Economics 361
Estimation

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

1 Overview

If the data generating process (DGP) is fully known, then statistical inference consists primarily of probability theory applications. The best predictor and best linear predictor exercises demonstrate this point. However, if the data generating process is known only up to some **parameters** – the parameters themselves unknown – then statistical inference requires data. The required data is namely a **sample** from the population characterizing the DGP.

The practice of using the sample to arrive at values of the unknown DGP parameters is **estimation.**

- **Point Estimation**: when we arrive at a **single value** for each unknown parameter
- **Interval Estimation**: when we arrive at a **set of values** for each unknown parameter

In this course, we will focus on point estimation.[1]

Note the similarity between estimation and prediction. In the latter, we use the sample to arrive at some value of the realization of the unobserved random variable Y. In the former, we use the sample to arrive at some value of the unobserved parameter governing the probabilistic relationship among the population random variables, e.g. X and Y.

As shown earlier, prediction depends on the parameters – neither $BP_{MSE}(Y|X)$ nor $BLP_{MSE}(Y|X)$ can be calculated without the parameters. So estimation precedes prediction in scenarios where the DGP is not fully known. We estimate the unknown parameters and **then** predict the values of the unknown Y using the estimated parameters. Clearly, the precision with which we predict Y will depend on the precision with which we estimate the parameters.

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

[1] Point versus interval estimation hinges on the issue of identification, which we discuss below. If the sample can confidently pin down a single value for each parameter, then point estimation is preferred. If the sample cannot or can only under fairly restrictive assumptions, then interval estimation is preferred

2 Estimation Precision

Earlier, we discussed two key methods used to arrive at parameter estimates

- **Analogy (Moment) Principle**: Method of Moments (or moment-based) Estimators
- **Likelihood Principle**: Maximum Likelihood (or likelihood-based) Estimators

We loosely discussed how to rationalize the estimators arrived by each principle. But we did not discuss the precision with which these estimators estimate the unknown parameters. The precision of the estimators will depend, generally, on

- Size and Type of the Sample
- Properties of DGP: joint distribution of the population random variables

We explore these issues using the multivariate ordinary least squares (OLS) model for some population $\{Y, X_1, \ldots, X_{k-1}\}$

2.1 Finite Random Sample

Consider some size N random sample of $\{Y, X_1, \ldots, X_{k-1}\}$ where $N \ll \infty$ (**finite sample**).

As the sample is random, the 2^{nd} Gauss Markov assumption (Spherical Errors) is satisfied: $\text{Var}(Y|X) = \sigma^2 I$. Let us consider the case where the 1^{st} Gauss Markov assumption (Linearity) is also satisfied:

$$E[Y|X] = X\beta$$

This suggests that the multivariate OLS model can be used to estimate β, the parameters of the $BP_{MSE}(Y|X)$.

$$b^{ols} = (X'X)^{-1}X'Y$$

where X is the $(N \times k)$ matrix that includes the vector on ones (ι).

The above analytical form of b^{ols} exists only if $(X'X)^{-1}$ exists. $(X'X)^{-1}$ exists only if X consists of linearly independent columns. This is essentially the 3^{rd} Gauss Markov assumption (full rank). Without a full rank X, b^{ols} is not well-defined.

To see this, consider the k first order conditions associated with b^{ols}

$$\frac{\partial SSR}{\partial b} = \begin{pmatrix} \frac{\partial SSR}{\partial b_0} \\ \vdots \\ \frac{\partial SSR}{\partial b_{k-1}} \end{pmatrix} = \underbrace{\begin{pmatrix} -2\sum_{i=1}^{N} X_{0i}(Y_i - \sum_{j=0}^{k-1} b_j X_{ji}) \\ \vdots \\ -2\sum_{i=1}^{N} X_{(k-1)i}(Y_i - \sum_{j=0}^{k-1} b_j X_{ji}) \end{pmatrix}}_{-2(X'Y)+2(X'X)b} = 0$$

We have the classic k linear equations for k parameters problem. As long as each equation is linearly independent, this system of linear equations can be solved to arrive at a single value for

each parameter estimate b_j. But if any of the linear equations (first order conditions) are linearly **de**pendent of the others, we have fewer independent equations than parameters to solve; we cannot assign a unique value to each parameter estimate b_j. The system of equations is **under-identified**.

The linear independence of the k first order conditions translate into the linear independence of the k column vectors in X. Therefore, if X is full rank, $\text{rank}(X) = k$, then

- k First Order Conditions (FOCs) are linearly independent
- $b^{ols} = (X'X)^{-1}X'Y$ is well defined $\Longrightarrow$ b^{ols} is **identified**

The 1^{st} (linearity) and 3^{rd} (full rank) Gauss-Markov assumptions are assumptions concerning the joint distribution of the population random variables $(Y, X_1, \ldots, X_{k-1})$.[2] The 2^{nd} assumption is an assumption concerning the sample – one that is satisfied if (but not only if) the sample is random.

Under all three Gauss Markov assumptions, we can show that

$$\begin{aligned}
\underbrace{E[b^{ols}|X]}_{\text{cond. mean}} &= E[(X'X)^{-1}X'Y|X] = (X'X)^{-1}X'\underbrace{E[Y|X]}_{X\beta} \\
&= \underbrace{(X'X)^{-1}X'X}_{I}\beta = \beta \quad \text{(Unbiasedness)} \\
\text{Var}(b^{ols}|X) &= \text{Var}((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'\underbrace{\text{Var}(Y|X)}_{\sigma^2 I}X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} = \underbrace{\sigma^2 (X'X)^{-1}\underbrace{X'IX}_{(X'X)}(X'X)^{-1}}_{I} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

Furthermore, from the Gauss Markov Theorem, we know that there are no linear unbiased estimators of β (conditional on X) that has a lower variance than $\sigma^2(X'X)^{-1}$.

What do the above statements tell us about the precision of b^{ols}? To answer this question, recall the notion of probability (frequentist) we are using [3]

> *Frequentist Probability*: Let N be the number of times the random experiment is repeated and N_A the number of times event A occurred in the repeated experiments. Then the probability of A is defined by $P(A) = \lim_{N\to\infty} \frac{N_A}{N}$

The above statements about the conditional mean and variance of b^{ols} are statements concerning the outcome from an *infinite* repetition of the controlled experiment, each repetition resulting in a size N sample and each sample yielding its own b^{ols} estimate of β.[4]

[2] It is technically possible for the sampled X not to be linearly independent even if the underlying population variables themselves are linearly independent. But this is a measure zero event – effectively zero probability

[3] N in the frequentist probability definition is not the same as the sample size N

[4] We are examining the conditional probability. So for each of these repetitions, we fix the value of X and allow only Y to be re-sampled. Thus, "controlled experiment" where the value of X is controlled. If we allow X to be re-sampled as well, then we are examining the *un*conditional probability

The unbiasedness result concerns the average value of b^{ols} across the repetitions and the variance result the spread of these different b^{ols} values across the repetitions. As the number of repetitions grows to infinity, the average value of b^{ols} converges to β and the spread, around the average, converges to $\sigma^2(X'X)^{-1}$.[5] The statements above tell us something about the precision of our estimator under infinite repetitions of the underlying (controlled) random experiment.[6]

- If $E[b^{ols}|X] \neq \beta$ then we say that b^{ols} is a biased estimate of β – one whose average value across infinite repetitions does not converge to the desired parameter value.
- $\text{Var}(b^{ols}|X)$ is an important value as it tells us how much any specific b^{ols} from a given repetition may differ from the average (again, under infinite repetitions). Larger values of $\text{Var}(b^{ols}|X)$ indicate that a specific b^{ols} can, probabilistically, stray farther from $E[b^{ols}|X]$

So, in terms of precision, we prefer an estimator that is unbiased with low variance. This is why the Gauss Markov Theorem plays an important role in many econometrics courses. It states that under the Gauss Markov assumptions, b^{ols} is the unbiased linear estimator of β that achieves the lowest variance.[7]

2.2 Asymptotic Random Sample

Consider, instead, the precision of b^{ols} given an "infinite sized" random sample.[8] This "asymptotic" random sample presents both complications and simplifications. The complications have to do with the values of $(X'X)^{-1}$ and $(X'Y)$.

Suppose that the third Gauss Markov assumption holds: $\text{rank}(X) = k$. $b^{ols} = (X'X)^{-1}X'Y$ may still not be well-defined as the values of $(X'X)^{-1}$ and/or $(X'Y)$ may not be well defined with an infinite number of observations. Elements of both $(X'X)$ and $(X'Y)$ involve the sum of an infinite number of terms. Even if each term is finite, the sum of an infinite number of finite terms may be infinite (or negatively infinite).

Therefore, we need additional assumptions concerning $(X'X)$ and $(X'Y)$ as the sample size $N \to \infty$. Specifically,

$$\lim_{N\to\infty} \frac{(X'X)}{N} = Q_{XX} \qquad \lim_{N\to\infty} \frac{(X'Y)}{N} = Q_{XY}$$

where Q_{XX} and Q_{XY} are some finite matrices (matrices whose elements consist only of finite values) and Q_{XX} is invertible (Q_{XX}^{-1} exists). Then

$$\lim_{N\to\infty} \underbrace{(X'X)^{-1}X'Y}_{b^{ols}} = \lim_{N\to\infty} \left(\frac{(X'X)}{N}\right)^{-1} \frac{(X'Y)}{N} = Q_{XX}^{-1}\, Q_{XY}$$

[5] To see this, consider a discrete random variable Y that takes on one of two values (y_A, y_B). $E[Y|X] = y_A\ P(y = y_A|X) + y_B\ P(y = y_B|X) = y_A \left(\lim_{N\to\infty} \frac{N_A}{N}\right) + y_B \left(\lim_{N\to\infty} \frac{N_B}{N}\right) = \lim_{N\to\infty} \frac{N_A y_A + N_B y_B}{N}$ which is the average of y across the N repetitions as $N \to \infty$. This intuition can be generalized.

[6] In this manner, they are *meta* analysis statements

[7] It may be possible to achieve a lower variance – but only with a *biased* estimator.

[8] More specifically, a random sample whose sample size asymptotes to infinite

The simplification has to do with the stochastic properties of Q_{XX} and Q_{XY}. Waving our hands a bit (using asymptotic theory), we can show that the above conditions concerning Q_{XX} and Q_{XY} combined with assumptions about the first two moments of the population random variables (necessary for Khinchine's Law of Large Numbers)

$$b^{ols} \xrightarrow{p} \beta$$

The ols estimator b^{ols} **converges in probability** to β. In the language of statistics, b^{ols} is a **consistent** estimator of β. This consistency result is not the same as the earlier unbiasedness result. The unbiasedness result concerned b^{ols} calculated using a finite sample. The consistency result above concerns b^{ols} calculated using an infinite sample.

[**ASIDE:** Why does this convergence in probability occur? Heuristically ... as the number of observations goes to infinity, every possible realization of $(Y, X_1, \ldots, X_{k-1})$ is represented in the random sample and in the proportions defined by the joint distribution $f(y, x_1, \ldots, x_{k-1})$. So the sample means collapse to the actual population means.]

When we have all three Gauss Markov assumptions, including linearity, the above consistency result is not very useful as we, in practice, deal with finite samples. However, the result is very useful if we are not sure that the linearity condition holds. Without the linearity condition, we cannot derive the unbiasedness result for b^{ols}. Moreover, we cannot define what β is. A non-linear conditional mean, $E[Y|X] \neq X\beta$, invalidates the interpretation of β as the coefficients of the linear $E[Y|X]$.

But with an asymptotic sample, we can re-interpret β. Instead of β as the parameters of the $BP_{MSE}(Y|X) = E[Y|X]$, we interpret β as the parameters of the $BLP_{MSE}(Y|X) = X\beta$. Note that $X\beta$ subsumes the intercept. So

$$b^{ols} \xrightarrow{p} \beta \quad \text{(the coefficients of the } BLP_{MSE}(Y|X)\text{)}$$

This result is explicitly developed in the "Regression Algebra" Handout.

Unfortunately, we do not have a general unbiasedness result concerning β the coefficient of $BLP_{MSE}(Y|X)$. If the linearity condition does not hold, then $E[b^{ols}|X] \neq \beta$ (the coefficients of the $BLP_{MSE}(Y|X)$).

> **Key Takeaway Point:** Without the linearity condition, we cannot generally discuss the unbiasedness of b^{ols}, to either the $BP_{MSE}(Y|X)$ or $BLP_{MSE}(Y|X)$. We can discuss the consistency of b^{ols} to the coefficients of the $BLP_{MSE}(Y|X)$.

We can also make some statement about $\text{Var}(b^{ols}|X)$ as $N \to \infty$.

$$\lim_{N\to\infty} \text{Var}(b^{ols}|X) \quad = \quad \lim_{N\to\infty} \sigma^2 \left(X'X\right)^{-1} \quad = \quad \sigma^2 \lim_{N\to\infty} \left(X'X\right)^{-1} \quad \approx \quad \sigma^2 \underbrace{\frac{Q_{XX}^{-1}}{N}}_{\to 0?}$$

We revisit this strange result when we discuss the "asymptotic distribution" of b^{ols}.

3 Distribution of an Estimator

In the first half of the 20^{th} Century, William Gosset and Ronald Fisher effected a revolution in statistics. In modern terms, their key insight can be summed up as follows:

- Estimators (and, as will be shown later, test statistics) are functions of the sample
- The sample consists of random variables and are governed by a sampling distribution
- So the estimators, as functions of random variables, must themselves be random variables with some distribution

Gosset's famous "Student t Distribution" article concerns the distribution of the moment-based estimator of $E[X]$ based on a size N random sample $\{X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N\}$ where the sampling distribution $f(x_1, x_2, \ldots, x_N)$ is multivariate Normal. This article by Gosset led to Fisher (and then other statisticians) to derive the distribution of other estimators (and test statistics) under various sampling distributions.[9]

We demonstrate the usefulness of this insight – the distribution of an estimator – using the multivariate OLS model.

3.1 OLS Model under Normality (Known σ^2)

Consider the OLS model with a finite random sample and where all three Gauss Markov assumptions hold. We showed that

- From full rank assumption, $b^{ols} = (X'X)^{-1}X'Y$
- From the full rank and linearity assumptions, $E[b^{ols}|X] = \beta$ (the coefficients of $E[Y|X]$)
- From the full rank and spherical errors assumptions, $\text{Var}(b^{ols}|X) = \sigma^2(X'X)^{-1}$
- From all three, b^{ols} is the best (minimum variance) linear unbiased estimator of β

Suppose we add one more assumption

- Normality: the conditional distribution of Y given X, where Y and X refers to the population random variables, is Normal

This (along with linearity and spherical errors) implies that

$$f(y|x_1, \ldots, x_{k-1}) \quad = \quad \frac{1}{\sqrt{2\Pi\sigma^2}}\, e^{-\frac{(y-x\beta)^2}{2\sigma^2}}$$

[9] For an accessible account of the the history of Gosset and Fisher, see "Gosset, Fisher, and the t Distribution," by Joan Fisher Box, *The American Statistician*, May 1981, 35(2), pp.61-66

If we combine this fourth assumption with an implicit fifth – the sample is random – we can derive the samplign distribution as

$$\begin{aligned} f(y_1, \ldots, y_N|X) &= \prod_{i=1}^{N} f(y_i|X) \quad \text{from independently distributed} \\ &= \prod_{i=1}^{N} f(y|X) \quad \text{from identically distributed} \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\Pi\sigma^2}} \, e^{-\frac{(y-x\beta)^2}{2\sigma^2}} \quad \text{from Normality assumption} \\ &= \left(\frac{1}{\sqrt{2\Pi\sigma^2}}\right)^N \, e^{-\frac{\sum_{i=1}^{N}(y_i - x_i\beta)^2}{2\sigma^2}} \end{aligned}$$

The above distribution is an example of the **multivariate Normal** distribution. The general version of the multivariate normal distribution (in matrix notation) for an $(N \times 1)$ vector of random variables Y is

$$f(y) = (2\Pi)^{-\frac{N}{2}} \, |\Sigma|^{-\frac{1}{2}} \, e^{\frac{-(y-\mu)'\Sigma^{-1}(y-\mu)}{2}}$$

where μ is an $(N \times 1)$ vector and Σ an $(N \times N)$ matrix. $|\Sigma|$ refers to the determinant of the matrix Σ. So $|\Sigma|^{-\frac{1}{2}}$ is the inverse of the square root of the determinant of Σ.[10]

If Y has the above distribution, we say that Y is distributed multivariate Normal with mean μ and variance/covariance matrix Σ: $Y \sim N(\mu, \Sigma)$. Goldberger Chapter 18 further discusses the multivariate Normal distribution, including some useful properties of the distribution.

Therefore, $Y|X$ is said to have a conditional distribution of $N(\mu, \Sigma)$ if

$$f(y|x) = (2\Pi)^{-\frac{N}{2}} \, |\Sigma|^{-\frac{1}{2}} \, e^{\frac{-(y-\mu)'\Sigma^{-1}(y-\mu)}{2}}$$

In terms of the OLS model under Normality and random sample, μ corresponds to $x\beta$ and Σ to $\sigma^2 I_N$. One can show that $|\sigma^2 I_N|^{-\frac{1}{2}} = \left(\frac{1}{\sqrt{\sigma^2}}\right)^N$

Recall that $b^{ols} = (X'X)^{-1}X'Y = AY$ where we use Goldberger's $A = (X'X)^{-1}X'$ matrix. So b^{ols} is a linear transformation of Y. Using the multivariate version of the "change of variables" result we showed earlier, we can show that b^{ols} is itself distributed multivariate Normal.[11]

$$f(b|X) = (2\Pi)^{-\frac{N}{2}} \, \left|\sigma^2(X'X)^{-1}\right|^{-\frac{1}{2}} \, e^{\frac{-(b-\beta)'\left(\sigma^2(X'X)^{-1}\right)^{-1}(b-\beta)}{2}}$$

In other words, $b^{ols}|X \sim N(\beta, \sigma^2(X'X)^{-1})$

[10] For this course, you do not need to "know" how to calculate determinants of a matrix. For those interested, see Amemiya Chapter 11 or the appropriate chapter in a linear algebra textbook

[11] Goldberger Chapter 18.2 (p.198) states this result fully

3.2 OLS Model under Normality (Unknown σ^2)

Note that the Normal distribution has two sets of parameters, the mean and the variance. In the multivariate setting, the two sets of parameters are the vector of means and the variance/covariance matrix. Thus far, we have only discussed the OLS estimator for the vector of means.

This implies that we must either know the value of $\sigma^2(X'X)^{-1}$ or arrive at some estimator of $\sigma^2(X'X)^{-1}$. We observe X and therefore know $X'X)^{-1}$. So the issue boils down to whether we know σ^2 or need an estimator for σ^2. For the latter, the estimator of σ^2 usually paired with b^{ols} is

$$s^2 \quad = \quad \frac{e'e}{N-k} \quad \text{where } e \equiv Y - X\ b^{ols}$$

$X\ b^{ols}$ is the OLS estimator of $E[Y|X]$. So e is the difference between the actual realized values Y and their corresponding OLS estimates of $E[Y|X]$. e is sometimes referred to as the **residual**. e is not the same as the "residuals" $\epsilon \equiv Y - X\ \beta$ introduced earlier. They are, however, analogous.

Consider

$$\text{Var}(\epsilon|X) \quad = \quad \text{Var}(Y - X\beta|X) \quad = \quad \text{Var}(Y|X) \quad = \quad \sigma^2$$

Here (again, abuse of notation) ϵ refers to the difference between the population random variable Y and the population random variables X matrix multiplied by β.

So a moment-based estimator of σ^2 may involve the above moment condition, replacing β with b^{ols}

$$\text{Var}(\epsilon|X) \quad = \quad E[\epsilon^2|X] \quad = \quad E[(Y - X\beta)^2|X] \quad \Longrightarrow \quad \frac{1}{N}\sum_{i=1}^{N}(\ \underbrace{Y_i - X_i b^{ols}}_{e_i}\ )^2 \quad = \quad \frac{e'e}{N}$$

The difference between the above estimator of σ^2 and s^2 is the denominator (N instead of $N-k$).[12] The two estimators are asymptotically equivalent as $\lim_{N\to\infty} \frac{N-k}{N} = 1$. But for finite samples, s^2 is preferred. This preference stems from the following result obtained via "change of variables"

$$\frac{e'e}{\sigma^2} \quad \sim \quad \chi^2_{N-k}$$

The ratio of the sum of squared residuals (SSR) evaluated using b^{ols} and the conditional variance of Y, σ^2, is distributed chi-squared with $N-k$ degrees of freedom.[13] This result can be combined with another result (again, change of variables) to show that

$$\frac{b_j^{ols} - \beta_j}{\sqrt{s^2(X'X)^{-1}_{jj}}} \quad \sim \quad t_{N-k} \quad \text{where } (X'X)^{-1}_{jj} \text{ is the (j,j) element of matrix } (X'X)^{-1}$$

where b_j^{ols} refers to the OLS estimate for the specific coefficient β_j.[14] t_{N-k} designates the "t" distribution with $N-k$ degrees of freedom.

[12] This difference was, in fact, the impetus for the initial correspondence between Gosset and Fisher.
[13] Goldberger Chapter 21.1 outlines the proof
[14] Goldberger Chapter 21.1 outlines the proof

In summary

- When we have all three Gauss Markov assumptions, the Normality assumption, and a random sample, we can show that $b^{ols} \sim N(\beta, \sigma^2(X'X)^{-1})$
- If σ^2 is unknown, we can use $s^2 = \frac{e'e}{N-k}$ and show that $\frac{b^{ols}-\beta_j}{\sqrt{s^2(X'X)^{-1}_{jj}}} \sim t_{N-k}$

3.3 OLS Model under Asymptotic Normality

If the Normality assumption cannot be maintained, then we may still be able to derive the distribution of b^{ols} but only for infinite sized (asymptotic) samples. With asymptotic samples, we can use the Central Limit Theorem (CLT).

If all three Gauss Markov assumptions are satisfied and we have a random sample, the multivariate version of the CLT can be used to show that

$$b^{ols} \quad \overset{a}{\sim} \quad N(\beta, \frac{\sigma^2}{N} Q_{XX}^{-1})$$

if $\lim_{N\to\infty} \frac{X'X}{N}$ and $\lim_{N\to\infty} \frac{X'Y}{N}$ converge to finite matrices Q_{XX} and Q_{XY} and Q_{XX} is invertible.

So if

- all three Gauss Markov assumptions are satisfied
- sample is random
- $\lim_{N\to\infty} \frac{X'X}{N}$ and $\lim_{N\to\infty} \frac{X'Y}{N}$ converge to finite matrices Q_{XX} and Q_{XY} and Q_{XX} is invertible

then the "asymptotic distribution" of b^{ols} is multivariate Normal with mean β and variance/covariance matrix $\frac{\sigma^2}{N} Q_{XX}^{-1}$.

Recall that the "asymptotic distribution" is an *approximation* we use when we have "large" sized samples. $\frac{\sigma^2}{N} Q_{XX}^{-1}$ is, itself, approximated by $\sigma^2(X'X)^{-1}$ ... which is essentially the value of $\frac{\sigma^2}{N} Q_{XX}^{-1}$ at finite values of N. So $\frac{\sigma^2}{N} Q_{XX}^{-1}$ is never meant to be evaluated at $N \to \infty$