Lecture 4

Why is the OLS a *good* estimator?

Ananya Iyengar

January 2026

St. Stephen's College

This corresponds to Chapter 4 of Studenmund (2017).

The previous lecture looked at estimating the coefficients of a linear regression mode using the OLS method. This lecture tries to understand how and under what conditions does the OLS estimator of β have properties which are favourable for inference and prediction.

Specifically, we will show that under some assumptions, the OLS estimator is the *best* among a larger class of estimators and we will characterise *best* in terms of our original goal – to estimate the population regression equation using the sample.

1 Assumptions of the Classical Linear Regression Model (CLRM)

1. *The regression model is linear in parameters, correctly specified and has an additive stochastic error term.*
   (a) The underlying model may be non-linear; log transformations often used.
   (b) For e.g., $Y = K^\alpha L^\beta e^\epsilon \iff lnY = \alpha lnK + \beta lnL + \epsilon$
2. *The population mean of the error term is 0 i.e.* $E[\epsilon] = 0$.
   (a) Note that the population ϵ is never observed!
   (b) If this assumption is violated, the non-zero mean is subsumed in the constant. The slope coefficient remains unchanged.

3. *The explanatory variables are uncorrelated with the stochastic error i.e.* $Cor(X, \epsilon) = 0$.

   (a) If this correlation is non-zero, the model would erroneously attribute variation in ϵ to variation in X.

   (b) When this happens, the estimated $\hat{\beta}$ is said to be biased. The difference between β and $\hat{\beta}$ would no longer be a product of *only* stochastic error. This problem is called endogeneity of the explanatory variables.

4. *Error term observations are uncorrelated with each other i.e.* $E[\epsilon_i \epsilon_j] = 0$ *for* $i \neq j$.

   (a) Recall – a random sample is one where all X_is are independently and identically distributed. If the stochastic error of some i is correlated with the error of some other observation j, then each observation in the sample is not "random" and the OLS estimator loses some precision (this will be formalised in upcoming lectures).

   (b) Non-zero correlation between the error terms is called autocorrelation or serial correlation, and is very pertinent when dealing with time series data.

5. *The conditional variance of the stochastic error term is a constant i.e.* $V(\epsilon/X) = \sigma^2$.

   (a) This assumption relates to the "identical" distribution part of i.i.d. The unexplained variation should have the same distribution for different X values.

   (b) Non-constant error variance is called heteroskedasticity; and this is often pertinent when dealing with cross-sectional data.

6. *No explanatory variable is a perfect linear function of any other explanatory variable(s) i.e. there is no perfect multicollinearity.*

   (a) Under perfect multicollinearity, OLS computation fails and β_1 is undefined.

   (b) As a very simple example, suppose in a simple regression $Y = \beta_0 + \beta_1 X + \epsilon$, we use $X = \alpha \times 1$. Then, the denominator of $\hat{\beta_1}$, which is $\sum(X - \bar{X})^2$ is 0! This follows in the multiple regression case and in the case of all affine transformations.

   (c) It is often the case that multicollinearity is imperfect i.e. two explanatory variables are highly correlated but the correlation is < 1. In that case, the coefficient can be estimated but the estimated coefficient runs into problems.

7. *The error term is normally distributed i.e.* $e \sim N(0, \sigma^2)$.

2 Sampling Distribution of $\hat{\beta}$

The *sampling distribution* is the probability distribution of $\hat{\beta}$ under different samples. Load this applet to play around with the concept of sampling distributions of the mean.

In practice we never observe the sampling distribution – we most often only sample once! However, the concept of the sampling distribution provides a way of thinking about desirable properties of the OLS estimators.

1. An estimator $\hat{\beta}$ is said to be *UNBIASED* if the mean of the sampling distribution of $\hat{\beta}$ is equal to the true population parameter β i.e.

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

   Of course, a given *estimate* of an unbiased estimator may be further away from the true population value than a given estimate of a biased estimator.

2. In the simple regression case, for the OLS estimator:
   $E[\hat{\beta}_1] = E[\frac{\sum x_i y_i}{\sum x_i^2}]$ where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. The Xs are non-stochastic. Therefore,
   $E[\hat{\beta}_1] = \frac{1}{\sum x_i^2} E[\sum x_i y_i] = \frac{1}{\sum x_i^2} E[\sum x_i (Y_i - \bar{Y})]$
   $= \frac{1}{\sum x_i^2} E[\sum x_i Y_i]$ since $\sum x_i \bar{Y} = 0$. $\implies E[\hat{\beta}_1] = \frac{1}{\sum x_i^2} E[\sum x_i (\beta_0 + \beta_1 X_i + e_i)]$
   $= \frac{1}{\sum x_i^2} E[\beta_0 \sum x_i + \beta_1 \sum x_i X_i + \sum e_i x_i]$
   $= \frac{1}{\sum x_i^2} E[0 + \beta_1 (\sum X_i^2 - n\bar{X}^2) + \sum e_i x_i]$
   $= \beta_1 \frac{1}{\sum x_i^2} (E[\sum x_i^2] + E[\sum e_i x_i]) = \beta_1$ from Assumption (3).

3. We also care about the variance of the sampling distribution – out of two unbiased estimators, we might prefer the one with a lower variance. $V(\hat{\beta})$ is the variance of the OLS estimator and the square root of the variance is called the *STANDARD ERROR* $SE(\hat{\beta})$.

4. Let us denote $\hat{\beta}_1$ as $\sum w_i Y_i$, where $w_i = \frac{x_i}{\sum x_i^2}$. We can do this because the Xs are non-stochastic. Then,
   $V(\hat{\beta}_1) = V(\sum w_i Y_i) = \sum w_i^2 V(Y_i)$ from Assumption (4) $\implies V(\hat{\beta}_1) = \sigma^2 \sum w_i^2$ from Assumption (5).

5. We will now show for the simple regression case that the OLS estimator has the lowest variance for a class of similar unbiased estimators.

6. Let $\sum c_i Y_i \equiv \hat{b}$ be any linear unbiased estimator of β_1. Since it is unbiased,
   $E[\sum c_i Y_i] = E[\sum c_i(\beta_0 + \beta_1 X_i + e_i)] = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$ For unbiasedness to hold, we need $\sum c_i = 0$ and $\sum c_i X_i = 1$.
   Now, $V(\hat{b}) = V(\sum c_i Y_i) = \sigma^2 \sum c_i^2$. Let $c_i = w_i + h_i$, where h_i is some additive term. Then,
   $\sigma^2 \sum c_i^2 = \sigma^2 \sum (w_i + h_i)^2$
   $= \sigma^2 \sum (w_i^2 + h_i^2 + 2w_i h_i)$
   $= \sigma^2 \sum w_i^2 + \sigma^2 \sum h_i^2 + 2\sigma^2 \sum w_i h_i.$
   Since $\sum h_i w_i = 0$[1],
   $V(\hat{b}) = \sigma^2 \sum w_i^2 + \sigma^2 \sum h_i^2 \geq V(\hat{\beta_2})\ \forall\ h_i$ Therefore, the OLS estimators have minimum variance within the class of linear unbiased estimators.

7. Note that the above two results hold *irrespective* of the sample size n. That is, we have currently enumerated the *finite sample properties* of the OLS estimators.

3 The Gauss Markov Theorem

Under Assumptions (1) to (6), the OLS estimator is ***BLUE*** *i.e. has the minimum variance ("best") among the class of all unbiased estimators of β.*

1. Normality of the error is not required for the Gauss Markov Theorem to hold.

2. Violations of the CLRM assumptions may affect either the variance or the bias of the estimator or both.

3. According to Hansen (2022), OLS is BUE i.e. has the minimum variance from the class of all unbiased estimators, not just linear ones.

4. Although not required for the Gauss Markov Theorem, the fact that the errors are assumed to have a normal distribution implies that the distribution of $\hat{\beta}$ is also normal, which has implications going forward.

For assessment: Questions 1-5 from the back of the text.

[1] Here, $\sum h_i w_i = \frac{1}{\sum x_i^2}(\sum h_i Xi - \bar{X} \sum h_i)$. Since $\hat{b}$ is unbiased and $\sum x_i = 0$ always, it follows that $\sum h_i = 0$. Also, $\sum h_i X_i = \sum (c_i - w_i) X_i = 1 - \frac{1}{x_i^2} \sum (X_i^2 - \bar{X} X_i) = 1 - \frac{1}{x_i^2}(\sum X_i^2 - \bar{X} \sum X_i) = 1 - \frac{1}{x_i^2}(\sum X_i^2 - n\bar{X}^2) = 1 - 1 = 0.$