Economics 361
Regression Algebra

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

1 Overview

We presented the bivariate version of the ordinary least squares model:

$$(a_{ols}, b_{ols}) \quad = \quad \text{argmin}_{a,b} \sum_{i=1}^{N} \; [\; Y_i - (a + bX_i) \;]^2$$

Consider now the multivariate version

Suppose we are faced with a population consisting of k random variables. Let those k random variables be denoted $(Y, X_1, X_2, \ldots, X_{k-1})$. We have a size N sample from this population

$$\{ \; (Y_1 = y_1, X_{11} = x_{11} \; , \; \ldots \; , \; X_{(k-1)1} = x_{(k-1)1}) \; \cdots \; (Y_N = y_N, X_{1N} = x_{1N} \; , \; \ldots \; , \; X_{(k-1)N} = x_{(k-1)N}) \; \}$$

In an abuse of notation, the sample is often depicted using the following $(N \times 1)$ column vectors

$$Y \quad = \quad \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \qquad X_j \quad = \quad \begin{pmatrix} X_{j1} \\ \vdots \\ X_{jN} \end{pmatrix}$$

In addition, we often combine the latter $k-1$ random variables with a $(N \times 1)$ column vector of ones (ι) to form a $(N \times k)$ matrix

$$\iota \quad = \quad \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad X \quad = \quad (\; \iota \; X_1 \; X_2 \; \cdots \; X_{k-1} \;) \quad = \quad \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{(k-1)1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \ldots & X_{(k-1)N} \end{pmatrix}$$

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

Note that

$$X'Y = \begin{pmatrix} 1 & \cdots & 1 \\ X_{11} & \cdots & X_{1N} \\ \vdots & \ddots & \vdots \\ X_{(k-1)1} & \cdots & X_{(k-1)N} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} X_{1i} Y_i \\ \vdots \\ \sum_{i=1}^{N} X_{(k-1)i} Y_i \end{pmatrix}$$

For notational simplicity, let $X_{0i} = 1$ for $i = 1$ to N and $X_0 = \iota$

We can now express the multivariate ordinary least squares model as

$$\begin{aligned} b^{ols} &= \text{argmin}_b \underbrace{\sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{k-1} b_j X_{ji} \right)^2}_{SSR} \\ &= \text{argmin}_b \ (Y - Xb)'(Y - Xb) \\ &= \text{argmin}_b \ \left(Y'Y - Y'Xb - b'X'Y + b'X'Xb \right) \end{aligned}$$

where both b and b^{ols} are $k \times 1$ column vectors

$$b = \begin{pmatrix} b_0 \\ \vdots \\ b_{k-1} \end{pmatrix} \quad \text{and} \quad b^{ols} = \begin{pmatrix} b_0^{ols} \\ \vdots \\ b_{k-1}^{ols} \end{pmatrix}$$

Consider the relevant first order conditions (in vector form)

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

Note: the 0 above is technically $\vec{0}$, a $k \times 1$ column vector consisting of zeroes.

2 Multivariate BP, BLP, and OLS

Let us first consider the "best predictor" under the MSE criterion. The best predictor of Y given X, $\tilde{Y}(X)$, is defined as

$$\tilde{Y}(X) \equiv \text{argmin}_{\tilde{Y}(X)} \underbrace{E[\,(\,Y-\hat{Y}(X)\,)'\,(\,Y-\hat{Y}(X)\,)\,]}_{\text{LF(MSE)}}$$

and X only refers to the $k-1$ random variables $\{X_1 \cdots X_{k-1}\}$

Let us rewrite the above objective function by "adding and subtracting" the **conditional expectation function** (CEF) of Y given X: $E(Y|X)$

$$\begin{aligned}
\text{LF(MSE)} &= E[\,(\,Y-E(Y|X)+E(Y|X)-\tilde{Y}(X)\,)'\,(\,Y-E(Y|X)+E(Y|X)-\tilde{Y}(X)\,)\,] \\
&= E[\,(\,Y-E(Y|X)\,)'\,(\,Y-E(Y|X)\,)+(\,E(Y|X)-\tilde{Y}(X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,) \\
&\quad +(\,Y-E(Y|X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,)+(\,E(Y|X)-\tilde{Y}(X)\,)'\,(\,Y-E(Y|X)\,)\,] \\
&= \underbrace{E[\,(\,Y-E(Y|X)\,)'\,(\,Y-E(Y|X)\,)}_{(A)} \\
&\quad +\underbrace{E[\,(\,E(Y|X)-\tilde{Y}(X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,)\,]}_{(B)} \\
&\quad +2\ \underbrace{E[\,(\,Y-E(Y|X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,)\,]}_{(C)}
\end{aligned}$$

We want to choose $\tilde{Y}(X)$ to minimize the above loss function. Note that the value of (A) does not vary with our choice of $\tilde{Y}(X)$ as it is purely a function of Y and $E(Y|X)$ and not $\tilde{Y}(X)$. But, perhaps surprisingly, neither does (C) vary with our choice of $\tilde{Y}(X)$ ... despite the fact that $\tilde{Y}(X)$ does appear in (C):

$$\begin{aligned}
(C) &= E[\,(\,Y-E(Y|X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,)\,] \\
&= E_X[\,E_{Y|X}[\,(\,Y-E(Y|X)\,)'\,(\,E(Y|X)-\tilde{Y}(X)\,)\,]\,] \\
&= E_X[\,E_{Y|X}[\,Y'E(Y|X)-Y'\tilde{Y}(X)-E(Y|X)'E(Y|X)+E(Y|X)'\tilde{Y}(X)\,] \\
&= E_X[\,(\,E(Y|X)'E(Y|X)-E(Y|X)'\tilde{Y}(X)\,)-(\,E(Y|X)'E(Y|X)-E(Y|X)'\tilde{Y}(X)\,)\,]=0
\end{aligned}$$

We rely upon three "tricks" in the derivation above. First, we use the **Law of Iterated Expectation**: $E(\ \cdot\ ) = E_X[\,E_{Y|X}[\ \cdot\ ]\,]$. Second, we note that $E_{Y|X}[\,E(Y|X)\,] = E(Y|X)$ as the outer expectation is redundant. And third, we note that $E_{Y|X}[\,\tilde{Y}(X)\,] = \tilde{Y}(X)$ as $\tilde{Y}(X)$ is a *deterministic* function of X that we *choose*. Therefore, the MSE loss function can be minimized by minimizing (B). But note that (B) is simply the expected value of a sum of *squared* terms. Therefore, the minimum value that (B) can take is zero. This is achieved when $\tilde{Y}(X) = E(Y|X)$. This leaves us with the following result:

• **The best predictor (BP) of Y given X under the mean-squared error (MSE) loss function is the conditional expectation function (CEF) of Y given X — namely $E(Y|X)$**

Although we would like to use the best predictor, oftentimes we do not have enough information to calculate either the conditional expectation function $E[Y|X]$. This function requires us to know the conditional distribution of Y given X. However, if we know the mean of X and Y (μ_X, μ_Y respectively), the variance of X (Σ_{XX}) and the covariance of X and Y (Σ_{XY}) then we can calculate the **best *linear* predictor** (BLP) of Y given X under the MSE criterion.

The BLP is the best predictor of Y given X when we limit ourselves to the class of estimators that are *linear* in X: $\tilde{Y}(X) = \alpha + X\beta$

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \text{argmin}_{\alpha,\beta} \underbrace{E[\ (\ Y - (\alpha + X\beta)\ )'(\ Y - (\alpha + X\beta)\ )\ ]}_{W}$$

We can solve for α, β from the appropriate first order conditions, derived using vector calculus

$$\begin{aligned} \frac{\partial W}{\partial \alpha} &= 2\ [\ \alpha - E(Y) + E(X)\beta\ ] \\ &\Longrightarrow \alpha = \underbrace{E(Y)}_{\mu_Y} - \underbrace{E(X)}_{\mu_X}\beta \\ \frac{\partial W}{\partial \beta} &= 2\ [\ E(X'X)\beta - E(X'Y) + E(X)'\alpha\ ] \\ &\Longrightarrow \beta = [\ \underbrace{E(X'X) - E(X)'E(X)}_{\Sigma_{XX}}\ ]^{-1}\ [\ \underbrace{E(X'Y) - E(X)'E(Y)}_{\Sigma_{XY}}\ ] \end{aligned}$$

This leaves us with the following result:

• **The best linear predictor (BLP) of Y given X under the mean-squared loss (MSE) loss function is $Y = \alpha + X\beta$ where $\alpha = \mu_Y - \mu_X \frac{\Sigma_{XY}}{\Sigma_{XX}}$ and $\beta = (\Sigma_{XX})^{-1}\ \Sigma_{XY}$**

Now consider the OLS estimators (a, b) when we split off the intercept (constant) from the other variables in X. The $b^{ols} = (X'X)^{-1}X'Y$ formulation subsumes the intercept in X. What we write below only looks different as we change the definition of X. We now define X as

$$X = (\ X_1\ X_2\ \cdots\ X_{k-1}\ ) \quad \text{and} \quad X_i = \left(\ X_{1i}\ X_{2i}\ \cdots\ X_{(k-1)i}\ \right)$$

and

$$b^{ols} = \begin{pmatrix} b_1^{ols} \\ \vdots \\ b_k^{ols} \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

where a is a scalar and b is a $(k-1) \times 1$ column vector.

Re-consider the multivariate least squares problem

$$
\begin{aligned}
(a,b) &\equiv \text{argmin}_{\,a,b} \sum_{i=1}^{N} (\, Y_i - a - X_i\, b\,)^2 \\
&= \text{argmin}_{\,a,b}\, (Y - a\iota - Xb)'(Y - a\iota - Xb) \\
&\quad \text{(where } \iota \text{ is a Nx1 vector of ones and } a \text{ is a scalar)} \\
&= \text{argmin}_{\,a,b} \underbrace{[\, Y'Y - aY'\iota - Y'Xb - a\iota'Y + a^2\iota'\iota + a\iota'Xb - b'X'Y + ab'X'\iota + b'X'Xb \,]}_{W^*} \\
\frac{\partial W^*}{\partial a} &= -2\, Y'\iota + 2\, a\iota'\iota + 2\, \iota'Xb \\
&\Longrightarrow a = \bar{Y} - \bar{X}'b \\
&\quad \text{where } \bar{Y} = \frac{1}{N}\, Y'\iota \;\text{ and }\; \bar{X} = \frac{1}{N}\, X'\iota \\
\frac{\partial W^*}{\partial b} &= -X'Y + aX'\iota - X'Y + aX'\iota + 2X'Xb \\
&= -2\, [\, X'Y - aX'\iota - X'Xb \,] \\
&\Longrightarrow b = (\frac{X'X}{N} - \bar{X}'\bar{X})^{-1}\, (\frac{X'Y}{N} - \bar{X}\bar{Y})
\end{aligned}
$$

But observe that the best linear predictor (under MSE) parameters (α, β) are very similar to the OLS estimators (a, b), with the OLS estimators replacing the true *population* moments with their *sample* analogs.

$$E(Y) \Rightarrow \bar{Y} \quad E(X) \Rightarrow \bar{X} \quad E(X'X) \Rightarrow \frac{X'X}{N} \quad E(X'Y) \Rightarrow \frac{X'Y}{N}$$

Under some fairly general conditions, we can show that the random variables X and Y are ergodic. Technically, this indicates that the probability limit (convergence in probability) of the sample moments of X and Y are the population moments. More loosely speaking, this implies that given a large enough sample, the sample moments will very closely approximate the population moments. Under such conditions, we can further show that $a \xrightarrow{p} \alpha$ and $b \xrightarrow{p} \beta$

• Under some general conditions, the OLS estimators (a, b) are consistent estimators of the parameters (α, β) in the best linear predictor (BLP) of Y given X

A sufficient condition for ergodicity is for each observation (Y_i, X_i) to be drawn independently from the identical (same) probability distribution. A sample of such observations is referred as an *i.i.d.* random sample with *i.i.d.* standing for "**i**ndependently and **i**dentically **d**istributed."

3 Linearity Condition

We have shown that the OLS estimator is a moment-based estimator of the $BLP(Y|X)$ under the MSE criterion. This implies that *if* $E[Y|X]$ is *linear* in X then the OLS estimator is a moment-based estimator of the $BP(Y|X)$ (under MSE) as well.

Consider the following

$$\begin{aligned} Y &= X\beta + \epsilon \\ &\text{where } \epsilon \equiv Y - X\beta \end{aligned}$$

ϵ is a $(N \times 1)$ column vector of "residuals," the difference between Y and some linear function of X. Here X includes the constant.

Suppose that $E[Y|X] = X\beta$. Then

$$E[\epsilon|X] = E[Y|X] - E[X\beta|X] = X\beta - X\beta = 0$$

Furthermore, consider the OLS estimator b^{ols}

$$E[b^{ols}|X] = E[(X'X)^{-1}X'Y|X] = (X'X)^{-1}X'E[Y|X] = (X'X)^{-1}X'(X\beta) = \beta$$

Simply put, if $E[Y|X]$ is linear in X then the expected value (mean) of the OLS estimator is β, the actual coefficient values for $E[Y|X]$. In statistics jargon, b^{ols} is said to be an **unbiased** estimate of β (conditional on X) when $E[b^{ols}|X] = \beta$.

This result did not require any asymptotic theory. What it did require was for $E[Y|X] = X\beta$. $E[Y|X] = X\beta$ is referred to as the **linearity condition**.[1]

If the linearity condition is satisfied, we do not require asymptotic theory to justify our moment-based estimator. Even with finite sized samples (**"small samples"**), the OLS estimator can be rationalized as a moment-based estimator of the $BP(Y|X)$ under MSE.

Additionally, under the linearity condition, the estimators have further interpretation that lends itself to economic applications. Note that

$$\beta_j = \frac{\partial E[Y|X]}{\partial X_j} \qquad \text{for } j = 1 \ldots k$$

where β_j is the j^{th} element of β.

So, under the linearity condition, the OLS estimator b^{ols} provides unbiased estimates of the **marginal effect** of each X_j on the conditional mean of Y. Again, this result did not require either the Law of Large Numbers (LLN) or the Central Limit Theorem (CLT).

[1] Some textbook will refer to the linearity condition as $Y = X\beta + \epsilon$ where $E[\epsilon|X] = 0$. This is somewhat misleading as ϵ is just an artificial construct. The random variables that make up the population are simply X and Y. ϵ has meaning only as a strictly defined function of X and Y, $\epsilon \equiv Y - X\beta$

Gauss Markov Theorem

If the linearity condition is combined with two more conditions, we can show an even stronger result for the OLS estimator. Suppose the following three conditions are satisfied

1. Linearity: $E[Y|X] = X\beta$
2. Spherical Errors: $\text{Var}(Y|X) = \sigma^2 I$
   where σ^2 is a scalar and I the $(N \times N)$ diagonal matrix of ones
3. Linear Independence: $\text{rank}(X) = k$

These three conditions, together, are known as the **Gauss-Markov assumptions**. The Goldberger text refers to them as the assumptions of the neoclassical model (Chapter 25.2).[2]

Consider the second assumption, spherical errors. The assumption governs the pairwise relationship between observations in the sample. $\sigma^2 I$ can be expanded to

$$\sigma^2 \, I \quad = \quad \underbrace{\begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}}_{(N \times N) \text{ diagonal matrix of } \sigma^2}$$

All of the off-diagonal elements are zero for $\sigma^2 I$. The above implies the following:

- Conditional variance of Y_i is the same for each observation i: $\text{Var}(Y_i|X) = \sigma^2$ for $i = 1 \ldots N$
- Conditional covariance between Y_i and Y_j for $i \neq j$ is zero: $\text{Cov}(Y_i, Y_j|X) = 0$ for $i \neq j$

Note that the above implications are satisfied if the sample is random.

Consider the third assumption, linear independence. The assumption governs the relationship among the conditioning random variables, X. $\text{rank}(X) = k$ implies that no conditioning random variable X_j can be a linear function of the other conditioning random variances.

From a practical point of view, $\text{rank}(X) < k$ implies that $(X'X)^{-1}$ does not exist as $(X'X)$ is no longer **non-singular**.[3] And if $(X'X)^{-1}$ does not exist, then b^{ols} is not well-defined.

From a conceptual point of view, a $\text{rank}(X) < k$ implies that some of the conditioning random variables are redundant – the information about Y in those redundant random variables are also contained in some combination of the other conditioning random variables.

[2] The difference between the classical model (Chapter 15) and the neoclassical model is simply whether we treat X as random. From a practical perspective, the main difference between the models is that we use the marginal distribution $f(y)$ for classical and the conditional distribution $f(y|x)$ for neoclassical.

[3] $\text{rank}(X) > k$ is not possible as rank counts the number of linearly independent columns in X and there are only k columns in X

Suppose $X_3 = c_1X_1 + c_2X_2$ where (c_1, c_2) are some fixed constants. Then observing X_{3i} provides no additional information on Y_i after observing X_{1i} and X_{2i} as X_{3i} can be constructed from X_{1i} and X_{2i}. Additionally note that it will be difficult to distinguish β_3 from β_1 and β_2

$$\begin{aligned} Y &= X\beta + \epsilon \\ &= \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 \cdots + \beta_kX_k + \epsilon \\ &= \beta_1X_1 + \beta_2X_2 + \beta_3(c_1X_1 + c_2X_2) + \beta_4X_4 + \cdots + \beta_kX_k + \epsilon \\ &= (\beta_1 + c_1\beta_3)X_1 + (\beta_2 + c_2\beta_3)X_2 + \beta_4X_4 + \cdots + \beta_kX_k + \epsilon \end{aligned}$$

This difficulty in distinguishing among the individual effects is known as an **identification problem**. We will discuss this issue in more detail later.

If all three Gauss-Markov assumptions hold, we can prove that the OLS estimator b^{ols} is the **best linear unbiased estimator (BLUE)** of β, the coefficients of the linear conditional expectation function. Here, "best" refers to the estimator that yields the **minimum variance** among linear unbiased estimators.[4] And "linear" means "linear in Y"

> **Gauss Markov Theorem:** Given the Gauss Markov assumptions, the ordinary least squares estimator b^{ols} is the best (minimum variance) linear unbiased estimator of β

The proof for the Gauss-Markov theorem (adapted from Goldberger Ch. 15.4) is reproduced below:

- Any linear estimator of β, say b^*, can be expressed as $b^* = A^*Y$ where Y is the $(N \times 1)$ column vector from the sample and A^* some $k \times N$ consisting of constants and/or elements from X.
- If the linear estimator is also unbiased, then $E[b^*|X] = \beta$. This implies that $E[A^*Y|X] = A^*E[Y|X] = \beta$. From the Linearity Assumption, $E[Y|X] = X\beta$. So an unbiased linear b^* must satisfy $A^*X\beta = \beta$. Or, simply, $A^*X = I$ where I is the $(k \times k)$ identity matrix
- $\text{Var}(b^*|X) = \text{Var}(A^*Y|X) = A^*\text{Var}(Y|X)A^{*'}$. From Spherical Errors assumption, $\text{Var}(b^*|X) = A^*(\sigma^2 I)A^{*'} = \sigma^2 A^*A^{*'}$
- Use Goldberger's "A" matrix: $A = (X'X)^{-1}X'$. Note that $(X'X)^{-1}$ and hence A are well defined only if $\text{rank}(X) = k$. Without loss of generality, $A^* = A + D$ where $D \equiv A^* - A$.
- So, $A^*X = I$ implies $(A+D)X = (AX+DX) = I$ But $AX = (X'X)^{-1}X'X = I$. So $DX = 0$ (matrix of zeroes). This further implies that $DA' = 0$ and $AD' = 0$.
- $\text{Var}(b^*|X) = \sigma^2(A + D)(A + D)' = \sigma^2(AA' + AD' + DA' + DD') = \sigma^2(AA' + DD')$.
- Some linear algebra can be applied to show that DD' is minimized when $D = 0$. By construction (we defined D not A), AA' cannot be minimized.
- So $\text{Var}(b^*|X)$ for a linear unbiased b^* is minimized when D=0. This implies $A^* = A$. But $AY = (X'X)^{-1}X'Y = b^{ols}$.

[4] Instead of BLUE, we sometimes use the acronym MVLUE – Minimum Variance Linear Unbiased Estimator. One can show that among LUEs, the MVLUE is also the LUE that minimizes. mean squared error (MSE). So the use of "best" here is consistent with its earlier use