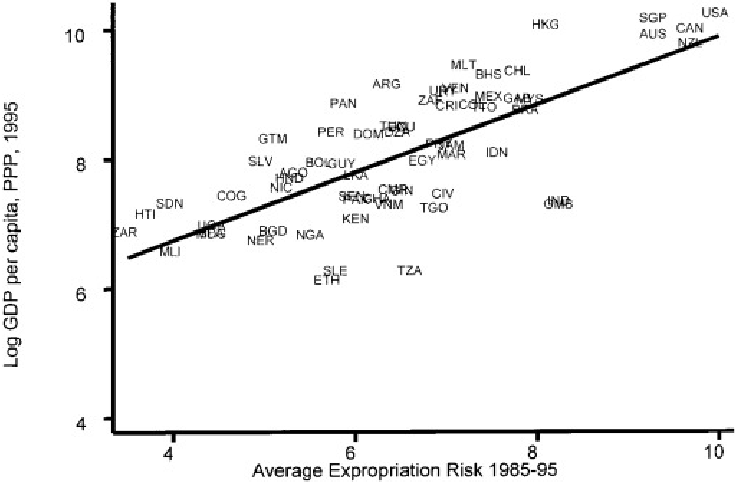
Naijia Liu

Spring 2025

Where are we? Where are we going?

- We saw how the population linear projection works.
- How can we estimate the parameters of the linear projection or CEF?
- Now: least squares estimator and its algebraic properties.
- After that: the statistical properties of least squares.

Acemoglu, Johnson and Robinson 2001



Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ is the **sample** and can be seen in two ways:
  - Numbers in your data matrix, fixed to the analyst.
  - From a statistical POV, they are realizations of a random process.
- Violations include time-series data and clustered sampling.
  - Weakening i.i.d. usually complicates notation but can be done.

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\beta + e$$

- Here β minimizes the **population** expected squared error:

$$\beta = \arg\min_{b \in \mathbb{R}^k} S(b), \quad S(b) = \mathbb{E}\left[\left(Y - \mathbf{X}'b\right)^2\right]$$

- Last time we saw that this can be written:

$$\beta = \left(\mathbb{E}[\mathbf{X}\mathbf{X}']\right)^{-1}\mathbb{E}[\mathbf{X}Y]$$

- How do we estimate β?

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i' b$
  - **Sum of squared residuals**, $SSR(b) = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' b)^2$
  - Total prediction error using b as our estimated coefficient.
- We can use these residuals to get a sample average prediction error:

$$\hat{S}(b) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' b)^2 = \frac{1}{n} SSR(b)$$

- $\hat{S}(b)$ is an estimator of the expected squared error, $S(b)$.

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\beta = \arg\min_{b \in \mathbb{R}^k} \mathbb{E}\left[\left(Y - \mathbf{X}'b\right)^2\right]$$

$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' b\right)^2$$

- In words: find the coefficients that minimize the sum/average of the squared residuals.
- After some calculus, we can write this as a plug-in estimator:

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i\right)$$

- $\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is the sample version of $\mathbb{E}[\mathbf{X}\mathbf{X}']$
- $\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i Y_i$ is the sample version of $\mathbb{E}[\mathbf{X}Y]$

Bivariate regressions

- **Bivariate regression** is the linear projection model with $\mathbf{X} = (1, X)$:

$$Y = \beta_0 + X\beta_1 + e$$

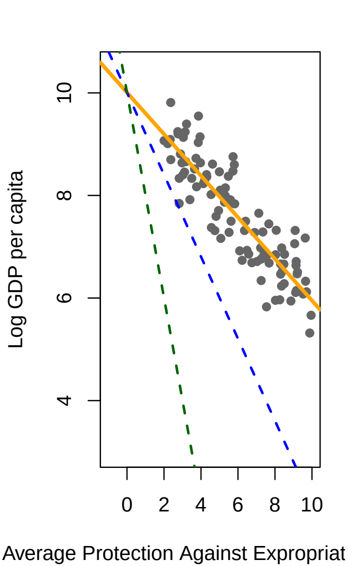
- Linear projection slope in the population from last times:

$$\beta_1 = \frac{\mathsf{Cov}(X, Y)}{\mathbb{V}[X]}$$

- We can show the OLS estimator of the slope is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\widehat{\mathsf{Cov}}(X, Y)}{\widehat{\mathbb{V}}[X]}$$

Visualizing Regression



Residuals

- **Fitted value** $\hat{Y}_i = \mathbf{X}_i'\hat{\beta}$ is what the model predicts at $\mathbf{X}_i$
  - Not really a prediction for Y_i since that was used to generate $\hat{\beta}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}_i'\hat{\beta}$$

  - We can write $Y_i = \mathbf{X}_i'\hat{\beta} + \hat{e}_i$
  - $\hat{e}_i$ are not the true errors e_i
- Key **mechanical properties** of OLS residuals:

$$\sum_{i=1}^{n} \mathbf{X}_i \hat{e}_i = 0$$

- Sample covariance between $\mathbf{X}_i$ and $\hat{e}_i$ is 0.
- If $\mathbf{X}_i$ has a constant, then $n^{-1}\sum_{i=1}^{n} \hat{e}_i = 0$

Prediction error

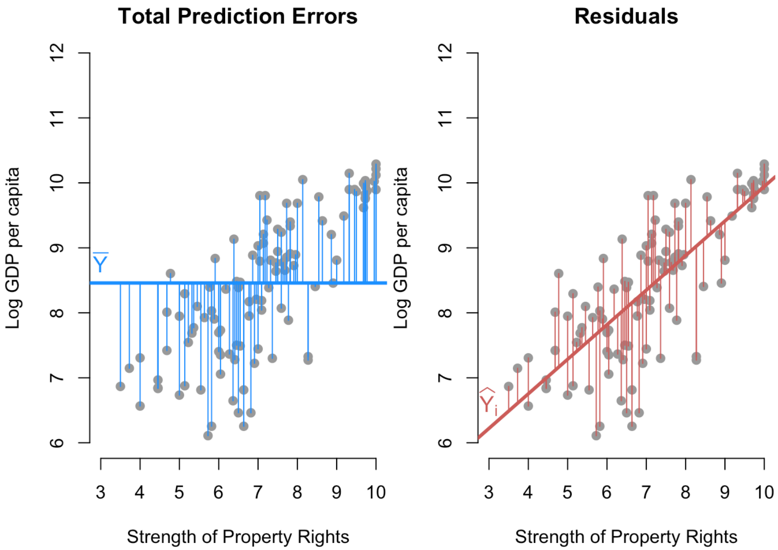
- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
  - Best prediction is the mean, $\bar{Y}$
  - Prediction error is called the total sum of squares (TSS) and would be:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **Prediction errors with $\mathbf{X}_i$:**
  - Best predictions are the fitted values, $\hat{Y}_i$
  - Prediction error is the sum of the squared residuals or SSR:

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

TSS and SSR


Total Prediction Errors
Log GDP per capita
6
7
8
9
10
11
12
$\overline{Y}$
3
4
5
6
7
8
9
10
Strength of Property Rights
Residuals
Log GDP per capita
6
7
8
9
10
11
12
$\widehat{Y}_i$
3
4
5
6
7
8
9
10
Strength of Property Rights

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$
- **Common interpretation**: R^2 is the fraction of the variation in Y_i "explained by" $\mathbf{X}_i$:
  - $R^2 = 0$ means no relationship
  - $R^2 = 1$ implies perfect linear fit
- Mechanically increases with additional covariates (better fit measures exist)

Linear model in matrix form

- Linear model is a system of n linear equations:

$$\begin{aligned} Y_1 &= \mathbf{X}_1'\beta + e_1 \\ Y_2 &= \mathbf{X}_2'\beta + e_2 \\ &\vdots \\ Y_n &= \mathbf{X}_n'\beta + e_n \end{aligned}$$

- We can write this more compactly using matrices and vectors:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

- Model is now just:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbf{X}'\mathbf{X}, \quad \sum_{i=1}^{n} \mathbf{X}_i Y_i = \mathbf{X}'\mathbf{Y}$$

- Implies we can write the OLS estimator as:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- Residuals:

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} 1\hat{\beta}_0 + X_{11}\hat{\beta}_1 + X_{12}\hat{\beta}_2 + \cdots + X_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + X_{21}\hat{\beta}_1 + X_{22}\hat{\beta}_2 + \cdots + X_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + X_{n1}\hat{\beta}_1 + X_{n2}\hat{\beta}_2 + \cdots + X_{nk}\hat{\beta}_k \end{bmatrix}$$

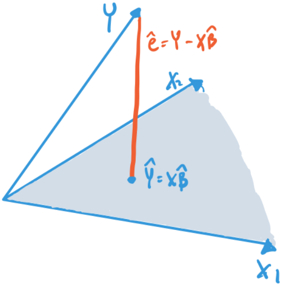
Least squares in matrix form

- OLS still minimizes sum of the squared residuals

$$\arg \min_{\mathbf{b} \in \mathbb{R}^{k+1}} \hat{\mathbf{e}}'\hat{\mathbf{e}} = \arg \min_{\mathbf{b} \in \mathbb{R}^{k+1}} (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

- We can write the covariate-residual orthogonality as $\mathbf{X}'\hat{\mathbf{e}} = 0$.

Projection



- OLS can be seen as a projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$, $\mathcal{S}(\mathbf{X})$.
  - Picture with $n = 3$ and $k = 2$: points in 3D space,
  - Column space of $\mathbf{X}$ is a plane in this space.
- Intuition: $\hat{\beta}$ defines the projection that gets is shortest distance between $\mathbf{Y}$ and prediction.

Projection/hat matrix

- We can define the transformation of $\mathbf{Y}$ that does the projection:

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- **Projection matrix**

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- Also called the **hat matrix**; it puts the "hat" on $\mathbf{Y}$:

$$\mathbf{P}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\beta} = \hat{\mathbf{Y}}$$

- Key properties:
    - $\mathbf{P}$ is an $n \times n$ symmetric matrix
    - $\mathbf{P}$ is **idempotent**: $\mathbf{P}\mathbf{P} = \mathbf{P}$
    - Projecting $\mathbf{X}$ onto itself returns itself: $\mathbf{P}\mathbf{X} = \mathbf{X}$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:
$$\mathbf{M} = I_n - \mathbf{P} = I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
- Also called the **residual maker**:
$$\mathbf{MY} = (I_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \hat{\mathbf{Y}} = \hat{\mathbf{e}}$$
- "Annihilates" any function in the column space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X})$:
$$\mathbf{MX} = (I_n - \mathbf{P})\mathbf{X} = \mathbf{X} - \mathbf{PX} = \mathbf{X} - \mathbf{X} = 0$$
- Properties:
    - $\mathbf{M}$ is a symmetric $n \times n$ matrix and is idempotent: $\mathbf{MM} = \mathbf{M}$
    - Admits a nice expression for the residual vector: $\hat{\mathbf{e}} = \mathbf{Me}$
- Allows the following orthogonal partition:
$$\mathbf{Y} = \mathbf{PY} + \mathbf{MY} = \text{projection} + \text{residual}$$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^2 + \cdots + \hat{a}_n^2}$
- Distance between two vectors:
  $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^{k+1}} \|\mathbf{Y} - \mathbf{X}b\|^2 = \arg \min_{b \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' b)^2$$

- Let $\mathcal{C}(\mathbf{X}) = \{\mathbf{X}b : b \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbf{X}$:
  - All n-vectors formed as a linear combination of the columns of $\mathbf{X}$
  - $k + 1$-dimensional subspace of $\mathbb{R}^n$
  - This is the space that OLS is searching over!
- Geometrically OLS is:
  - Find coefficients that minimize distance between the $\mathbf{Y}$ and $\mathbf{X}b$
  - Find the point in $\mathcal{C}(\mathbf{X})$ that is closest to $\mathbf{Y}$

Projection

- Finding closest point in $\mathcal{C}(\mathbf{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.
  - Column space of $\mathbf{X}$ is a plane in this space.
- Residual vector $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ is **orthogonal** to $\mathcal{C}(\mathbf{X})$
  - Shortest distance from $\mathbf{Y}$ to $\mathcal{C}(\mathbf{X})$ is a straight line to the plane, which will be perpendicular to $\mathcal{C}(\mathbf{X})$.
  - Implies that $\mathbf{X}'\hat{\mathbf{e}} = 0$

Multicollinearity

- Hidden assumption: $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'$ is invertible.
  - Equivalent to $\mathbf{X}$ being **full column rank**.
  - Equivalent to columns of $\mathbf{X}$ being **linearly independent**.
- Full column rank if $\mathbf{X}b = 0$ if and only if $b = 0$.

$$b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \cdots + b_{k+1}\mathbf{X}_{k+1} = 0 \quad \Longleftrightarrow \quad b_1 = b_2 = \cdots = b_{k+1} = 0$$

- Typically reasonable but can be violated by user error:
  - Accidentally adding the same variable twice.
  - Including all dummies for a categorical variable.
  - Including fixed effects for group and variables that do not vary within groups.