Simple Econometric Theory Review

Applied Econometrics Resources

Kevin Li

Model Specification

For independent variables $X_1, X_2, \ldots, X_p$, and outcome variable Y for units $i = 1, 2, \ldots, n$:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}_{\mathbb{E}[Y_i|X_i]} + \varepsilon_i$$

$\beta_0, \beta_1, \ldots, \beta_p$ are parameters that describe the deterministic part of the relationship between Y and $X_1, \ldots, X_p$.

- Read: the part of Y explained by $X_1, \ldots, X_p$.

ε_i error term is the non-deterministic relationship between Y and $X_1, \ldots, X_p$.

- Read: part of Y **not** explained by $X_1, \ldots, X_p$.
- $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$

Matrix Form

Condensed form:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{pmatrix}$$

Even more condensed matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Sum of Squared Errors

Naturally, we want to choose the values $b_0, \ldots, b_p$ for the unknown $\beta_0, \ldots, \beta_p$ that **minimise** the sum (squared) error of predicted $\hat{Y}_i$ in respect to the true population.

- Actual true Y values: Y_i, with unknown β
- Predicted $\hat{Y}$ values, with some choice of β value of b.

Thus, the sum (squared) error is the sum of the differences between actual Y_i and predicted $\hat{Y}_i$:

$$\mathrm{SSE} = \sum (Y_i - \hat{Y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

Why squared?

1. gets rid of direction, only keeps magnitude
2. Easier for calculus as absolute value function is non-differentiable at vertex.
3. Nice properties (see later in the slides).

Ordinary Least Squares

Re-arrange SSE:

$$\begin{aligned}\text{SSE} &= (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}\end{aligned}$$

We want to minimise the SSE, so take the derivative in respect to $\mathbf{b}$ and set equal to 0:

$$\frac{\partial \text{SSE}}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{Xb} = 0$$

Re-arrange the equation to get

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \implies \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

Estimator Properties

When we estimate β (or any parameter), we typically use a sample of the population.

- What if we used a different sample to calculate the parameter? We would get a slightly different $\hat{\beta}$ estimate since the sample data is slightly different.

Sampling distribution is the distribution of all estimated $\hat{\beta}$ from different samples, taking an infinite number of samples.

- Imagine you take one sample, and estimate $\hat{\beta}$. Then, take another sample and estimate $\hat{\beta}$. Then again and again. Plot all of the $\hat{\beta}$ in a distribution to get the sampling distribution.

Unbiasedness is if the expected value of the sampling distribution equals the true population value of β. In other words: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Standard Error is the standard deviation of the sampling distribution.

Unbiasedness of OLS (1)

Theorem: Part of the **Gauss-Markov Theorem** states that under 4 conditions, the OLS estimate of β is **unbiased**: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

1. The population model can be expressed as a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
2. i.i.d sampling from population.
3. No perfect multicollinearity. Basically, $\mathbf{X}$ must be full-rank.
4. **Strict Exogeneity**: Formally defined as $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$.

   This implies that $\text{Cov}(\varepsilon, X_j) = 0$ for any explanatory variable $X_j = X_1, \dots, X_p$.

Violations of strict exogeneity often caused by omitted confounders (see causal frameworks).

Unbiasedness of OLS (2)

Proof:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon
\end{aligned}$$

Now we want to prove $\mathbb{E}[\hat{\beta}] = \beta$. So we want to take the expected value of $\hat{\beta}$:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\
\implies \mathbb{E}[\hat{\beta}|\mathbf{X}] &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\varepsilon|\mathbf{X}]
\end{aligned}$$

Unbiasedness of OLS (3)

$$\mathbb{E}[\hat{\beta}|\mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\varepsilon|\mathbf{X}]$$

Recall Gauss-Markov condition (4), strict exogeneity: $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$. Thus:

$$\mathbb{E}[\hat{\beta}|X] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(0) = \beta$$

Finally, law of iterated expecations (LIE) gets us:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}]] = \beta$$

Thus, we have shown $\mathbb{E}[\hat{\beta}] = \beta$, proving OLS is an unbiased estimator of the true β population parameters under 4 gauss-markov conditions.

Variance of OLS (1)

Start with our solution:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\end{aligned}$$

β is a constant population value, so it is not the variance. Thus, the variance of the estimator comes from 2nd term:

$$\begin{aligned}\text{Var}[\hat{\beta}|\text{X}] &= \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\text{X}] \\ \implies \text{Var}[\hat{\beta}|\text{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}[\varepsilon|\mathbf{X}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon]^{-1} \\ \implies \text{Var}[\hat{\beta}|\text{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}[\varepsilon|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Variance of OLS (2)

Homoscedasticity assumption:

$$\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\mathbf{I} = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots \\ 0 & \sigma^2 & 0 & \dots \\ 0 & 0 & \sigma^2 & \vdots \\ \vdots & \vdots & \dots & \ddots \end{pmatrix}$$

- Read: no matter the value of $\mathbf{X}$, the error term ε has the same constant variance σ^2.

If homoscedasticity assumption is true, we can plug this into our OLS variance formula:

$$\begin{aligned} \text{Var}[\hat{\boldsymbol{\beta}}|\text{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}[\boldsymbol{\varepsilon}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Variance of OLS (3)

Alternatively, we can weaken this assumption to **heteroscedasticity**: where the error term variance depends on unit i's $\mathbf{X}$ values:

$$\text{Var}[\varepsilon|\mathbf{X}] = \sigma^2\mathbf{I} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_i^2 & \vdots \\ \vdots & \vdots & \dots & \ddots \end{pmatrix}$$

Our variance of OLS once plugging in is:

$$\text{Var}[\hat{\beta}|X] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_i^2 & \vdots \\ \vdots & \vdots & \dots & \ddots \end{pmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Hypothesis Testing

We do not know the values of σ^2 or σ_i^2. Thus, we use estimates of them involving our residuals $\hat{\varepsilon}_i$.

Using these estimates, we can find the estimated variance and standard error. From this, we can conduct hypothesis testing with t-tests.

$$t = \frac{\hat{\beta}_j - H_0}{\widehat{se}(\hat{\beta}_j)}, \quad \text{for} \quad \hat{\beta}_j \in \hat{\beta}_0, \dots, \hat{\beta}_p$$

- Where H_0 is the null (usually 0).

We can then calculate p-value: probability the null is true given our estimate $\hat{\beta}_j$.

Note: hypothesis testing is only approximate if ε is not normally distributed (will be achieved in large sample sizes due to CLM). Consider bootstrap inference for small samples.

Geometrics of OLS (1)

Our predicted values of $\hat{Y}_i$ are defined as following:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Let us define projection matrix $\mathbf{P}$ as:

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

- $\mathbf{P}$ is symmetrical $\mathbf{P}' - \mathbf{P}$, and idempotent $\mathbf{P}\mathbf{P} = \mathbf{P}$.

Thus, we can rewrite our predicted values as:

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$$

Thus, $\mathbf{P}$ is projecting $\mathbf{y} \to \hat{\mathbf{y}}$.

Geometrics of OLS (2)

Let us define residual maker matrix $\mathbf{M}$:

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

- $\mathbf{M}$ is also symmetrical and idempotent.
- $\mathbf{M}$ is orthogonal to $\mathbf{P}$ and $\mathbf{X}$, meaning $\mathbf{PX} = \mathbf{MX} = 0$. You can prove this on your own, it is pretty simple.

Our error between Y_i and $\hat{Y}_i$ is $\hat{\varepsilon}_i$:

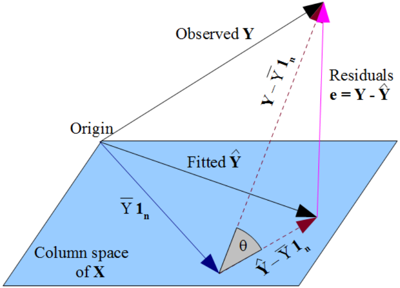
$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Py} \\ &= (\mathbf{I} - \mathbf{P})\mathbf{y} \\ &= \mathbf{My} \end{aligned}$$

Thus, $\mathbf{M}$ is projecting $\mathbf{y} \to \hat{\boldsymbol{\varepsilon}}$.

Geometrics of OLS (3)

We know that predicted $\hat{\mathbf{y}}$ is some linear combination of $\mathbf{X}$ (explanatory variables $X_1, \dots, X_p$), since $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.

Thus, $\mathbf{P}$ projects $\mathbf{y}$ into a vector $\hat{\mathbf{y}}$ that is in a space spanned by $\mathbf{X}$ (column space of $\mathbf{X}$).



$\mathbf{M}$ projects vector $\mathbf{y}$ into vector $\mathbf{e}$ (error), which is perpendicular to the column space of $\mathbf{X}$.

- Read: strict exogeneity: error term should not be correlated with $\mathbf{X}$.