Classical Linear Regression Model & Ordinary Least Squares

Table of contents

1 Overview

In this handout we will revisit the Classical Linear Regression Model (CLRM) (see Wooldridge 2010, chaps. 1–2). The goal of this week's lecture is to:

1. understand the model specification;
2. it's underlying assumptions;

3. and the appropriate interpretation;
4. the OLS estimator, using linear algebra;
5. the geometry of OLS and partitioned regression result.

2 Model Specification

The linear population regression model is given by,

$$\begin{aligned} Y_i =& X_i'\beta + \varepsilon_i \\ =& \beta_1 \mathbf{1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + \varepsilon_i \end{aligned}$$

for $i = 1, 2, ..., n$. Where,

- i: unit of observation; e.g. individual, firm, union, political party, etc.
- $Y_i \in \mathbb{R}$: scalar random variable.
- $X_i \in \mathbb{R}^k$: k-dimensional (column[1]) vector of regressors, with $k < n$.[2]
- β: k-dimensional, non-random vector of unknown population parameters.
- ε_i: *unobserved*, random error term.[3]

The linear population regression equation is **linear in parameters**. This is an important assumption that does NOT restrict the model from being non-linear in regressors. For example, the equation

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i2}^2 + \varepsilon_i$$

non-linear in X_{i2}, but still linear in parameters. In contrast, the equation

$$Y_i = \beta_1 + \beta_2 X_{i2} + (\beta_2 \beta_3) X_{i3} + \varepsilon_i$$

is non-linear in parameters.

[1] My notation assumes that X_i is a column vector, which makes $X_i'\beta$ a scalar. Wooldridge (2010) uses the notation $X_i\beta$, implying that X_i is a row vector. This is a matter of preference.

[2] You might also refer to the vector of regressors or explanatory variables. The terms covariates or control variables are more common in Microeconometrics literature, where regressors are typically included for identification of causal effects. Some texts will use the term independent variables, but this name implies a specific relationship between Y and X that need not hold. Note, we will asssume in this term that $n > k$; i.e. this is "small" data.

[3] This is **NOT** the residual.

2.1 Intercept

The constant (intercept) in the equation serves an important purpose. While there is no *a priori* reason for the model to have a constant term, it does ensure that the error term is mean zero.

Proof. Suppose $E[\varepsilon_i] = \gamma$.

We can then define a new error term, $\upsilon_i = \varepsilon_i - \gamma$, such $E[\upsilon_i] = \gamma$. The population regression model can be rewritten as,

$$\begin{aligned} Y_i =& X_i'\beta + \upsilon_i + \gamma \\ =& \underbrace{(\beta_1 + \gamma)}_{\tilde{\beta}_1}\mathbf{1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + \upsilon_i \end{aligned}$$

The model has a new intercept $\tilde{\beta}_1 = \beta_1 + \gamma$, but the other parameters remain unchanged. □

2.2 Matrix notation

For a sample of n observations, we can stack the unit-level linear regression equation into a vector,

$$Y = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_n'\beta \end{bmatrix}}_{n\times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & ... & X_{1k} \\ X_{21} & X_{22} & & \\ \vdots & & \ddots & \\ X_{n1} & & & X_{nk} \end{bmatrix}}_{n\times k} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{k\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = X\beta + \varepsilon$$

Notice, in matrix notation, you lose the transpose from $X_i'\beta$. Apart from the absence of the i subscript, this is a useful way of knowing the dimension of the equation (in my notes). You MUST always write $X\beta$ and not βX. For the scalar case, $X_i'\beta = \beta' X_i$, but for the vector case βX is not defined since β is $k \times 1$ and X is $n \times k$.

3 CLRM Assumptions

Assumption CLRM 1. Population regression equation is linear in parameters:

$$Y = X\beta + \varepsilon$$

Assumption CLRM 2. Conditional mean independence of the error term:

$$E[\varepsilon|X] = 0$$

Assumption CLRM 2. is stronger than $E[\varepsilon_i|X_i]$ (mean independence for unit i). If all units were independent, then $E[\varepsilon_i|X_i]$ would imply $E[\varepsilon|X] = 0$. However, since we have not (yet) assumed this, we need this stronger exogeneity assumption. Consider, if i represented units of time (t), as in time-series models, independence across i will not hold.

Together, CLRM 1. and CLRM 2. imply that

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

This means that the Conditional Expectation Function is known and linear in parameters.

Conditional mean independence implies - by the Law of Iterated Expectations - mean independence of the error term,

$$E[\varepsilon|X] = 0 \Rightarrow E\big[E[\varepsilon|X]\big] = E[\varepsilon] = 0$$

and uncorrelatedness,

$$E[\varepsilon|X] = 0 \Rightarrow E[\varepsilon X] = 0$$

Note, neither of the above statements hold the other way around. Mean independence does not imply conditional mean independence and uncorrelatedness (zero correlation/covariance) does not imply conditional mean independence.

Uncorrelatedness rules out linear relationships between the regressors and error term while conditional mean independence rules out non-linear relationships too.

In general, distributional independence implies mean independence which then implies uncorrelatedness.

In the case joint-normally distributed random variables, uncorrelatedness implies independence. That is, if

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

Then $\sigma_{12} = \sigma_{21} = 0 \iff f_1 * f_2 = f_{12}$.

We will later show that uncorrelatedness is sufficient for consistency of the Ordinary Least Squares estimater, while conditional mean independence is required for unbiasedness of OLS.

Assumption CLRM 3. Homoskedasticity: $Var(\varepsilon|X) = E[\varepsilon\varepsilon'|X] = \sigma^2 I_n$

CLRM 3. states that the variance of the error term is independent of X and constant across units. The diagonal nature of the covariance matrix also implies that the error terms are uncorrelated across units in the data. Note, this does not imply independence of the error terms across units.

Models with heteroskedasticity relax the assumption of constant variance, allowing for a richer variance-covariance matrix that typically depends on X.

This assumption is unlikely to hold in time-series models where units represent repeated observations across time. Such violations are referred to as serial correlation or autocorrelation.

Even in cross-sectional data settings, you can have non-zero correlations across units in the data. A common instance of this is the case of clustering. Clustering can occur when units experience common/correlated 'shocks'; for example, the data contains groups of students from the same classroom who have a the same teacher. This can also be the result of clustered sampling, a common practice in multi-stage survey design.

Assumption CLRM 4. Full rank: $rank(X) = k \quad a.s.$ a.s.[4]

Since X is a random variable we should add to the assumption: $rank(X) = k$ *almost surely* (abbreviated a.s.). This means that the set of events in which X is not full rank occur with probability 0. The reason for this addition is that such a set of events (in the sample space) may not be empty.

CLRM 4. is some time referred to as the absence of perfect (or exact) collinearity. Do not confuse this with multicollinearity. Multicollinearity occurs when regressors are highly (linearly) correlated with one another, yielding imprecise estimates.

Assumption CLRM 5. Normality of the error term: $\varepsilon|X \sim N(0, \sigma^2 I_n)$

Assumption CLRM 6. Observations $\{(Y_i, X_i) : i = 1, ..., n\}$ are independently and identically distributed (iid).

CLRM 5 & 6 are not part of the Classical assumptions, but do simplify the problem of inference. Note, CLRM 5 implies independence across error terms, not implied by CLRM 3.

3.1 Non-random X

There is an alternative version of the CLRM in which X is a non-random, matrix of regressors/predictors. With X fixed, the error term is the only only random variable in the model. CLRM assumptions 1 and 4 remain the same, while CLRM 2, 3, 5, and 6 become:

Assumption CLRM 2$^{\mathbf{a}}$. Mean independence of the error term:

$$E[\varepsilon] = 0$$

Assumption CLRM 3$^{\mathbf{a}}$. Homoskedasticity: $Var(\varepsilon) = \sigma^2 I_n$

Assumption CLRM 5$^{\mathbf{a}}$. Normality of the error term: $\varepsilon \sim N(0, \sigma^2 I_n)$

Assumption CLRM 6$^{\mathbf{a}}$. Observations $\{\varepsilon_i : i = 1, ..., n\}$ are independently and identically distributed (iid).

[4] See extra material on Linear Algebra to read more on rank.

3.2 Identification

CLRM 1,2 and 4. are the *identifying* assumptions of the model. These assumptions allow us to write the parameter of interest as a set of 'observable' moments in the data. We can demonstrate this as follows.

Proof. Start with CLRM 2.

$$E[\varepsilon_i|X_i] = 0$$

Pre-multiply by the vector X_i,

$$X_i E[\varepsilon_i|X_i] = 0$$

Since the expectation is conditional on X_i, we can bring X_i inside the expectation function,

$$E[X_i\varepsilon_i|X_i] = 0$$

This conditional expectation is a random-function of X_i. If we take the expectation of this function w.r.t. X, we achieve the aforementioned result that conditional mean independence implies zero covariance,

$$E\left[E[X_i\varepsilon_i|X_i]\right] = E[X_i\varepsilon_i] = 0$$

Now substitute in for ε_i using the linear regression model from CLRM 1 and separate the resulting two terms,

$$\begin{aligned} E[X_i(Y_i - X_i'\beta)] &= 0 \\ \Rightarrow E[X_iX_i']\beta &= E[X_iY_i] \end{aligned}$$

Since β is a non-random vector, we can remove it from the expectation function.

Now we have a system of linear equations (of the form $Av = b$) with a unique solution if and only if the matrix $E[X_iX_i']$ is invertible. For the inverse of $E[X_iX_i']$ to exist, we require CLRM 4, since $rank(X) = k \quad a.s. \Rightarrow rank(E[X_iX_i']) = k$.[5]

$$\beta = E[X_iX_i']^{-1}E[X_iY_i]$$

□

[5]For n large, $rank(E[X_iX_i']) = k \Rightarrow rank(X) = k$. This follows from Law of Large Numbers, since $plim(n^{-1}X'X) = E[X_iX_i']$.

We cannot compute β because we do not know the joint distribution of (Y_i, X_i) needed to solve for the variance-covariance matrices. However, β is (point) identified because both Y and X are observed in the data and the parameters are "pinned down" by a unique set of 'observable' moments in the data.

β is not identified if the above system of linear equations does not have a unique solution. This will occur if two or more of the regressors are perfectly colinear.[6] β is also not be identified if the resulting expression for β includes 'objects' (moments, distribution/scale parameters) that are not 'observed' in the data. For example, if the error is not mean independent, the above expression will include a bias term that depends on $E[X_i'\varepsilon_i]$.

In this instance, the identification of β is scale dependent. That is, if we multiply Y_i by a scalar, β is multiplied by the same scalar. For example, in cases where a researcher is modelling standardized test-scores.

4 Interpretation

In this linear regression model each slope coefficient has a partial derivative interpretation,

$$\beta_j = \frac{\partial E[Y_i|X_i]}{\partial X_{ij}}$$

or, as a vector,

$$\beta = \frac{\partial E[Y_i|X_i]}{\partial X_i} = \begin{bmatrix} \frac{\partial E[Y_i|X_i]}{\partial X_{i1}} \\ \vdots \\ \frac{\partial E[Y_i|X_i]}{\partial X_{ik}} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Note, the derivative is expressed in terms of changes in the *expected* value of Y_i (conditional on X_i), not Y_i itself. This is because Y_i is a random variable, but under CLRM 1 & 2

$$E[Y_i|X_i] = X_i'\beta$$

For a given value of X_i, the above expression is non-random.

As β_j is a partial derivative, its interpretation is one that "holds fixed" the value of other regressors (i.e. *ceteris paribus*). Because of this, many researchers apply the experimental language of control variables when interpretting regression coefficients. However, this is dependent on the *assumed* linearity of the CEF.

[6] There are many such failures of (parametric) identification in models that include dummy variables (or fixed effects). Earlier we saw that the intercept is not separately identified from the mean of the error term. Mean independence of the error term, $E[\varepsilon_i] = 0$, is required for us to separately 'identify' β_1.

5 Ordinary Least Squares

OLS is *an* estimator for β. As will become evident in Lecture 3, it is not the only estimator for β.

The OLS estimator is the solution to,

$$\min_b \sum_{i=1}^{n} (Y_i - X_i' b)^2$$

Using vector notation, we can rewrite this as

$$\begin{aligned} &\min_b (Y - Xb)'(Y - Xb) \\ =&\min_b Y'Y - Y'Xb - b'X'Y + b'X'Xb \\ =&\min_b Y'Y - 2b'X'Y + b'X'Xb \end{aligned}$$

From line 2 to 3 we use the fact that $Y'Xb$ is a scalar and therefore symmetric: $Y'Xb = b'X'Y$.[7]

Differentiating the above expression w.r.t. the vector b and setting the first-order conditions to 0, we find that the following condition must hold for $\hat{\beta}$, the solution.

$$\begin{aligned} 0 &= -2X'Y + 2X'X\hat{\beta} \\ \Rightarrow X'X\hat{\beta} &= X'Y \end{aligned}$$

How did we get this result? Deriving the first order conditions requires knowledge of how to solve for the derivative of a scalar respect to a column vector (in this case $b \in R^k$). The extra material on Linear Algebra has some notes on vector differentiation.

[7] When working with vectors and matrices it is important to keep track of their size. You can only multiply two matrices/vectors if their column and row dimensions match. For example, if A and B are both $n \times k$ matrices ($n \neq k$), then AB is not defined since A has k columns and B n rows. For the same reason BA is also not defined. However, you can pre-multiply B with A' as A' is a $k \times n$ matrix: $A'B$ is therefore a $(k \times n) \cdot (n \times k) = k \times k$ matrix. Similarly, $B'A$ is defined, but is a $n \times n$ matrix.

Order matters when working with matrices and vectors. Pre-multiplication and post-multiplication are not the same thing.

Keep track of the size of each term to ensure they correspond to one another. In this instance, each term should be a scalar. For example, $-2b'X'Y$ is the multiplication of a scalar (-2: size 1×1), row vector (b': size $1 \times k$), matrix (X': size $k \times n$), and column vector (Y: size $n \times 1$). Thus we have a $(1 \times 1) \cdot (1 \times k) \cdot (k \times n) \cdot (n \times 1) = 1 \times 1$.

We can ignore the first term $Y'Y$ as it does not depend on b. The second term is $-2b'X'Y$. Here we can use the rule that,

$$\frac{\partial z'a}{\partial z} = \frac{\partial a'z}{\partial z} = a$$

In this instance, $a = X'Y \in R^k$. Thus,

$$\frac{\partial - 2b'X'Y}{\partial b} = -2\frac{\partial b'X'Y}{\partial b} = -2X'Y$$

The third term is $b'X'Xb$. This is what is commonly referred to as a quadratic form: $z'Az$. We know that the derivative of this form is,

$$\frac{\partial z'Az}{\partial z} = Az + A'z$$

and if A is symmetric, the result simplies to $2Az$. In this instance, $A = X'X$ is symmetric and the derivative is given by,

$$\frac{\partial b'X'Xb}{\partial b} = 2X'X$$

In order to solve for $\hat{\beta}$ we need to move the $X'X$ term to the right-hand side. If these were scalars we would simply divide both sides by the same constant. However, as $X'X$ is a matrix, division is not possible. Instead, we need to pre-multiply both sides by the inverse of $X'X$: $(X'X)^{-1}$. Here's the issue: the inverse of a matrix need not exist.

Given a *square* $k \times k$ matrix A, its inverse exists *if and only if* A is non-singular. For A to be non-singular its rank must have full rank: $r(A) = k$, the number of rows/columns. This means that all k columns/rows must be linearly independent. (See Material on Linear Algebra for a more detailed discussion of all these terms.)

In our application, $A = X'X$ and

$$r(X'X) = r(X) = colrank(X) \leq k$$

To insure that the inverse of $X'X$ exists, X must have full column rank: all column vectors must be *linearly independent*. In practice, this means that no regressor can be a *perfect* linear combination of others. However, we have this from

CLRM 4: $rank(X) = k$

You may know this assumption by another name: the absence of perfect colinearity between regressors.

The rank condition is the reason we exclude a base category when working with categorical variables.

Recall, most linear regression models are specified with a constant. Thus, the first column of X is

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

a $n \times 1$ vector vector of 1's, denoted here as ℓ. Suppose you have a categorical - for example, gender in an individual level dataset - that splits the same in two. The categories are assumed to be exhaustive and mutually exclusive. If you create two dummy variables, one for each category,

$$X_2 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{and} \qquad X_3 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

it is evident that $X_2 + X_3 = \ell$. (Here I have depicted the sample as sorted along these two categories.) If $X = [X_1 \; X_2 \; X_3]$, then it is rank-deficient: $r(X) = 2 < 3$, since $X_3 = X_1 - X_2$. Thus, we can only include two of these three regressors. We can even exclude the constant and have $X = [X_2 \; X_3]$.

If X is full rank, then $(X'X)^{-1}$ exists and,

$$\hat{\beta} = (X'X)^{-1}X'Y$$

This relatively simple expression is the solution to least squares minimization problem. Just think, it would take less than three lines of code to programme this. That is the power of knowing a little linear algebra.

We can write the same expression in terms of summations over unit-level observations,

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Y_i$$

Note, the change in position of the transpose: X_i is a column vector $\Rightarrow X_i'X_i$ is a scalar while $X_i X_i'$ is a $k \times k$ matrix. To match the first expression, the term inside the parenthesis must be a $k \times k$ matrix. Similarly, $X'Y$ is a $k \times 1$ vector, as is $X_i Y_i$.

5.1 Univariate case

Undergraduate textbooks all teach a very similar expression for the OLS estimator of a univariate regression model (with a constant); typically, something like,[8]

$$Y_i = \beta_1 + \beta_2 X_{i2} + \varepsilon_i$$

We know that the OLS estimators are give by,

$$\begin{aligned} \tilde{\beta}_2 =& \frac{\sum (Y_i - \bar{Y})(X_{i2} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{i2} - \bar{X}_2)^2} \\ \text{and} \qquad \tilde{\beta}_1 =& \bar{Y} - \tilde{\beta}_2 \bar{X}_2 \end{aligned}$$

I am deliberately using the notation $\tilde{\beta}$ to distinguish these two estimators from the expression below. Let us see if we can replicate this result, using vector notation. The the model is,

$$\begin{aligned} Y =& X\beta + \varepsilon \\ =& \begin{bmatrix} 1 & X_{12} \\ 1 & X_{22} \\ \vdots & \vdots \\ 1 & X_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon \\ =& \begin{bmatrix} \ell & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} =& (X'X)^{-1} X'Y \\ =& \left(\begin{bmatrix} \ell' \\ X_2' \end{bmatrix} \begin{bmatrix} \ell & X_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \ell' \\ X_2' \end{bmatrix} Y \\ =& \begin{bmatrix} \ell'\ell & \ell' X_2 \\ X_2'\ell & X_2' X_2 \end{bmatrix}^{-1} \begin{bmatrix} \ell' Y \\ X_2' Y \end{bmatrix} \end{aligned}$$

[8] Once you are familiar with vector notation, it is relatively easy to tell whether a model is uni- or multi-variate. This is because the notation $\beta_2 X_{i2}$ is not consistent with X_{2i} being a vector (row or column).

If X_{i2} is a $k \times 1$ vector then so is β_2. Thus, $\beta_2 X_{i2}$ is $(k \times 1) \cdot (k \times 1)$, which is not defined.

If X_{i2} is a row vector (as in Wooldridge, 2011), $\beta_2 X_{i2}$ will then be $(k \times 1) \cdot (1 \times k)$, a $k \times k$ matrix. This cannot be correct since the model is defined at the unit level.

Thus, if you see a model written with the parameter in front of the regressor, you know that this must be a single regressor. This is subtle, yet imporant, distinction that researchers often use to convey the structure of their model. Whenever X_{i2} is a vector, researchers will *almost always* use the notation $X_{i2}'\beta$ or $X_{i2}\beta$, depending on whether X_{i2} is assumed to be a column or row vector.

I went through this rather quickly, using a number of linear algebra rules that you may not be familiar with. Do not worry, the point of the exercise is not become a linear algebra master, but instead to focus on the element of each of each matrix/vector. Each element is a scalar (size 1×1).

If we right them each down as sums you they might be a little more familiar. First consider the 2×2 matrix:

- element [1,1]: $\ell'\ell = \sum_{i=1}^{n} 1 = n$
- element [1,2]: $\ell' X_2 = \sum_{i=1}^{n} X_{i2} = n\bar{X}_2$
- element [2,1]: $X_2'\ell = \sum_{i=1}^{n} X_{i2} = n\bar{X}_2$ (as above, since scalars are symmetric)
- element [2,2]: $X_2'X_2 = \sum_{i=1}^{n} X_{i2}^2$

Next, consider the final 2×1 vector,

- element [1,1]: $\ell' Y = \sum_{i=1}^{n} Y_i = n\bar{Y}$
- element [2,1]: $X_2'Y = \sum_{i=1}^{n} Y_i X_{i2}$

Our OLS estimator is therefore,

$$\hat{\beta} = \begin{bmatrix} n & n\bar{X}_2 \\ n\bar{X}_2 & \sum_{i=1}^{n} X_{i2}^2 \end{bmatrix}^{-1} \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^{n} Y_i X_{i2} \end{bmatrix}$$

We now need to solve for the inverse of the 2×2 matrix. You can easily find notes on how to do this online. Here, I will just provide the solution.

$$\hat{\beta} = \frac{1}{n\sum_{i=1}^{n} X_{i2}^2 - n^2\bar{X}_2^2} \begin{bmatrix} \sum_{i=1}^{n} X_{i2}^2 & -n\bar{X}_2 \\ -n\bar{X}_2 & n \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^{n} Y_i X_{i2} \end{bmatrix}$$

Remember, this is still a 2×1 vector. We can now solve for the final solution:

$$
\begin{aligned}
\hat{\beta} =& \frac{1}{n\sum_{i=1}^{n} X_{i2}^2 - n^2\bar{X}_2^2} \begin{bmatrix} n\bar{Y}\sum_{i=1}^{n} X_{i2}^2 - n\bar{X}_2\sum_{i=1}^{n} Y_i X_{i2} \\ n\sum_{i=1}^{n} Y_i X_{i2} - n^2\bar{X}_2\bar{Y} \end{bmatrix} \\
=& \frac{1}{n\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2} \begin{bmatrix} n\bar{Y}\sum_{i=1}^{n} X_{i2}^2 + n^2\bar{Y}\bar{X}^2 - n^2\bar{Y}\bar{X}^2 - n\bar{X}_2\sum_{i=1}^{n} Y_i X_{i2} \\ n\sum_{i=1}^{n}(Y_i-\bar{Y})(X_{i2}-\bar{X}_2) \end{bmatrix} \\
=& \frac{1}{n\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2} \begin{bmatrix} n\bar{Y}\sum_{i=1}^{n}(X_{i2}-\bar{X})^2 - n\bar{X}_2\sum_{i=1}^{n}(Y_i-\bar{Y})(X_{i2}-\bar{X}_2) \\ n\sum_{i=1}^{n}(Y_i-\bar{Y})(X_{i2}-\bar{X}_2) \end{bmatrix} \\
=& \begin{bmatrix} \bar{Y} - \frac{n\sum_{i=1}^{n}(Y_i-\bar{Y})(X_{i2}-\bar{X}_2)}{n\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2}\bar{X}_2 \\ \frac{n\sum_{i=1}^{n}(Y_i-\bar{Y})(X_{i2}-\bar{X}_2)}{n\sum_{i=1}^{n}(X_{i2}-\bar{X}_2)^2} \end{bmatrix} \\
=& \begin{bmatrix} \bar{Y} - \tilde{\beta}_2\bar{X}_2 \\ \tilde{\beta}_2 \end{bmatrix} \\
=& \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix}
\end{aligned}
$$

The math is a little involved, but it shows you these solutions are are the same. Unfortunately, the working gets even more arduous in a multivariate context. However, there are useful tools to help us with that we will discuss next.

5.2 Geometry of OLS

In the last section we saw how the OLS estimator can, more generally, be described as a linear transformation of the Y vector.

$$\hat{\beta} = (X'X)^{-1}X'Y$$

We also saw that in order for there to be a (unique) solution to the least squared problem, the X matrix must be full rank. This rules out any perfect colinearity between columns (i.e. regressors) in the X matrix, including the constant.

Given the vector of OLS coefficients, we can also estimate the residual,

$$
\begin{aligned}
\hat{\varepsilon} =& Y - X\hat{\beta} \\
=& Y - X(X'X)^{-1}X'Y \\
=& (I_n - X(X'X)X^{-1})Y
\end{aligned}
$$

by plugging the definition of $\hat{\beta}$. Thus, the OLS estimator separates the vector Y into two components:

$$\begin{aligned} Y =& X\hat{\beta} + \hat{\varepsilon} \\ =& \underbrace{X(X'X)^{-1}X'}_{P_X} Y + (\underbrace{I_n - X(X'X)X^{-1}}_{I_n - P_X = M_X})Y \\ =& P_X Y + M_X Y \end{aligned}$$

The matrix $P_X = X(X'X)^{-1}X'$ is a $n \times n$ *projection* matrix. It is a linear transformation that projects any vector into the span of X: $S(X) \subset \mathbb{R}^n$. (See for more information on these terms.) $S(X)$ is the vector space spanned by the columns of X. The dimensions of this vector space depends on the rank of P_X,

$$dim(S(X)) = r(P_X) = r(X) = k$$

The matrix $M_X = I_n - X(X'X)^{-1}X'$ is also a $n \times n$ projection matrix. It projects any vector into X's *orthogonal* span: $S^{\perp}(X)$. Any vector $z \in S^{\perp}(X)$ is orthogonal to X. This includes the estimated residual, which is by definition orthogonal to the predicted values and, indeed, any column of X (i.e. any regressor). The dimension of this orthogonal vector space depends on the rank of M_X,

$$dim(S^{\perp}(X)) = r(M_X) = r(I_n) - r(X) = n - k$$

The orthogonality of these two projections can be easily shown, since projection matrices are idempotent ($P_X P_X = P_X$) and symmetric ($P'_X = P_X$). Consider the inner product of these two projections,

$$P'_X M_X = P_X(I_n - P_X) = P_X - P_X P_X = P_X - P_X = 0$$

The least squares estimator is a projection of Y into two vector spaces: one the span of the columns of X and the other a space orthogonal to X.

Why is this useful? Well, it helps us understand the "mechanics" (technically geometry) of OLS. When working with linear regression models, we typically assume either strict exogeneity - $E[\varepsilon|X] = 0$ - or uncorrelatedness - $E[X'\varepsilon] = 0$ - where the former implies the latter (but not the other way around).

When we use OLS, we estimate the vector $\hat{\beta}$ such that,

$$X'(Y - X\hat{\beta}) = X'\hat{\varepsilon} = 0 \quad always$$

This is true, *not just in expectation*, but by definition. The relationship is "mechanical": the regressors and estimated residual are perfectly uncorrelated. This can be easily shown:

$$\begin{aligned} X'\hat{\varepsilon} =& X'M_X Y \\ =& X'(I_n - P_X)Y \\ =& X'I_n Y - X'X(X'X)^{-1}X'Y \\ =& X'Y - X'Y \\ =& 0 \end{aligned}$$

You are essentially imposing the assumption of uncorrelatedness between the explained and unexplained components of Y on the data. This means that if the assumption is wrong, so is the projection.

5.3 Partitioned regression

The tools of linear algebra can help us better understand partitioned regression. Indeed, I would go as far to to say that it is quite difficult to understand partitioned regression without an understanding of projection matrices. Moreover, we need to understand partitioned regression to really understand multivariate regression. The partitioned regression result is referred to as Frisch-Waugh-Lovell Theorem (FWL).

Theorem 5.1. *FWL says that if you have two sets of regressors,* $[X_1, X_2]$*, then* $\hat{\beta}_1$*, the OLS estimator for* β_1*, from the regression,*

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

is also given by the regression,

$$M_2 Y = M_2 X_1 \beta_1 + \xi$$

We can demonstrate the FWL theorem result using projection matrices. Two simplify matters, we will divide the set of regressors into two groups: X_1 a single regressor and X_2 a $n \times (k-1)$ matrix. We can rewrite the linear model as,

$$Y = X\beta + \varepsilon = \beta_1 X_1 + X_2\beta_2 + \varepsilon$$

Let us begin by applying our existing knowledge. From above, we know that the residual from the regression of X_1 on X_2 is,

$$\hat{v} = M_2 X_1$$

where $M_2 = I_n - X_2(X_2'X_2)^{-1}X_2'$. It turns out, it does not matter if we residualize Y too. **Can you see why?** Thus, the model we estimate in the second step, is

$$Y = \gamma_1 \underbrace{M_2 X_1}_{\hat{v}} + \xi$$

We know that $\hat{\gamma}_1 = (\hat{v}'\hat{v})^{-1}\hat{v}'Y$. Replacing the value of the residual, we get,

$$\begin{aligned}\hat{\gamma}_1 =&(\hat{v}'\hat{v})^{-1}\hat{v}'Y \\ =&(X_1'M_2M_2X_1)^{-1}X_1'M_2Y \\ =&(X_1'M_2X_1)^{-1}X_1'M_2Y\end{aligned}$$

Here, we use both the symmetric and idempotent qualities of M_2. Next we want to show that $\hat{\beta}_1$ is given by the same value. This part is more complicated. Let's start with by reminding ourselves of the following:

$$\begin{aligned} X'X\hat{\beta} &= X'Y \\ \begin{bmatrix} X_1 & X_2 \end{bmatrix}' \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \begin{bmatrix} X_1 & X_2 \end{bmatrix}' Y \\ \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix} \end{aligned}$$

We could solve for $\hat{\beta}_1$ by solving for the inverse of $X'X$; however, this will take a long time. An easier approach is to simply verify that $\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$. Recall, $\hat{\beta}$ splits Y into two components:

$$Y = \hat{\beta}_1 X_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$

If we plug this definition of Y into the above expression we get,

$$\begin{aligned}&(X_1'M_2X_1)^{-1}X_1'M_2(\hat{\beta}_1X_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}) \\ =&\hat{\beta}_1 \underbrace{(X_1'M_2X_1)^{-1}X_1'M_2X_1}_{=I_n} \\ &+ \underbrace{(X_1'M_2X_1)^{-1}X_1'M_2X_2\hat{\beta}_2}_{=0} \\ &+ \underbrace{(X_1'M_2X_1)^{-1}X_1'M_2\hat{\varepsilon}}_{=0} \\ =&\hat{\beta}_1\end{aligned}$$

In line 2, I use the fact that $\hat{\beta}_1$ is a scalar and can be moved to the front (since the order of multiplication does not matter with a scalar). In line 3, I use the fact that $M_2 X_2 = 0$ by definition. Line 4 uses the fact that $M_2\hat{\varepsilon} = \hat{\varepsilon}$ which means that $X_1' M_2 \hat{\varepsilon} = X_1'\hat{\varepsilon} = 0$.

The OLS estimator solves for β_1 using the variance in X_1 that is orthogonal to X_2. This is the manner in which we "hold X_2 constant": the variation in $M_2 X_1$ is orthogonal to X_2. Changes in $M_2 X_1$ are *uncorrelated* with changes in X_2; *as if* the variation in $M_2 X_1$ arose independently of X_2. However, uncorrelatedness does NOT imply independence.

References

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT press.