1 Introduction

In statistics, an estimator is considered as a random variable that depends on the sample data. In our case of OLS, β is a function of X and y. **The expected value of an estimator represents the average value that the estimator would take if we were to repeat the estimation process many times with different samples from the same population. The variance of an estimator is a measure of the spread or dispersion of the estimator around its expected value.**

These concepts are crucial in statistical inference, as it allows us to evaluate the performance of an estimator and make statements about its properties (i.e. how does the estimator behave as the sample changes.

1.1 Bias of an Estimator

The bias of an estimator is defined as:

$$\text{Bias}(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i - \beta \tag{1}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i$ is the expected value of $\hat{\beta}$, we can rewrite the equation as:

$$\text{Bias}(\hat{\beta}) = E[\hat{\beta}] - \beta \tag{2}$$

where β is the true value of the parameter and $E[\hat{\beta}]$ is the expected value of the estimator.

- If the bias is zero, the estimator is said to be unbiased.
- If the bias is not zero, the estimator is said to be biased.

In the case of OLS, the estimator $\hat{\beta}$ is a function of X and y, and it can be proved (under some assumptions) that the expected value is $E[\hat{\beta}] = \beta$, which means that the OLS estimator is unbiased.

1.2 Variance of an Estimator

The variance of an estimator is a measure of the spread or dispersion of the estimator around its expected value. It is defined as:

$$\text{Var}(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])^2] \tag{3}$$

A small variance indicates that the estimator is concentrated around its expected value, which means that it is a good estimator. A large variance indicates that the estimator is spread out around its expected value, which means that it is not a good estimator.

In the case of OLS (and again under assumptions), the variance of the estimator $\hat{\beta}$ is given by:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{4}$$

where σ^2 is the variance of the error term and $(X'X)^{-1}$ is the inverse of the matrix $X'X$.

2 Gauss-Markov Theorem

The Gauss-Markov theorem states that under some assumptions, OLS is the Best Linear Unbiased Estimator (BLUE). This means that the estimator has the smallest variance among other unbiased and linear estimators.

- B = Best: with the lowest variance $(Var(\hat{\beta}_{\mathbf{OLS}}) < Var(\tilde{\beta}_i))$
- L = Linear: among family of linear estimators
- U = Unbiased: $E[\hat{\beta}]) = \beta$
- E = Estimator

2.1 Recall on Linear Regression

- Theoretical model of population: $y = \beta_0 + \beta_1 X + \epsilon$
- Sample estimation: $y = \hat{\beta}_0 + \hat{\beta}_1 X$
- We can estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ thanks to the OLS

2.2 OLS Assumptions

1. The relationship between x and y must be linear. i.e. the model is "linear in parameters".

$$y = \beta_0 + \beta_1^2 X + \epsilon \text{ (not linear in parameters, OLS 1 violated)} \tag{5}$$

$$y = \beta_0 + \beta_1 X + \epsilon \text{ (linear in parameters)} \tag{6}$$

$$y = \beta_0 + \beta_1 X^2 + \epsilon \text{ (linear in parameters, even though } X \text{ is transformed)} \tag{7}$$

2. **Random sampling** : data are randomly sampled

   The sample data is selected randomly from the population, and every individual in the population has an equal chance of being included in the sample. This assumption helps to ensure that the sample is representative of the population, and that the results of the analysis can be generalized to the population.

   Random sampling helps to minimize the risk of **selection bias**, where certain individuals or groups are more likely to be included in the sample than others.

   Example : Suppose you want to study the relationship between the number of hours spent studying and the GPA of students at KBTU.

   You decide to collect data from two sub-samples of students:

   - Sub-sample 1: Students from the IT department
   - Sub-sample 2: Students from the Literature department

   You collect data from 50 students in each sub-sample, resulting in a total sample size of 100 students. You then run a linear regression analysis to examine the relationship between the number of hours spent studying and GPA.

   Here's a potential problem: if you don't randomize the sample selection, your sub-samples may not be representative of the larger population of KBTU students.

   For example:

   - IT students may have different study habits and academic backgrounds than Literature students. IT students may be more likely to spend long hours programming and working on projects, while Literature students may be more likely to spend time reading and analyzing texts.
   - IT students may have higher GPAs on average than literature students, due to differences in academic preparation or motivation.
   - The relationship between study hours and GPA may be different for IT students versus literature students. For example, IT students may see a stronger relationship between study hours and GPA, since their coursework requires more hands-on programming and problem-solving.

   In this case, our estimator $\hat{\beta}$ would not be efficient, i.e. would have a large variance, i.e. we could find another linear unbiased estimator more efficient (according to Gauss Markov theorem).

3. **No perfect collinearity** : no independent variable is a perfect linear combination of the other independent variables.

   The assumption of no perfect collinearity means that:

   - The matrix X is full rank, i.e., $Rank(X) = p$, where p is the number of columns.
   - The columns of X are linearly independent, i.e., no column can be expressed as a linear combination of the other columns.

   Example : Still with the example of GPU score of KBTU students, lets say we estimate the GPU score by different explanatory variables such as :

   $$\widehat{\text{GPU}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Nb of lecture attended} + \hat{\beta}_2 \text{Hours in Room 388} + \hat{\beta}_4 \text{IQ} + \hat{\epsilon} \tag{8}$$

   In this regression, Nb of lecture attended and Hours in Room 388 are strongly correlated, as the lecture take place in the room 388. This means that students who attend more classes also tend to spend more hours in room 388, and vice versa. In this case, our feature matrix might not be of full rank, i.e. can't be invertible.

4. **Independence of errors :** The errors (or residuals) in the model are independent of each other, i.e. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$

   Example : Back to the same example, we now assume the regression :

   $$\widehat{\text{GPU}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Nb of lecture attended} + \hat{\beta}_3 \text{Hours of Independent Study} + \hat{\beta}_4 \text{IQ} + \hat{\epsilon} \tag{9}$$

   Here we added the *Hours of Independent Study* variable. Suppose that some of the students in the sample are friends and tend to study together. They often meet up in the library or at each other's apartments to study and discuss the material. As a result, their study habits and approaches to learning are similar, and their exam scores tend to be correlated. This **Omitted variable**[1] can create correlation in the errors, because the independence of errors assumption is no longer met. The errors are correlated, because the students who study together tend to have similar results, i.e. residuals.

   Correlation in the residuals can lead to unstable coefficient, i.e. inefficient (with high variance).

[1] Relevant variable not included in the model

5. **Homoskedasticity :** The variance of residual is independent on explanatory variables level. i.e. $\text{Var}(\varepsilon \mid X) = \sigma^2 I_n$

   Homoskedasticity refers to the condition where the variance of the residuals (or errors) in a regression model is constant across all levels of the predictor variables. In other words, the spread of the residuals is the same at all points of the predictor variables.

   Example :

   $$\widehat{\text{GPU}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Hours of Independent Study} + \hat{\epsilon} \tag{10}$$

   Again in the case of predicting *GPU*, we could think that there exists a strict linear relationship between hours of studies and *GPU*. In fact, according to the method of study, a student may stagnate or not study well (bad learning methodology), so that his GPU will not evolve with the *hours of independent* study. However, it is almost certain that the *GPU* will be low if he never studies. In this case, the variance of the residuals will be heteroskedastic (and violates the homoskedasticity assumption)

6. **Zero conditional mean :**

   $$E[\varepsilon \mid X] = 0 \tag{11}$$

   the zero conditional mean assumption ensures that the regression model is unbiased, meaning that it doesn't systematically overestimate or underestimate the response variable for any particular set of predictor variables.

   If this assumption is met, it implies that the regression line accurately represents the relationship between the predictor variables and the response variable, and that the errors are randomly distributed around this line.

   If this assumption and the homoskedasticity are met, then ϵ is normally distributed :

   $$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{12}$$