Gov 2001 Section 10, 2025

This section focuses on the OLS estimator. We observe data $(\mathbf{X}_{1:n}, Y_{1:n})$, $(\mathbf{x}_{1:n}, y_{1:n})$ if crystallized. Each $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})^T$. We offer three ways to derive and interpret it.

OLS as Plug-in Estimator

We first consider the case without intercept: $Y = \mathbf{X}'\beta + \epsilon$. Notice that here $\mathbf{X}$ is a random vector $(X_1, X_2, \ldots, X_k)^T$. All elements in the vector are random variables. We can also consider $X_1 = 1$ to include the intercept. We have:

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

Notice that β is a column vector with k elements. With matrix calculus, the result is

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

We use a plug-in estimator, replacing the theoretical expected value with the sample mean:

$$\hat{\beta} = \frac{\sum_{i=1}^n \mathbf{X}_i Y_i}{\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i'}$$

If we consider the intercept (separated from the covariates' matrix), we have

$$\beta, \beta_0 = \arg\min_{b, b_0 \in \mathbb{R}^k} E[(Y - \mathbf{X}'b - b_0)^2]$$

Again, β and β_0 are column vectors. With matrix calculus, we solve

$$-2E[Y - \mathbf{X}'b - b_0] = 0, \quad -2E[\mathbf{X}(Y - \mathbf{X}'b - b_0)] = 0$$

We have

$$b_0 = E[Y] - E[\mathbf{X}']b, \quad E[\mathbf{X}\mathbf{X}']b = E[\mathbf{X}Y] - E[\mathbf{X}']b_0$$

Therefore,

$$\beta = [\mathrm{Var}(\mathbf{X})]^{-1} \mathrm{Cov}(\mathbf{X}, Y)$$

Again, we use a plug-in estimator,

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

OLS as Minimization of the Residual

This is the most frequently used OLS interpretation, especially by econometricians and computer scientists. We use the sampled data $(\mathbf{X}_{1:n}, Y_{1:n}) \overset{i.i.d.}{\sim}$ same DGP, trying to fit a line that minimizes the total distances between predicted outcomes and true outcomes, i.e., the sum of squared residuals. We first consider the case without intercept. The predicted outcomes are $\hat{Y}_i = \mathbf{X}_i'\hat{\beta}$, while the true outcomes are Y_i. We are thus solving:

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

Taking the derivative, we have

$$-2\sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i'b) = 0$$

Therefore,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\mathbf{X}_i Y_i}{\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'}$$

With the intercept term, we solve

$$\beta, \beta_0 = \arg\min_{b, b_0 \in \mathbb{R}^k} \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'b - b_0)^2$$

We have

$$b_0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'b) = \bar{Y} - \bar{\mathbf{X}}'b, \quad \sum_{i=1}^{n}\mathbf{X}_i(Y_i - \mathbf{X}_i'b - b_0) = 0$$

Therefore,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(\mathbf{X}_i Y_i - n\bar{\mathbf{X}}\bar{Y})}{\sum_{i=1}^{n}(\mathbf{X}_i\mathbf{X}_i' - n\bar{\mathbf{X}}^2)} = \frac{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'}$$

OLS as Linear Projection

This approach reveals the geometric interpretation of OLS: given the data ($\mathbf{X} \in \mathbb{R}^{n\times k}, Y \in \mathbb{R}^n$), we are finding a vector $\hat{\beta} \in \mathbb{R}^k$ such that it minimizes the total Euclidean distance (L2 norm) from the predicted $\hat{Y} = \mathbf{X}\hat{\beta}$ to the true Y:

$$\hat{\beta} = \arg\min_{b} ||Y - \mathbf{X}b||^2$$

In other words, we are projecting Y onto the column space of $\mathbf{X}$. Let the projection matrix be P, such as $\hat{Y} = PY = \mathbf{X}\hat{\beta}$. The residual vector $\hat{\epsilon} = Y - \mathbf{X}\hat{\beta} = Y - PY$ must be orthogonal to the column space of $\mathbf{X}$. By matrix calculus, we can derive $\hat{\beta}$ by solving:

$$-2\mathbf{X}'(Y - \mathbf{X}b) = 0, \quad \text{so } \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

Since $PY = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$, we have $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We have the following properties:

1. P is an orthogonal projection matrix, so it is idempotent $P^2 = P$ and symmetric $P = P'$.
2. Let $M = I - P$. P makes the predicted outcomes, while M makes the residuals: $\hat{Y} = PY$, $\hat{\epsilon} = MY$, $Y = PY + MY$. M is also symmetric and idempotent.
3. $PX = X$, $MX = 0$.
4. Important assumption for projection: $\mathbf{X}'\mathbf{X}$ must be invertible (full-rank, i.e., column vectors linearly independent).

   Note: When adding unit fixed effects for a regression on a panel data with m units, you are essentially adding $m - 1$ binary covariates as unit indicators (adding all m would lead to multicollinearity). Adding both unit and time fixed effects is the standard practice and usually would not cause multicollinearity. However, adding the (factorized) intersection term of unit and time would cause perfect multicollinearity.

Frisch–Waugh–Lovell (FWL) Theorem

$$\mathbf{Y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\epsilon} \tag{1}$$

The coefficient $\hat{\beta}_2$ from the full OLS regression of $\mathbf{Y}$ on $(\mathbf{X}_1, \mathbf{X}_2)$ is equal to the coefficient from:

1. Regressing $\mathbf{Y}$ on $\mathbf{X}_1$, saving the residuals $\tilde{\epsilon}_1 = \mathbf{M}_1\mathbf{Y}$
2. Regressing $\mathbf{X}_2$ on $\mathbf{X}_1$, saving the residuals $\tilde{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2$
3. Regressing $\mathbf{M}_1\mathbf{Y}$ on $\mathbf{M}_1\mathbf{X}_2$:

$$\hat{\beta}_2 = (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}(\tilde{\mathbf{X}}_2'\tilde{\epsilon}_1) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}$$

   where $\mathbf{M}_1 = \mathbf{I} - \mathbf{P}_1$ is the annihilator matrix for $\mathbf{X}_1$.

Proof: Just multiply both sides of (1) by $\mathbf{M}_1$, we are essentially fitting the following regression,

$$\mathbf{M}_1\mathbf{Y} = \mathbf{M}_1\mathbf{X}_1\hat{\beta}_1 + \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_1\hat{\epsilon} = \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \hat{\epsilon}$$

We can do $\mathbf{M}_1\hat{\epsilon} = \hat{\epsilon}$ because when fitting the regression, the residual must be orthogonal to $\mathbf{X}_1$, and thus $\mathbf{P}_1\hat{\epsilon} = 0$, and $\mathbf{M}_1\hat{\epsilon} = (\mathbf{I} - \mathbf{P}_1)\hat{\epsilon} = \hat{\epsilon}$.