This section handout is devoted to the linear regression model. It introduces the general notation used in linear regression models, presents the different assumptions we impose on these models, and discusses key properties of the estimators derived from these models.

1 Introduction and Notation

1.1 Context

An economic model consists of mathematical equations that describe various relationships. In lecture, we have started studying a framework to describe a student's test score in relation to the size of their classroom:

$$Y = f(X)$$

where $Y =$ test scores and $X =$ class size.

After we specify an economic model, we need to turn it into what we call an econometric model. The form of the function $f(\cdot)$ must be specified: $f(\cdot)$ could simply be X as in the case with the examples we have seen yet, or it could be some transformation of such as X^2 or $\log(X)$. We also have to think about how to deal with variables that cannot reasonably be observed, such as "parental resources". We can specify the following econometric model for example:

$$TestScore = \beta_0 + \beta_1 ClassSize + \epsilon$$

where $TestScore$ =some measure of students' test scores, $ClassSize$ = the size of their classroom. Please note that the ϵ term contains all of the unobserved variables that also explain test scores (things like motivation, innate ability, other resources available to the students, characteristics of school, peer group, etc). We will learn how to deal with "unobservables" later in the course.

The constants β_0 and β_1 are the parameters of the econometric model, and they describe the directions and strengths of the relationship between test score and class size used to determine test scores in the model.

1.2 Regression in the population

Let's consider a model where Test Score (Y) is only a function of Class Size (X),

$$Y = f(X, \epsilon) = \beta_0 + \beta_1 X + \epsilon$$

We assume that this is the true data generating process. Note that we have assumed a linear function.

$\epsilon = Y - \beta_0 - \beta_1 X$: The variable ϵ is called the **error term**, or disturbance in the relationship, and represents factors other than X that affect Y. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model. We make an important assumptions about ϵ that we will explain more in the next section of these notes: $E(\epsilon|x) = E(\epsilon) = 0$.

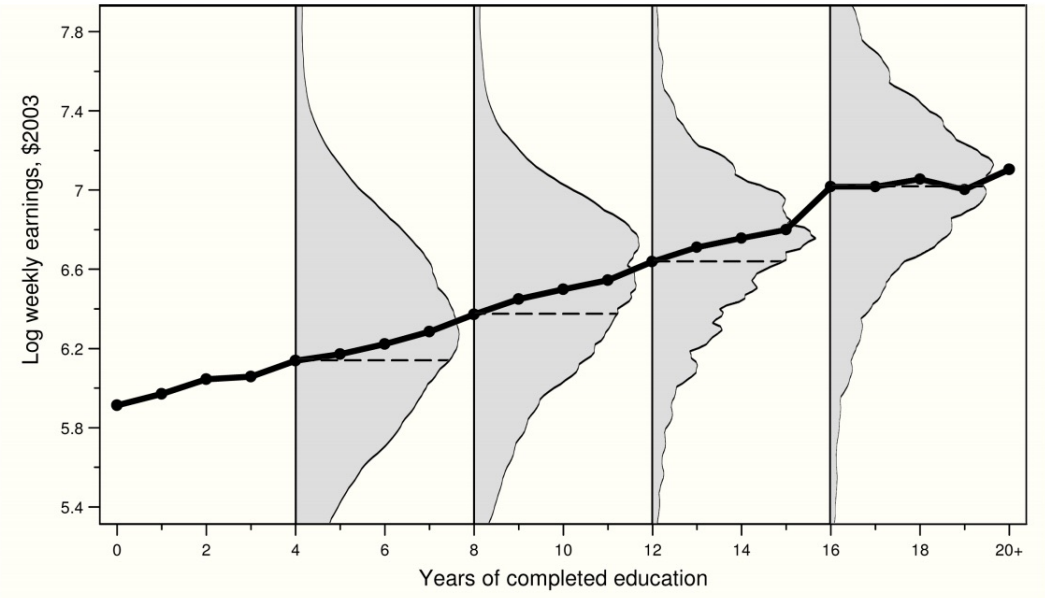
$E(Y|X) = \beta_0 + \beta_1 X$: This is the **population regression function** (PRF). $E(Y|X)$ tells us how the population average of one variable changes as we move the conditioning variable over the different values this variable might assume. For every value of the conditioning variable, we get a potentially different average of the dependent variable Y. The collection of all such averages is called the population regression function. In this class, we will assume the PRF is linear in the parameters β_0 and β_1. This linearity implies that a one-unit increase in X (or its transformation) changes the *expected* value of Y by the amount of β_1. For any given value of X, the distribution of Y is centered around $E(Y|X)$.

Note that this definition relies on the assumption (which we will investigate later) that $E(\epsilon|X) = E(\epsilon)$, which is essentially saying that the value of X doesn't convey any information on average about the value of the error term.[1] We can also write:

$$Y_i = E(Y|X) + \epsilon_i$$

[1] This assumption will allow us to interpret the β coefficient (in the population) as the causal effect of an additional unit of X on the expected value of Y. We can still fit a line to our data without this assumption, but we won't be able to interpret the estimate as causal. Think of investigating the impact of education on income. If there is something unobservable such as ability that varies

This says that any variable Y_i can be decomposed into a piece that is explained by X, $E(Y|X)$, and some piece that is left over, ϵ, which we don't observe.



The figure above plots the population regression function of log weekly wages given schooling for men from the 1980 US census. The distribution of earnings is also plotted for a few key values: 4, 8, 12, and 16 years of schooling.

The PRF tells us how the average value of Y changes with X: it does not say that Y equals $\beta_0 + \beta_1 X$ for all units in the population. For example, suppose X=4, then on average this implies log weekly earnings of 5.9 dollars. This does not mean that everyone with 4 years of schooling makes 5.9 dollars.

In this picture the PRF isn't actually linear, but for the purposes of this class we often assume that it is.

1.3 Regression in a sample

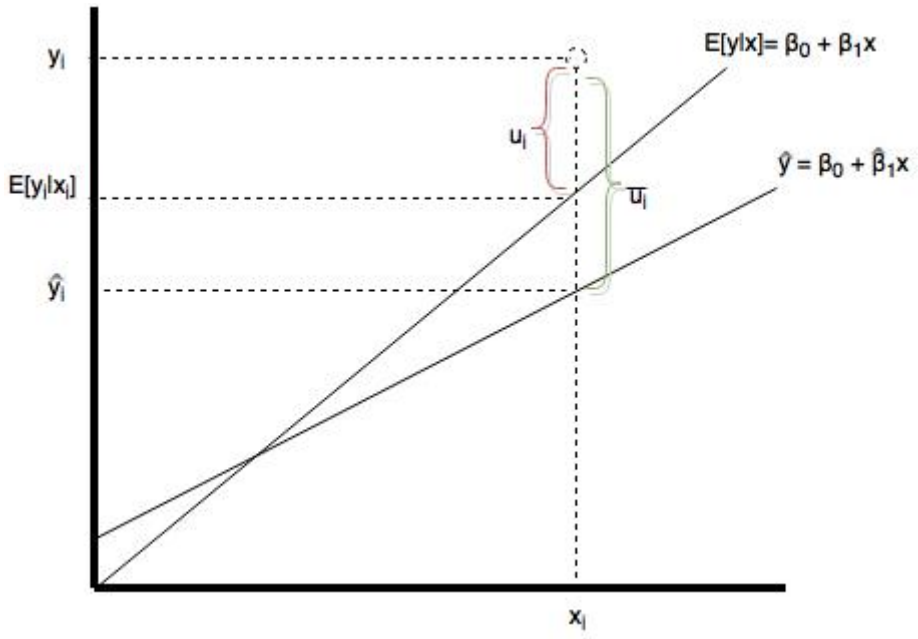
As we have mentioned, we often work with samples rather than the entire population of data. In this case, we do our best to approximate this population regression function.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$: This is the **fitted regression line**. It can be thought of as our best guess for Y given a certain value of X. This equation is also called the **sample regression function** (SRF) because it is the estimated version of the PRF.

$e = Y - \hat{Y}$: The variable e is called the **residual**, it can be thought of as the deviations between the real Y_i value and the predicted $\hat{Y}_i$ value.

$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e = \hat{Y} + e$: This is now our estimated model. The hat symbol above our beta's indicate that these are calculated estimates of the true beta value they represent. Again we see how we can decompose *Yi* into two parts: a fitted value (best guess) and a residual.

with the level of education (e.g. higher educated people also have more ability) such that $E[\epsilon|X] \neq 0$, then we won't be able to say that the coefficient associated with education reveals the true effect of education on income because we are confounding the effect of ability and education. We will discuss this in more details later.



2 Assumptions of Linear Regression Model

2.1 Summary Table

The linear regression model includes of a set of five key assumptions that allow us to study the properties of the parameters estimated by the model. These five assumptions we consider are as follows:

No.	Formal Assumption	Intuition
1	$\mathbb{E}[\epsilon \mid X] = 0$	For any given value of X, the mean of ϵ is zero. This implies that ϵ and X are uncorrelated.
2	$(X_i, Y_i)_{i=1}^n$ are independent and identically distributed (i.i.d)	The data drawn from the population are mutually independent and follow a common distribution (i.e. they are sampled randomly).
3	The matrix $\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^T]$ is invertible (i.e. has full rank) and is finite	The variable X is not constant: $Var(X) \neq 0$
4	$\mathbb{E}(\mathbf{X}_i \epsilon_i (\mathbf{X}_i \epsilon_i)^T) = \mathbb{E}(\mathbf{X}_i \epsilon_i^2 \mathbf{X}_i^T)$ is non-singular	Condition imposed to be able to use the Central Limit Theorem
5	The random variables (X_i, Y_i) have finite fourth moments	Large outliers are unlikely in the data

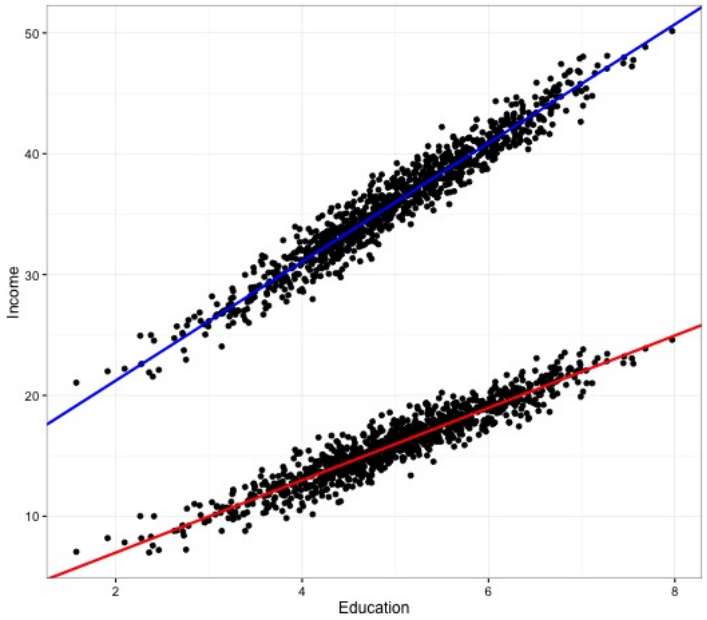
2.2 Intuition

In this sub-section, we elaborate on intuition points to help understand the meaning and the importance of each of the five assumptions stated above.

1. This first assumption is called the Zero Conditional Mean assumption. In words it says that no observations on X convey any information about the expected value of the error term.

   The graph below tries to give some intuition. Essentially, this assumption will allow us to interpret the β coefficient as the causal effect of an additional unit of X on the expected value of Y. We can still fit a line to our data without this assumption, but we won't be able to interpret the estimate as causal. Think of investigating the impact of education on income in the US. If we could account for everything that affect income (ability, education, parent's education, private versus public school attendance) then we could control for all the variables that affect income and isolate the effect of education on it's own. This scenario corresponds to the red line in the graph.

   Unfortunately we live in a world where we don't observe everything. Most notably we don't usually observe measures of ability/IQ. If this unobservable characteristic (ability) varies with the level of education (higher educated people also have more ability), and we can't tease the two effects apart, then running a regression of income on education will give us coefficients that *cannot* be interpreted as the causal effect of education. This scenario corresponds to the blue line in the graph. In other words, we won't be able to say that the coefficient associated with education reveals the true effect of education on income because we are confounding the effect of ability and education.



2. This second assumption states that the pairs of n points that we draw from our sample must be mutually independent and must follow a common distribution.

   A set of independent and identically distributed points can also be thought of as a random sample. This random sampling is important to maintain since we want to be able to say something about the population at large. For example, if we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. When is this condition violated? Suppose for example that we are interested in studying factors that influence the accumulation of family wealth. This is a sensitive topic and while we may chose a set of families to interview at random, some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families.

   The independent and identical distribution assumptions (or the random sampling assumption) are important since they allow us to use common laws and theorems in order to infer the properties of our parameters as our sample size changes.

3. This third assumption ensures that we are not studying a variable whose value does not vary in the population. A lack of variation in a variable of interest X would mean we cannot study its relationship with any other variable Y.

   If X varies in the population, random samples of X will typically contain variation unless the population variation is minimal or the sample size is small. This non-zero variance is thus important to maintain in the linear regression model since it allows us to have well-behaved estimators (more on this in the next section).

4. This fourth assumption is imposed to allow us to be able to use the Central Limit Theorem in the derivation of the asymptotic behavior of the variance of the estimated $\hat{\beta}$ in our linear regression model.

5. This fifth assumption relates to the fourth moment of a random variable X, which measures the fatness of the tail of the distribution of X (i.e. how large the difference between the density function of the variable and 0 is, as X approaches ∞ and $-\infty$. This fourth moment is always non-negative since it is the expectation of a fourth power. Assumption 5 thus requires the two random variables of our linear regression model X and Y to have thin tails and to have their density function converge to 0 as they approach ∞ and $-\infty$. This means that we assume that there will not be many outliers in the sample that we draw from the population.

   The convergence of the density function is the condition that ensures that the fourth moment is finite (since the expectation of the variable raised to the fourth power will be some large positive number.)

3 Properties of $\hat{\beta}$

We use the vector $\hat{\beta}$ to refer to the estimated intercept of the linear regression model β_0 and the estimated slope of the same linear regression model β_1 jointly.

3.1 Derivation of $\hat{\beta}$

We saw in lecture Handout 1 and in Section notes 2 that the population parameters:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}(X^2) \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix} \tag{1}$$

We now define the following vectors:

$$\mathbf{X} = \begin{bmatrix} 1 \\ X \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2}$$

We can use the vectors in equation (2) to rewrite the different elements of equation (1) as follows:

$$\begin{bmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}(X^2) \end{bmatrix} = \mathbb{E}(\mathbf{X}\mathbf{X}') \text{ and } \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix} = \mathbb{E}(\mathbf{X}Y) \tag{3}$$

Using the expressions in equation (3), equation (1) can now be rewritten as follows:

$$\beta = \left(\mathbb{E}(\mathbf{X}\mathbf{X}')\right)^{-1} \mathbb{E}(\mathbf{X}Y) \tag{4}$$

By the analogy principle, we have replace the expectations $\mathbb{E}$ by sample averages (since we do not have distributions to calculate expectations). We thus replace $\mathbb{E}(\mathbf{X}\mathbf{X}')$ by

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

and $\mathbb{E}(\mathbf{X}Y)$ by

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

where

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ X_i \end{bmatrix}$$

With these replacements, our guess for β, which we call $\hat{\beta}$) can be written as follows:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i \tag{5}$$

3.2 Bias and Consistency

We define terms that are used to qualify the behavior of estimator $\hat{\beta}$.

Unbiasedness: A random variable X is unbiased for a constant m if $\mathbb{E}(X) = m$. If $\mathbb{E}(X) \neq m$, then we say that W is biased for m. In our study of the linear regression model, we are interested in checking whether our estimator $\hat{\beta}$ is unbiased for the true parameters β, which would be the same as checking whether our estimate $\hat{\beta}$ is correct on average. In general, unbiased estimators are preferred to biased estimators.

Consistency: A sequence of random variable $\{X_n\}_{n=1}^{\infty}$ converges in probability to a constant m if for all $\tau > 0$, we have:

$$\lim_{n\to\infty} P(||X_n - m|| > \tau) = 0$$

In plain English, this definition states that a sequence of random variables converges to a constant m if the realization of the random variables approaches the constant m as n gets larger.

More elaborately, in order to check consistency, we need two things: a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ and a number m towards which the sequence might converge. Once this is clear, we choose a given number τ, which can be as small or large as we want to be. For each realization of the sequence of the random variable, we calculate the difference between the realized value and m and check whether this difference is greater than our chosen threshold τ. If the likelihood that this difference is greater than τ becomes closer and closer to 0 as the sample size increases, and if this holds for even infinitesimal values of τ, then we say that the sequence of random variables converges to m. In that case, we say X_n **is consistent for** m.

3.3 Law of Large Numbers and Central Limit Theorem

In this class, we talk a lot about samples versus populations. In an ideal world, we would have data on the universe of individuals in order to estimate a relationship between X and Y. In other words, we would have data on the population at large. However, it's almost always the case that we do not have all this data at our disposal. Rather, we have a sample of individuals (e.g. Berkeley students rather than the universe of students at American universities). We can calculate statistical properties of these smaller samples, which we call sample mean and sample variance. The **sample mean** is given below:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The **law of large numbers** says that if we draw a sample consisting of n realizations of our random variable, take the average of this sample, then this sample mean will approach the population mean as n approaches infinity.

As an example, let X be the roll of a die, which can take on values 1, 2, 3, 4, 5 or 6. The population average (or the mathematical expectation) is the average if we were to throw the die infinitely many

times:

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

Next, let's calculate the sample mean, which is the value we get after only 100 throws of the die:

n	x_j	$\bar{X}_n$
2	6, 6	$\frac{12}{2} = 6$
3	1, 2, 2	$\frac{5}{3} = 1.67$
4	1, 1, 6, 3	$\frac{11}{4} = 2.75$
⋮	⋮	⋮

As n$\to \infty$ any irregularities that occur due to the small sample size are muted, and the sample mean will converge to the population mean.

The **Central Limit Theorem (CLT)** states that the distribution of the difference between the sample mean and the population mean approximates the normal distribution with mean 0 and standard deviation σ, when multiplied by the factor $\sqrt{n}$. This convergence holds regardless of the shape of the distribution of the initial random variable. [2]

Exercise 1: Suppose that $\{Y_i\}_{i=1}^n$ is an i.i.d. sample from a distribution with mean μ and variance σ^2

1. Let $\bar{Y}$ denote the sample mean. Characterize the behaviour of this quantity when the sample size is large and state the theorem you use.

2. Consider the object $\sqrt{n}\left(\bar{Y} - \mu\right)/\sigma$. Characterize the limiting distribution of this object and state the theorem you use to characterize it.

3.4 Properties of $\hat{\beta}$

Recall from equation (5) above that we can write our $\hat{\beta}$ as follows:

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

Substituting Y_i with $\mathbf{X}_i'\beta + \epsilon_i$ in the last term yields:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\mathbf{X}_i'\beta + \epsilon_i\right) \tag{6}$$

$$\tag{7}$$

We can break down the term $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\mathbf{X}_i'\beta + \epsilon_i\right)$ into two terms: $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\beta$ and $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i$. We can rewrite equation (6) as follows:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)\beta + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i \tag{8}$$

[2] This holds for independent and identically distributed random variables, but can still be established under certain conditions when these two prerequisites are not met.

The term $\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)$ is equal to the identity matrix I since it is the product of a matrix and its inverse. So we replace $\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)\beta$ with $I\beta$ or simply β. Equation (9) thus become:

$$\hat{\beta} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i \tag{9}$$

Property 1: $\hat{\beta}$ ***is unbiased for*** β

We take the expected value of all terms in equation 9 above yields:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}\left[\mathbb{E}[\hat{\beta}|\mathbf{X_1},\ldots,\mathbf{X_n}]\right] && (10)\\
&= \mathbb{E}\left[\mathbb{E}\left[\beta + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i|\mathbf{X_1},\ldots,\mathbf{X_n}\right]\right] && (11)\\
&= \mathbb{E}\left[\beta + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbb{E}[\epsilon_i|\mathbf{X_1},\ldots,\mathbf{X_n}]\right] && (12)\\
&= \mathbb{E}[\beta + 0] && (13)\\
&= \beta && (14)
\end{aligned}$$

$\mathbb{E}(\hat{\beta}) = \beta$ since this β is a vector of true parameters whose value does not change. Assumption 3 stated at the beginning of this section tells us that we can assume that $(\mathbb{E}(\mathbf{X}_i\mathbf{X}_i'))^{-1}$ is a finite number and so the sample analogue of this term also does not go to infinity. Assumption 2 implies that $\mathbb{E}[\epsilon_i|\mathbf{X_1},\ldots,\mathbf{X_n}] = \mathbb{E}[\epsilon_i|\mathbf{X}_i]$ since ϵ_i is distributed independently of $\mathbf{X}_j$ for all $j \neq i$. Assumption 1 (the zero conditional mean assumption) also implies that $\mathbb{E}[\epsilon_i|\mathbf{X}_i] = 0$. So the last term of equation (12) is the product of a finite number and 0, which mean that this last term is equal to 0.

In the end, we are left with $\mathbb{E}(\hat{\beta}) = \beta$, which means that $\hat{\beta}$ is unbiased for β.

Property 2: $\hat{\beta}$ ***converges towards*** β

We reconsider all terms of equation (9) and check whether they converge to anything.

The true β does not converge to anything since its elements are true parameters not subject to any probability. Using assumption2 and by the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'$ converges in probability to $\mathbb{E}(\mathbf{X}_i\mathbf{X}_i')$, and we know from assumption 3 that this is finite. Using Assumption 2, by the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i$ converges in probability to $\mathbb{E}(\mathbf{X}_i\epsilon_i)$ which is 0. So, for the last term of equation (9) $\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i$ converges to 0. This implies that $\hat{\beta}$ converges in probability to β. This convergence in probability simply means that in large samples, the estimator should be close to the true β with a high probability.

Property 3: $\sqrt{n}(\hat{\beta}-\beta)$ ***converges to a normal distribution (or achieves asymptotic normality)***

Under the five assumptions stated in the previous section, we can derive the following result using multiple iterations of the Central Limit Theorem:

$\sqrt{n}\left(\hat{\beta}-\beta\right)$ is distributed *exactly* as $\mathcal{N}\left(0,\hat{V}\right)$ for some variance V. [3]

This property is particularly useful if we want to make inference and answer questions of the sort: Is the true β higher than or lower than some number? Is the true β different from a certain number? These

[3] We refer the interested readers to the detailed step-by-step explanations presented in the lectures notes to get the exact variance formula.

questions are often what policymakers are most interested in. We can leverage the properties above to answer them without knowing the true β. We will discuss this point in more details in the next section handout on hypothesis testing.

4 Conclusion

So to close off this section note, while the assumptions and the different derivations might seem too vague or out of touch, they help put some structure on the linear model to ensure that the estimators we derive from them are well behaved (as the sample size increases). This good behavior of the estimators allows us to use these estimators to answer practical questions about the populations of interest without needing to observe the entire information about the population of interest. This model can thus be a great tool to avoid the expensive or impossible task of collecting information about an entire population just to be able to address a policy issue.

However, the good behavior of the model estimators is only as good as the soundness of the assumptions needed to derive these estimators. In future weeks, we will revisit some of these key assumptions and discuss what strategies to adopt when they do not hold.