Economics 361
Sampling

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

1 Data Generating Process

Thus far, we have not discussed the role of "data" in statistical inference. This is because we have assumed that we know the **data generating process (DGP)** fully. For random experiments that can be properly characterized by random variables, the DGP underlying those experiments can be properly characterized by the **joint distribution** of the relevant random variables. If the joint distribution of the relevant random variables are known, then statistical inference may be conducted without any data; we only need to apply probability theory to our known joint distribution. The best predictor and best linear predictor results emphasize this point.

Consider a game of chance like a coin flip. If we know that the coin and flip are both fair – head and tail are equally likely – then statistical inference on the coin flip does not require data (the archive of the outcomes of past coin flips). We can derive the proper distribution underlying the coin flip and use probability theory to conduct most any statistical inference on the coin flip. Let X be a binary random variable equal to 1 if the coin flip yields head (and zero if tail). Then

$$X = \begin{cases} 1 & \text{if head} \\ 0 & \text{if not head} \end{cases} \qquad f(x) = \begin{cases} \frac{1}{2} & \text{for } x = 1 \\ \frac{1}{2} & \text{for } x = 0 \\ 0 & \text{for } x \neq 0 \text{ and } x \neq 1 \end{cases}$$

The distribution of X, $f(x)$, characterizes the data generating process underlying a single flip of a fair coin. Most any statistical inference concerning a single flip of a fair coin can be conducted by applying probability theory to $f(x)$. For example, predicting the outcome of the next coin flip could be done by calculating best predictor of Y given X for $f(x)$ and the desired loss function.

Suppose we are interested in the random experiment involving two coin flips. Let X_1 and X_2 be defined similar to X above, with X_1 denoting the first coin flip and X_2 the second coin flip. Here, the data generating process would be characterized by the joint distribution of X_1 and X_2. If

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

we use the same fair coin and ensure that each flip is independent of each other, then the joint distribution simplifies to

$$f(x_1, x_2) \quad = \quad f(x_1)\ f(x_2) \quad = \quad \begin{cases} \frac{1}{4} & \text{for } x_1 = 1, x_2 = 1 \\ \frac{1}{4} & \text{for } x_1 = 1, x_2 = 0 \\ \frac{1}{4} & \text{for } x_1 = 0, x_2 = 1 \\ \frac{1}{4} & \text{for } x_1 = 0, x_2 = 0 \\ 0 & \text{for all other } x_1, x_2 \end{cases}$$

But more generally,

$$f(x_1, x_2) \quad = \quad f(x_1)\ f(x_2) \quad = \quad \begin{cases} \theta_{11} & \text{for } x_1 = 1, x_2 = 1 \\ \theta_{10} & \text{for } x_1 = 1, x_2 = 0 \\ \theta_{01} & \text{for } x_1 = 0, x_2 = 1 \\ \theta_{00} & \text{for } x_1 = 0, x_2 = 0 \\ 0 & \text{for all other } x_1, x_2 \end{cases}$$

where $\{\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00}\}$ are non-negative real values such that $\theta_{11} + \theta_{10} + \theta_{01} + \theta_{00} = 1$. If we know $\{\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00}\}$ then we can use the joint distribution $f(x_1, x_2)$ to conduct our statistical inference.

But we do not always fully know the DGP. Consider a twist on the above example: we know that each flip is independent of each other using the same coin; but we do not know whether the coin is fair (head and tail may not be equally likely). So X_1 and X_2 are **independently** and **identically** distributed. But we do not fully know this common distribution.

$$f(x_1) \quad = \quad f(x_2) \quad = \quad \begin{cases} \theta & \text{for } x = 1 \\ 1 - \theta & \text{for } x = 0 \\ 0 & \text{for } x \neq 0 \text{ and } x \neq 1 \end{cases}$$

$$f(x_1, x_2) \quad = \quad f(x_1) f(x_2) \quad = \quad \begin{cases} \theta^2 & \text{for } x_1 = 1, x_2 = 1 \\ \theta(1-\theta) & \text{for } x_1 = 1, x_2 = 0 \\ (1-\theta)\theta & \text{for } x_1 = 0, x_2 = 1 \\ (1-\theta)^2 & \text{for } x_1 = 0, x_2 = 0 \\ 0 & \text{for all other } x_1, x_2 \end{cases}$$

So, in this situation, we know the relevant DGP up to the unknown parameter θ (the probability of head on a single coin flip). But we cannot conduct statistical inference without θ. To solve this dilemma, we introduce data. We will use data to help us "estimate" the unknown parameters of the relevant DGP.

2 Population and Sample

Consider a random experiment characterized by the random variables (X, Y) that has been repeated N times. The outcomes for each of the N repetitions has been observed and stored as follows:

$$\{ (X_1 = x_1, Y_1 = y_1) \cdots (X_N = x_N, Y_N = y_N) \} = \{ (X_i = x_i, Y_i = y_i) \}_{i=1}^{N} = \{ (x_i, y_i) \}_{i=1}^{N}$$

The above set of observations is known as a **sample** of the random experiment (X, Y). More specifically, $\{ (x_i, y_i) \}_{i=1}^{N}$ is a **size** N **sample** of the random experiment (X, Y).

Note that $\{ (X_1, Y_1) \cdots (X_n, Y_N) \}$ is a series of $2N$ random variables. Instead of thinking of the sample as the result of N repetitions from a single random experiment involving the **two** random variables (X, Y), the sample can be thought of as the result of a single repetition of a single random experiment involving the $\mathbf{2}N$ random variables $\{ (X_1, Y_1) \cdots (X_n, Y_N) \}$.

Therefore, the relevant **DGP** underlying this sample is characterized by the $2N$ joint distribution

$$f(x_1, \ldots, x_N, y_1, \ldots, y_N)$$

Each repetition of the random experiment (X, Y) is called an **observation** of (X, Y)

Contrasting the sample is the concept of a **population.** The population is the random variables that characterize the underlying random experiment (X, Y). Thus, $\{ (X_1 = x_1, Y_1 = y_1) \cdots (X_N = x_N, Y_N = y_N) \}$ is the size N sample of the population (X, Y).

Perhaps confusingly, the term "sample" is used in multiple inconsistent manners.

- "Sample" appears in "sample space" – the set of all possible outcomes
- "Sample" refers to the N observations from the population (X, Y)
- "Sample" appears in "sample moments" – an analogous but different concept from moments

In order to alleviate confusion associated with the last usage, moments associated with the random variables making up the population (i.e. X, Y) are sometimes referred to as **population moments**.

The sample moment of $g(X, Y)$ is simply the average of value of $g(X, Y)$ across the N observations:

$$\text{Sample Moment of } g(X,Y) = \frac{1}{N} \sum_{i=1}^{N} g(x_i, y_i)$$

Three key sample moments are the **sample mean**, **sample variance**, and **sample covariance**

$$\text{Sample Mean of } X = \bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\text{Sample Var. of } X = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X}_N) \qquad \text{Sample Cov. of } X, Y = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X}_N)(y_i - \bar{Y}_N)$$

The estimation principle known as "Method of Moments" involves forming a link between sample moments and population moments. We explore this link later.

2.1 Random Sample

Suppose that each observation of the random experiment (X, Y) are **statistically independent** of each other. Then

$$f(x_1, \ldots, x_N, y_1, \ldots, y_N) \;=\; f(x_1, y_1)\; f(x_2, y_2)\; \cdots\; f(x_N, y_N) \;=\; \prod_{i=1}^{N} f(x_i, y_i)$$

The joint distribution for the sample can be "blocked" into the product of the joint distribution of the N observations $(X_i = x_i, Y_i = y_i)$ as each of the observations is distributed *independently* of each other.

Additionally, if each observation of the random experiment (X, Y) are **statistically identical** to each other

$$f(x_1, \ldots, x_N, y_1, \ldots, y_N) \;=\; \prod_{i=1}^{N} f(x, y) \;=\; (\; f(x, y)\;)^N$$

as $f(x_i, y_i) = f(x_j, y_j) = f(x, y)$ for $i, j = 1\ \ldots\ N$: the joint distribution for each of the N observations $(X_i = x_i, Y_i = y_i)$ is distributed *identically* to each other.

A size N sample where each of the N observations is **independently and identically distributed (*i.i.d.*)** is called a **random sample**.

2.2 Independently and/or Identically

A size N sample may not be random. Observations in the sample may be independently distributed but not identically

$$f(x_1, \ldots, x_N, y_1, \ldots, y_N) \;=\; f(x_1, y_1)\; f(x_2, y_2)\; \cdots\; f(x_N, y_N) \;=\; \prod_{i=1}^{N} f(x_i, y_i)$$

or identically distributed but not independently

$$f(x_i, y_i) \;=\; f(x_j, y_j) \;\text{ for } i \neq j \;\text{ but }\; f(x_i, x_j, y_i, y_j) \neq f(x_i, y_i)\; f(x_j, y_j)$$

In the case where the observations are independently distributed, we often use the shorthand

$$f_i(x,y) \;\equiv\; f(x_i, y_i)$$

Note that the **DGP** for the size N sample of (X, Y) can always be represented as $f(x_1, \ldots, x_N, y_1, \ldots, y_N)$. It can only be simplified if the sample also consists of observations that are independently and/or identically distributed. In the case where the sample is random, the joint distribution requires only knowledge of the bivariate joint distribution $f(x, y)$.

A Note on Notation: $f(x_1, \ldots, x_N, y_1, \ldots, y_N)$ will sometime be written as $f(x_1, y_1, \ldots, x_N, y_N)$. Do not let the ordering of the random variables confuse you.

2.3 Importance of Being Random

Why are **random samples** so important? There are at least two reasons, a minor and a major one. The minor reason is computational simplicity. The joint distribution for the entire sample simplifies immensely when the sample is random

$$\begin{aligned} f(x_1, y_1, x_2, y_2, \ldots, x_N, y_N) &= \prod_{i=1}^{N} f(x_i, y_i) \quad \text{By Independence} \\ &= \prod_{i=1}^{N} f(x, y) \quad \text{By Identical Distribution} \end{aligned}$$

But this is not a major reason. As long as we know $f(x_1, y_1, \ldots, x_N, y_N)$, we can still derive and calculate the various best predictors and best linear predictors. Advances in computer power have made estimation with such distributions quite feasible – though possibly tedious.

The more important reason revolves around the situation where we do not know the full joint distribution of the sample. If we know a few aspects of the full joint distribution, especially for a random sample, we might know how the full joint distribution evolves as the sample size grows to infinity, $N \to \infty$.

The (limiting) joint distribution when the sample grows to infinite size is known as the **asymptotic distribution**. A popular convention is to use the asymptotic distribution as a close **approximation** to the actual, "finite sample" distribution. The idea is that if the sample is "large enough" the asymptotic distribution should resemble the actual distribution of the sample. This is something of a "hand waving" and a main criticism of frequentist statistical approaches by Bayesian statisticians.[1] We explore this criticism in more detail later.

2.4 Asymptotic Theory

There are 2 important asymptotic results associated with a random sample with finite mean and variance. Although I have defined both in terms of the univariate case, the theorems generalize to the multivariate case. For more details, see Chapters 8-10 in the Goldberger text.

Khinchine's Law of Large Number (LLN) : If $\{X_i\}$ for $i = 1, \ldots, N$ is a random sample from a population X with finite mean μ and finite variance σ^2 then the sample mean converges to the population mean μ: $\bar{X}_N \xrightarrow{p} \mu$

Lindeberg-Levy Central Limit Theorem (CLT) : If $\{X_i\}$ for $i = 1, \ldots, N$ is a random sample from a population X with finite mean μ and finite variance σ^2 then the standardized mean converges in distribution to the standard Normal: $\frac{\sqrt{N}\ (\bar{X}_N - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$

[1] The trade-off is uncertainty about speed of convergence to asymptotic distribution in classical settings and concerns about sensitivity to choice of prior in Bayesian settings

The notation $\xrightarrow{p}$ refers to "convergence in probability." In the context of Khinchine's LLN, $\bar{X}_N \xrightarrow{p} \mu$ refers to the sample mean $\bar{X}_N$ converging in probability to the population mean μ. More specifically, it means that for any $\epsilon > 0$ and $\delta > 0$, there exists some positive integer T such that for all $N > T$

$$P_X(\ |\bar{X}_N - \mu| < \epsilon\ ) \quad > \quad 1 - \delta$$

The above heuristically translates into the statement that there is some (finite) minimum sample size at which the sample mean and the population mean are probabilistically indistinguishable.

The notation $\xrightarrow{d}$ refers to "convergence in distribution." In the context of the Lindberg-Levy CLT, $\frac{\sqrt{N}\ (\bar{X}_N - \mu)}{\sigma} \xrightarrow{d} N(0,1)$ refers to the "standardized" version of the sample mean converging in distribution to a random variable with the standard normal distribution. More specifically, it means that

$$\lim_{N \to \infty} F_N(t) \quad = \quad \int_{-\infty}^{t} \frac{1}{\sqrt{2\Pi}}\ e^{-\frac{1}{2}z^2}\ dz \qquad \text{for all } t \text{ for which } F_N(t) \text{ is continuous}$$

where $F_N(t)$ is the cdf of $\frac{\sqrt{N}\ (\bar{X}_N - \mu)}{\sigma}$ evaluated at $\{\frac{\sqrt{N}\ (\bar{X}_N - \mu)}{\sigma} \leq t\}$.

Note that μ and $\sigma = \sqrt{\sigma^2}$ might not be known. Therefore, we use a "cheating" application of the change of variables and say that the sample mean is "asymptotically distributed" as

$$\bar{X}_N \quad \overset{a}{\sim} \quad N(\mu, \frac{\sigma^2}{N})$$

If $\frac{\sqrt{N}\ (\bar{X}_N - \mu)}{\sigma}$ had an *actual* distribution of N(0,1) rather than simply *converging in distribution* to N(0,1), then the above change of variables would be correct. The "rationale" for this asymptotic distribution is that for a sufficiently large sized sample, the actual distribution of the sample mean can be "approximated" by the asymptotic distribution.

This idea of using the asymptotic distribution as an approximation for the desired actual distribution is something of a "desperation" move. But it is a move commonly (ab)used. We will discuss the use of asymptotic distribution later.

For both the LLN and CLT, the only information we need about the population and sample are

- Sample must be random (distributed *i.i.d.*)
- Mean (first moment) must be well defined, finite, and known
- Variance (second central moment) must be well defined, finite, and known

Based on just this information, we are tempted to use the sample moments to approximate the population moments (due to the LLN) and the asymptotic distribution to approximate for the population distribution (due to the CLT) ... at least for "large" sized samples.

But what is missing from this discussion is a measure of how "large" the sample size (N) must be in order for the population distribution underlying the sample to resemble closely the asymptotic distribution of the sample. This is an issue of "speed of convergence." Unfortunately, without knowing

the true finite sample distribution (not just the asymptotic distribution), the speed of convergence cannot accurately be measured. The asymptotic equivalents may be a poor approximation of the finite sample counterparts even for rather large sized samples.

Concerns about the speed of convergence have not stopped statisticians/econometricians from using the sample moments to approximate for the population moment and the asymptotic distribution to approximate for the population distribution. Moreover, the following useful results courtesy of the **Slutsky Theorem** and **Delta Method** have furthered the practice:

- Slutsky Theorem
  1. If $X_n \xrightarrow{p} c$ and $h(X_n)$ is continuous at c then $h(X_n) \xrightarrow{p} h(c)$
  2. If $X_n \xrightarrow{p} c_1$ and $Y_n \xrightarrow{p} c_2$ and $h(X_n, Y_n)$ is continuous at (c_1, c_2) then $h(X_n, Y_n) \to h(c_1, c_2)$
  3. If $Z_n \xrightarrow{p} c$ and X_n has a limiting distribution, then the limiting distribution of $Z_n + X_n$ is the same as that of $c + X_n$
  4. If $Z_n \xrightarrow{p} c$ and X_n has a limiting distribution, then the limiting distribution of $Z_n\ X_n$ is the same as that of $c\ X_n$
- Delta Method: If $\sqrt{N}\ (X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ and $Y_n = h(X_n)$ is continuously differentiable at μ then $\sqrt{N}\ (Y_n - h(\mu)) \xrightarrow{d} N(0, [h'(\mu)]^2\sigma^2)$

The above extends the approximation technique from X to continuous functions of X, $h(X)$.

3 Principles of Estimation

In frequentist statistics, we are left with resorting to the asymptotic approximation in cases where we do not know the true sample distribution. While a random sample is desirable, the random sampling assumption can be relaxed.[2] That said, we will work primarily with random samples (or random sample after stratification) in this course. Given a random sample, we want to think about how we can use the random sample to infer information about the DGP underlying the sample. There are two major principles that guide most estimation procedures based on random samples:

Analogy (Moments) Principle

This is, perhaps, the oldest and most intuitive. The idea is that we set the sample moments equal to their respective population moments. So the sample mean would be equated to the population mean ($\frac{1}{N}\sum_{i=1}^{N} X_i = \mu_X$) and so forth. The LLN provides some justification for this approach. The estimation procedure based on this principle is known as the **method of moments**.

Likelihood Principle

This popular approach is based on the idea that we choose the distribution that associates the largest "likelihood" of observing the given random sample. This principle requires more information than the analogy principle: we must know not only the moments but the form of the distribution (up to some parameters). Consequently, some people will rely upon the CLT and use the asymptotic distribution (Normal) to approximate the form of the distribution. The estimation procedure based on this principle is referred to as **maximum likelihood**.

A "Non-Parametric" Footnote

The convention adopted by most econometricians is that we "know" the DGP up to some parametric form. This means that we "know" the functional form of the distribution but not the values of the parameters that define the function. For example, we might assume that the sample $\{X_i, Y_i\}$ is distributed Normal but remain agnostic about the parameters of this multivariate Normal.

That said, it is possible to do statistical inference while being agnostic even about the general form of the DGP. These methods, depending on the degree to which we remain agnostic, are known as **"non-parametric"** or **"semi-parametric"** approaches. These approaches use the observed **sample** of realizations from the DGP to infer properties of the DGP. In a parametric approach, we use the sample to infer the values of the parameters governing the assumed distribution of the DGP. In non-parameteric and semi-parametric approaches, we use the sample to infer more general characteristics of the DGP.

A good primer on "non-parametric" and "semi-parametric" approaches is

- "Nonparametric Density and Regression Estimation," by John DiNardo and Justin Tobias, *Journal of Economic Perspective*, Vol. 15 No. 4 (Fall 2001), pp.11-28

[2] There are versions of the LLN and CLT under other sets of assumptions. For example, the "identically distributed" assumption can be substituted with assumptions concerning the first three moments of the distribution.

3.1 Example: Coin Flip

Consider a *random* sample consisting of N flips of the same coin: $\{ X_1 = x_1 \ , \ \ldots \ , \ X_N = x_N \}$

$$\text{Let } X_i \ = \ \begin{cases} 1 & \text{if Heads on } i^{th} \text{ flip} \\ \\ 0 & \text{if Tails on } i^{th} \text{ flip} \end{cases} \quad \text{and} \quad f(x_i) \ = \ \begin{cases} \theta & \text{if } x_i = 1 \\ (1-\theta) & \text{if } x_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

So θ, the probability of landing Heads, is the unknown parameter of the DGP.

Note that this random sample is generated from a sample with finite mean and variance. So the LLN and CLT can be used.

Analogy (Moment) Principle

Note that

$$E[X] \ = \ \sum_{x=0}^{1} x \cdot f(x) \ = \ 0 \cdot (1-\theta) + 1 \cdot \theta \ = \ \theta$$

So the population mean of X represents the unknown parameter θ.

Using the analogy principle, we set the *sample* mean of X equal to the population mean of X

$$\frac{1}{N} \sum_{i=1}^{N} X_i \ = \ E[X] \ = \ \theta$$

Therefore, a method of moment estimator of θ that can be rationalized using the LLN is

$$\hat{\theta} \ = \ \frac{1}{N} \sum_{i=1}^{N} X_i$$

Likelihood Principle

The sampling distribution of our random sample of coin flips is

$$\begin{aligned} f(x_1, c \ldots, x_N) \ &= \ \Pi_{i=1}^{N} f(x_i) \qquad \text{by independence} \\ &= \ (\theta)^M \ \ (1-\theta)^{N-M} \qquad \text{by identical distribution} \end{aligned}$$

Let M be the number of flips landing Heads and, therefore, $N - M$ the number of flips landing Tails

The Likelihood function for this sample is

$$L(\theta; x_1 \ldots x_N) \ = \ (\theta)^M \ \ (1-\theta)^{N-M}$$

and the maximum likelihood estimator of θ is the value of θ that maximizes the above Likelihood function

First order condition (FOC)

$$\begin{aligned}\frac{\partial L(\theta; x_1 \ldots x_N)}{\partial \theta} &= M(\theta)^{M-1}(1-\theta)^{N-M} + (N-M)(-1)(\theta)^{M}(1-\theta)^{N-M-1} = 0 \\ &= (\theta)^{M-1}(1-\theta)^{N-M-1}\ (M - N\theta) = 0\end{aligned}$$

There are three values of θ that satisfies the above FOC. Two of the values, $\theta = 0$ and $\theta = 1$, are values that *minimize* the likelihood.[3] The third, $\theta = \frac{M}{N}$ is the value that *maximizes* the likelihood.

So the maximum likelihood estimator of θ is

$$\hat{\theta}_{ML} = \frac{M}{N}$$

Note that $\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{M}{N}$. So both the moment-based estimator $\hat{\theta}$ and the likelihood based estimator $\hat{\theta}_{ML}$ are the same!

This is because, for this particular random experiment, the moment used to build $\hat{\theta}$ has the same information as the sampling likelihood used to build $\hat{\theta}_{ML}$. The sampling distribution is entirely defined by the single parameter θ and the single parameter θ is nailed down by the single moment $E[X]$. In general, estimators derived from the two principles will differ as the information contained in the moments will not be the same as the information contained in the likelihood.

[3] You can show this using the second order condition. Or you can intuit it – $\theta = 1$ or 0 implies that the coin always lands Heads or always lands Tails ... which means any sample involving both Heads and Tails have zero likelihood

3.2 Example: Ordinary Least Squares (OLS)

Consider a size N sample from the population (X, Y)

$$\{ \ (X_1 = x_1, Y_1 = y_1) \ \cdots \ (X_N = x_N, Y_N = y_N) \ \}$$

The German mathematician Carl Friedrich Gauss proposed the following predictor of Y given X using the above sample: $\hat{Y}(X) = a_{ols} + b_{ols}X$ where

$$(a_{ols}, b_{ols}) \quad = \quad \text{argmin}_{a,b} \qquad \underbrace{\sum_{i=1}^{N} \ [\ Y_i - \hat{Y}(X_i) \ ]^2}_{\text{Sum of Squared Residuals (SSR)}}$$

Gauss proposed that we use a *linear* predictor of Y given X that minimizes the sum of squared errors across the sample.[4] This approach has since become known as **Ordinary Least Squares (OLS)** and is currently the workhorse model in economics and many other empirical disciplines.

Note that

$$SSR \quad = \quad \sum_{i=1}^{N} \ [\ Y_i - \hat{Y}(X_i) \ ]^2 \quad = \quad \sum_{i=1}^{N} \ [\ Y_i^2 - 2Y_i(a + bX_i) + (a + bX_i)^2 \ ]$$

Let us solve for a_{OLS} and b_{OLS} using the appropriate first order conditions

$$\begin{aligned}
\frac{\partial SSR}{\partial a} \quad &= \quad -2\sum_{i=1}^{N} Y_i + 2\sum_{i=1}^{N}(a + bX_i) \quad = \quad 0 \\
&\Rightarrow \quad 2Na \quad = \quad 2\left(\sum_{i=1}^{N} Y_i - b\sum_{i=1}^{N} X_i\right) \\
&\Rightarrow \quad a_{ols} \quad = \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N} Y_i}_{\bar{Y}_N} \ - \ b_{ols}\underbrace{\frac{1}{N}\sum_{i=1}^{N} X_i}_{\bar{X}_N} \quad = \quad \text{Sample Mean of } Y - b_{ols} \times \text{Sample Mean of X} \\
\frac{\partial SSR}{\partial b} \quad &= \quad -2\sum_{i=1}^{N} X_i Y_i + 2\sum_{i=1}^{N} X_i(a + bX_i) \quad = \quad 0 \\
&\Rightarrow \quad 2a\sum_{i=1}^{N} X_i + 2b\sum_{i=1}^{N} X_i^2 \quad = \quad 2\sum_{i=1}^{N} X_i Y_i \\
&\Rightarrow \quad b\sum_{i=1}^{N} X_i^2 \quad = \quad \sum_{i=1}^{N} X_i Y_i - \underbrace{\left(\bar{Y}_N - b\bar{X}_N\right)}_{a_{OLS}}\sum_{i=1}^{N} X_i \\
&\Rightarrow \quad b_{ols} \quad = \quad \frac{\frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}_N\bar{Y}_N}{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - (\bar{X}_N)^2} \quad = \quad \frac{\text{Sample Covariance of } X, Y}{\text{Sample Variance of } X}
\end{aligned}$$

[4]The above representation is for the case of a bivariate population; the approach can be extended to the general multivariate case, as will be seen later in the course

Note the similarity between the OLS estimator (a_{ols}, b_{ols}) and the parameters of the best linear predictor under mean squared error $BLP_{MSE}(Y|X)$. The OLS estimator can be thought of as the moment-based estimator of the best linear predictor where we substitute the required population moments with their sample analogs

$$\begin{aligned} \text{Population} &\iff \text{Sample} \\ E[X] &\iff \bar{X}_N \\ E[Y] &\iff \bar{Y}_N \\ E[X^2] &\iff \frac{1}{N}\sum_{i=1}^{N} X_i^2 \\ E[XY] &\iff \frac{1}{N}\sum_{i=1}^{N} X_i Y_i \end{aligned}$$

Recall that $\text{Var}(X) = E[X^2] - (E[X])^2$ and, similarly, $\text{Cov}(X,Y) = E[XY] - E[X]\ E[Y]$

The OLS estimator may also be thought of as a maximum likelihood estimator if the sample is random and (X, Y) are jointly distributed Normal – which implies that the conditional distribution of Y given X is Normal too.[5]

$$f(y_i \mid x_i) = \frac{1}{\sqrt{2\Pi\sigma^2_{y|x}}}\, e^{-\frac{\left(y_i - \mu_{y|x}\right)^2}{2\sigma^2_{y|x}}}$$

$$\text{where } \mu_{y|x} = \underbrace{\left(\mu_Y - \frac{\sigma_{XY}}{\sigma^2_X}\mu_X\right)}_{\alpha} + \underbrace{\frac{\sigma_{XY}}{\sigma^2_X}}_{\beta} x_i \text{ and } \sigma^2_{y|x} = \sigma^2_Y - \underbrace{\left(\frac{\sigma_{XY}}{\sigma^2_X}\right)^2}_{\beta^2}\sigma^2_X$$

As will be explicitly shown later, the OLS estimators are, in this case, the maximum likelihood estimators of (α, β).

[5]This result is discussed in various textbooks, including Amemiya Chapter 5 and Goldberger Chapter 7