EC338 Pre-reading

Neil Lloyd

2023-10-03

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This e-book provides some pre-reading material for the *EC338: Econometrics 2 - Microeconometrics* module. In line with the Applied Microeconomics and Microeconometrics literatures, EC338 has evolved over time to focus on causal inference, the identification strategies that underpin various casual estimands, and their corresponding estimators.

As a result of this shift in focus, less time is given to more advanced material relating to the linear estimators used in this literature. As the relevant material is covered in *EC226: Econometrics 1*, you should be able to follow EC338 with a little revision and study. However, the change in notation may 'trip you up' and certain topics will be easier to understand with a richer understanding of linear estimators. For this reason, I have prepared this material as a brief (re-)introduction to ordinary least squares (OLS).

In EC338 we work extensively with dummy variables. As you know from EC226, estimating a linear regression model with discrete independent variables is relatively simple and the interpretation of the coefficient is *typically* straight-forward. However, the use of dummy variables requires a careful consideration of (perfect) collinearity and to understand collinearity (or rank conditions) it helps to think of data as a matrix, or system of column vectors. This becomes even more important when we start to consider models with various dimensions of fixed-effects.

These notes begin by revisiting the basic linear regression model and OLS estimator using vector notation. Next we revisit the properties of OLS, using this same notation, but without extensive proofs. Proofs should be available from a range of textbooks, including Wooldridge (2011). The material that will be least familiar to you will be the geometry of OLS. Here we will cover projection matrices and how they can be used to understand partitioned regression. Finally, we discuss dummy variables, their projections, and issues of colinearity.

Remember, this is still the same material covered in EC226, just using vector notation. In many instances this simplifies the notation, as summations over n and/or t can be replaced with a simple inner product of vectors. For example, consider the average of random variable Y_i,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

If we define two $n \times 1$ column vectors,

$$\ell = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \text{and} \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

then the inner produce of these vectors is,

$$\langle \ell, Y \rangle = \ell' Y = 1 \cdot Y_1 + 1 \cdot Y_2 + \cdots + 1 \cdot Y_n = \sum_{i=1}^{n} Y_i$$

In addition,

$$\langle \ell, \ell \rangle = \ell' \ell = 1 \cdot 1 + 1 \cdot 1 + \cdots + 1 \cdot 1 = \sum_{i=1}^{n} 1 = n$$

Thus, the average can be expressed as,

$$(\ell' \ell)^{-1} \ell' Y = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

It turns out, this linear transformation of Y,

$$AY = (\ell' \ell)^{-1} \ell' Y$$

has a structure that is common to many linear estimators; a structure that derives from the projection matrix of the ℓ-vector: $P_\ell = \ell(\ell'\ell)^{-1}\ell'$. We will discuss this in Chapter Chapter 4 when we explore the geometry of OLS.

At the end of this e-book is a compendium on linear algebra basics. I may reference these during the e-book. The compendium is not exhaustive and does not include proofs. Please consult a linear algebra text book for further reading if you require.

The material in this e-book will not be examined directly, but will assist with your understanding of the material covered during term.

2 The Linear Regression Model

2.1 Vector notation

Most undergraduate textbooks discuss data in terms of random variables: the dependent or outcome variable (typically denoted Y_i) and various independent or explanatory variables $(X_{i1}, X_{i2}, \dots, X_{ik})$. There's nothing wrong with this language, but to understand the geometry of OLS we will need to think in terms of random *vectors*.

When working with cross-sectional data, we think of a random sample as a collection of n realizations of the same random variable. Each observation represents a different unit (e.g., individual, firm, classroom, etc.) and we typically add the assumption that the data is *i.i.d.* (independently and identically distributed) across units of observation. It makes no difference then if we think of this random sample as a collection of $n \times 1$ random vectors, where each *row* represents a different unit and the unit of observation is maintained across random vector.

For each unit you observe the outcome (Y_i) and a vector of explanatory variables (or regressors):

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{bmatrix}$$

Take note of the ordering of the subscripts: the first denotes the unit of observation (i) and the second the number of the regressor $(j = 1 \dots k)$. The pair (Y_i, X_i) represents an *observation*, where Y_i is a single random variable and X_i a random (column) vector.[1] A collection of observations forms a sample.

You are, no doubt, familiar with the linear regression model. A simple univariate model is typically written as,

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

[1] In these notes, as in the remainder of EC338, I treat X_i as a column vector. Some texts, including Wooldridge (2011), will treat X_i as a row vector. This distinction is not significant, but will affect your notation. I will point this out at a later stage.

Where β_1 is the constant (intercept) and β_2 the slope coefficient. In EC338, we will discuss in more detail the justification for this model. For now, let us focus on notation.

The linear regression model is linear *in parameters*, which means we can express the outcome as a linear transformation of a finite set of parameters (i.e. $k \times 1$ vector of parameters). These (population) parameters are assumed to be constants and unknown to the econometrician.

We can rewrite the above equation using vectors,

$$\begin{aligned} Y_i &= \begin{bmatrix} 1 \; X_{i2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon_i \\ &= \begin{bmatrix} 1 \\ X_{i2} \end{bmatrix}' \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon_i \\ &= X_i'\beta + \varepsilon_i \end{aligned}$$

Where X_i is a column vector including the number 1 in the first row (for the constant/intercept) and X_{i2} in the second row.

> **[important]** You may find my notation slightly unusual. A lot undergraduate textbooks use different letters (α, β, γ, etc.) to denote the constant and slope coefficients. As we are collecting all coefficients in a single vector, it helps to use indexes instead of different letters.
>
> Then there is the issue of the where to start the index: 0 or 1. This decision is somewhat arbitrary and hinges on whether the the k regressors included in the model includes the constant term. At the top of the page I described an observation as an outcome and a random vector of regressors. As the constant is not *random* it is natural to think of it apart from the set of regressors included in the model. However, this decision is somewhat arbitrary. You could simply set $X_{i1} = 1 \; \forall \; i$ and the constant would be included in k.
>
> **In EC226, you indexed from 0.** Recall, the linear model as $k + 1$ parameters; the $+1$ for the constant. When computing the degrees of freedom for the RSS, you solved for $n - k - 1$: k regressors plus the constant.
>
> **Here we will index from** 1**.** This distinction makes it easier to keep track of the size of the matrix. It is also a more natural notation if you consider that the model need not have a constant. The choice of including a constant is therefore no different to including any other regressor. Moreover, when we consider models with fixed effects the constant typically drops from the model.
>
> The key thing to remember is that you need to keep track of the number of parameters in the model, that includes the constant *if there is one*.

We can easily extend this notation to the case of multivariate regression. For example, consider a model with a constant and $k-1$ regressors.

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

Regardless of the number of regressors, the notation remains the same. Take note of the fact that the X_i is a column vector which means that the notation must include a transpose: $X_i'\beta$. Excluding the transpose is incorrect since you cannot multiple two $k \times 1$ column vectors. You can multiply a $1 \times k$ row vector with a $k \times 1$ column vector giving you a 1×1 scalar. The result should be a scalar since Y_i is a scalar.

This notation is not universal. For example, Wooldridge (2011) treats X_i as a row vector. For this reason, the linear regression model can be expressed as $X_i\beta$. Both notations are used in the applied literature, but I am more familiar with and prefer the column-vector notation.

2.2 Matrix notation

The above expressions for the linear regression model all describe a single unit of observation from the sample. Consider, each line included the subscript i. Since the model is assumed to be the same for each observation, this an accurate depiction of the linear regression model. However, we also need to think about the correct notation for the entire sample. To do so, we will have to worker with both vectors and matrices.

Since the model is the same for each observation in the sample, we could imagine “stacking” all n observations on top of one another to form a vector. Consider first the outcome variable,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$

$$

Y is a $n \times 1$ column vector of all outcomes in the sample. You can distinguish the vector Y from the scalar Y_i by the absence of a subscript.

Similarly, we can stack the right-hand side of the equation.

$$\begin{aligned} Y &= \begin{bmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ X_n'\beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= X\beta + \varepsilon \end{aligned}$$

Like Y, ε is a $n \times 1$ vector. X is a $n \times k$ matrix and β remains a $k \times 1$ vector of parameters. The product of a $n \times k$ matrix and $k \times 1$ vector is a $n \times 1$ vector: the same size vector as Y and ε. As it is important to understand the structure of X, let us write it out in detail.

$$X = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & & \\ \vdots & & \ddots & \\ X_{n1} & & & X_{nk} \end{bmatrix}$$

The X matrix has n rows, each representing a different unit of observation, and k columns, each representing a different regressor. Recall, one of these regressors *may* be a constant. If the model includes a constant then $X_{i1} = 1 \ \forall \ i$. This means that the first column of X is a vector of 1's. Each subsequent column represents another regressor.

If you are familiar with rectangular datasets from with working in STATA or R, then you may have notices that X is essentially the "dataset" (excluding the outcome variable). In a rectangular dataset, each row represents a different observation and each column a different variable. That's what we have here.

Why is this noteworthy? When assert that there must be an absence of perfect colinearity between the variables in the model, we are actually saying that the columns of X must be linearly independent. The formal way of expressing this is that X must have *full* column rank; or $r(X) = k$ (see Chapter 6 for a definition of linear dependence and rank). This is why the OLS condition included in EC226 as "no perfect colinearity" is sometimes referred to as the rank condition. Without full rank, we cannot estimate the linear regression model.

3 Ordinary Least Squares

You are, no doubt, familiar with the ordinary least squares (OLS) estimator from your previous studies in Econometrics (EC226 or otherwise). OLS is *an* estimator for β, it is not the only one. Indeed, you could use maximum likelihood methods to estimate 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 face that $Y'Xb$ is a scalar and therefore symmetric: $Y'Xb = b'X'Y$.

> [**note**] 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 $(1 \times 1) \cdot (1 \times k) \cdot (k \times n) \cdot (n \times 1) = 1 \times 1$.

Differentiating the above expression w.r.t. the vecotr 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 $b \in R^k$). Chapter 6 has some notes on vector differentiation.

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,

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

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

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

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,

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

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

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

What are the dimensions of this equation? Well, $X'X$ is a $k \times k$ matrix and $\hat{\beta}$ a $k \times 1$ vector. Thus the left-hand side of the equation is a $k \times 1$ vector, as is the right-hand side.

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 Chapter 6 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. This gives rise to one of the key linear regression model assumptions:

> [**assumption**] *rank condition:* $r(X) = k$

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

> [**comment**] The rank condition is the reason we exclude a base category when working with categorical variables. We will revisit this subject in more detail in Chapter 5.
>
> Recall, most linear regression models are specified with 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 $X_3 = X_1 - X2$. Thus, we can only include two of these three regressors. We can even exclude the constant and have $X = [X_2\ X_3]$. In Chapter 4 we will show

> that the projection remains the same regardless of which category we exclude, even the constant.

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.

> [**comment**] You may recognise the above expression from the Chapter 1 where we used vectors to define the mean estimator. It turns out, the mean estimator is the simplest of OLS estimators. It is a regression of Y against a constant alone: $X = \ell$.

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

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

Note, the change in position of the transpose: X_i is a column vector $\Rightarrow X_i'X_i$ is a scalar and X_iX_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_iY_i.

3.1 The uni-variate case

Undergraduate textbooks all teach a very similar expression for the OLS estimator of a uni-variate regression model (with a constant), such as

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

> [**note**] 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.

We know that,

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

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. When written in vector notation, 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,

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

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. (Take a look back at Chapter 1 to remind yourself of some of these steps). 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,

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

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.

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

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.

4 The 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:

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

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: $S(X)$. (See Chapter 6 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 work 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 covariates 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.

4.1 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.

Let us 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,

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

Partitioned regression is typically taught as follows. The OLS estimator for β_1 can achieved by first regressing X_1 on X_2,

$$X_1 = X_2\gamma_2 + \upsilon_1$$

Next, you regress Y on the residual from the above model,

$$Y = \gamma_1\hat{\upsilon}_1 + \xi$$

The partitioned regression result states that the OLS estimator for $\hat{\gamma}_1 = \hat{\beta}_1$. This aids in our understanding of β_1 as the partial effect of X_1 on Y, holding X_2 constant.

We can show this using projection matrices. 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,

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

where $M_2 = I_n - X_2(X_2'X_2)^{-1}X_2'$. Thus, the model we estimate in the second step, is

$$Y = \gamma_1 M_2 X_1 + \xi$$

We know that $\hat{\gamma}_1 = (\hat{\upsilon}'\hat{\upsilon})^{-1}\hat{\upsilon}'Y$. Replacing the value of the residual, we get> [note] We use both the symmetry and idempotent quality of M_2.

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

Next we want to show that $\widehat{\beta}_1$ is given by the same value. This part is more complicated. Let's start with by reminding ourselves of the following:

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

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

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

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

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

In line 2, I use the fact that $\widehat{\beta}_1$ 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 $M_2X_2 = 0$ by definition. Line 4 uses the fact that $M_2\widehat{\varepsilon} = \widehat{\varepsilon}$ which means that $X_1'M_2\widehat{\varepsilon} = X_1'\widehat{\varepsilon} = 0$.

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_2X_1 is orthogonal to X_2. Changes in M_2X_1 are *uncorrelated* with changes in X_2; *as if* the variation in M_2X_1 arose independently of X_2. However, uncorrelatedness does NOT imply independence.

5 Dummy Variables

Dummy variables are used extensively in Microeconometrics. They are used to model discrete public policy and experimental ‘treatments’ in Microeconomics and are the building blocks of “fixed effects” used widely in Microeconomics models.

Any dummy variable, $D_i = \mathbf{1}\{true\}$, splits the sample into two groups: true (e.g. treated) and false (e.g. control/untreated). In a basic setting the use a of dummy variable might be relatively straight forward. For example, consider a linear model using to assess a single treatment from a randomized control trial,

$$Y_i = \beta_{10} + \beta_{11} D_{1i} + \varepsilon_i$$

where $D_{1i} = \mathbf{1}\{treated\}$ identifies those who are treated in the sample.

For ease of demonstration, let us assume that the data is sorted on D_i: the first N_0 observations are the untreated control group, and the next $N_1 = N - N_0$ the treated. From previous chapters, we know that the matrix of regressors in this model is given by,

$$X_1 = [\ell, D_1] = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

The above matrix has rank 2, as the columns are linearly independent. Likewise, the rank of the 2 × 2 matrix $X_1'X_1$ is also 2 (i.e. full rank) and is therefore invertible. If $N = N_0$ - i.e., no units were treated - the matrix would be rank deficient. Similarly, if $N_0 = 0$ - i.e., all units were treated - D_1 would be a column of 1’s, colinear with the constant.

Consider then the regressor $D_{2i} = 1 - D_{1i} = \mathbf{1}\{control\}$. What would happen if we included D_{2i} in the above model? The matrix of regressors would be,

$$X = [\ell, D_1, D_2] = \begin{bmatrix} 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{bmatrix}$$

The rank of matrix X is not 3; it remains 2. This because the three columns of X are linearly dependent:

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

You can have ANY two of the columns in the model, but not all three. Moreover, regardless which two you choose, the projection remains the same. Consider, the three possible models,

$$\begin{aligned} Y_i =& \beta_{10} + \beta_{11} D_{1i} + \varepsilon_i = X_1 \beta_1 + \varepsilon_i && \text{where} \quad X_1 = [\ell \; D_1] \\ Y_i =& \beta_{20} + \beta_{22} D_{2i} + \varepsilon_i = X_2 \beta_2 + \varepsilon_i && \text{where} \quad X_2 = [\ell \; D_2] \\ Y_i =& \beta_{01} D_{1i} + \beta_{02} D_{2i} + \varepsilon_i = X_0 \beta_0 + \varepsilon_i && \text{where} \quad X_0 = [D_1 \; D_2] \end{aligned}$$

It turns out that,

$$P_1 = X_1 (X_1' X_1)^{-1} X_1' = P_2 = P_0$$

Similarly,

$$M_1 = I_n - X_1 (X_1' X_1)^{-1} X_1' = M_2 = M_0$$

Thus, the vector of residuals estimated by OLS for the above three models are all the same.

Why does this matter? Consider, the case where you include a categorical covariate ('control' variable) in your model. We know from EC226 that a categorical variable must be included in the model as a set of dummy variables: one for each category.

$$Y_i = \alpha + \beta Z_i + \sum_{j=2}^{J} \gamma_j D_{ji} + \upsilon_i$$

We must exclude 1 of the J categories since the model contains a constant and all the dummy variables (for the J categories) add up to 1 (or the vector ℓ). Here, I exclude the first category, $j = 1$.

We know from the previous discussion on partitioned regression that we can estimate the scalar parameter β by first regressing Z_i on all other regressors,

$$Z_i = \delta + \sum_{j=2}^{J} \psi_j D_{ji} + \nu_i$$

Next we regress the outcome, Y_i, on the residuals from the above equation. Using vector notation, this model is given by,

$$Y_i = \beta \hat{\nu_i} + \zeta_i = \beta M_{D_1} Z_i + \zeta_i$$

where M_{D_1} is the orthogonal projection matrix from the regression of Z_i on the full set of dummy variables (excluding category 1) and constant. The OLS estimator for β is given by,

$$\hat{\beta} = (Z' M_{D_1} Z)^{-1} Z' M_{D_1} Y$$

There are two things to take from this equation. First, we get the same result whether or not we first orthogonalize Y_i on the full set of covariates, since M_D is idempotent: $M_{D_1} = M_{D_1} M_{D_1}$. Thus,

$$(Z' M_{D_1} Z)^{-1} Z' M_{D_1} M_{D_1} Y = (Z' M_{D_1} Z)^{-1} Z' M_{D_1} Y = \hat{\beta}$$

Second, the choice of base category among the control variables is irrelevant. Indeed, even if we dropped the constant and included a dummy variable for all J categories, we would achieve the same estimator. This is because,

$$M_{D_0} = M_{D_1} = M_{D_2} ... = M_{D_J}$$

where M_{D_0} is the orthogonal projection matrix from a regression with all J categories (and no constant) and M_{D_j} is the orthogonal projection matrix from a regression with the j-th category excluded (and a constant).

5.1 Fixed Effects

One way of interpreting a model with an (exhaustive) set of dummy variables is as a model with a group-specific constant. Since, it makes no difference to parameter of interest whether you include the constant and $J-1$ dummies or exclude the constant and include all J dummy variables, we can consider the model,[1]

$$Y_i = \beta Z_i + \sum_{j=1}^{J} \gamma_j D_{ji} + \upsilon_i$$

A common short-hand notation for such a set-up is,

$$Y_i = \beta Z_i + \gamma_j + \upsilon_i$$

where γ_j signifies the presence of J group fixed effects. For each group the constant, denoted by γ, changes. However, it is important to remember that this notation is a short-hand. The full set of regressors includes J dummy variables (or a constant and $J-1$ dummy variables). When applying this shorthand it is standard to drop the constant term.

5.2 Multi-level Fixed Effects

Suppose you had two categorical variables, where one was a proper subset of the other. For example, an individual level dataset that contained information on the country in the UK where an individual lived (i.e. England, Scotland, Wales, Northern Ireland) and their county (i.e. Warwickshire, Oxfordshire, Cambridgeshire, etc.). Since county borders do not overlap country borders in the UK, an individuals county perfectly predicts their country.

Consider then the model,

$$Y_i = \beta Z_i + \gamma_j + \delta_k + \upsilon_i$$

where j represents country fixed effects and k county fixed effects. Let's assume the data is sorted by country and then county, and that there are two countries, each with 2 counties. (An obvious simplification.) If we consider the matrix of fixed effects, they can be written as,

[1] The choice of base category makes a difference when the variable of interest is a categorical variable. In this instance, the choice of base category changes the interpretation of the coefficient. This discussion relates more to instances where the categorical variable is included as a control variable.

$$\begin{bmatrix}1 & 0 & 1 & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \vdots & \vdots & 1 & 0 & 0 & 0\\ \vdots & \vdots & 0 & 1 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0 & 1 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ \vdots & \vdots & 0 & 0 & 1 & 0\\ \vdots & \vdots & 0 & 0 & 0 & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 1 & 0 & 0 & 0 & 1\end{bmatrix}$$

We can immediately see that the matrix is rank deficient. Columns 3 and 4 add up to give you column 1. Likewise, columns 5 and 6 add up to give you column 2. This implies that the matrix is not invertible and we cannot separately identify all fixed effects.

If we include column 1 (i.e. dummy variable for country $j = 1$), we must exclude *either* column 3 or 4 (i.e. one of the counties in country $j = 1$). For the same reasons discussed above (when selecting between the inclusion of the constant and dummy variables), it makes no difference whether we include column 1 and drop either column 3 or 4; or keep columns 3 and 4, and exclude column 1. Thus, we can only identify the model,[2]

$$Y_i = \beta Z_i + \tilde{\delta}_k + \upsilon_i$$

The general rule is, you retain the "highest" level of fixed effect. In this application, the model with county fixed effects.

5.3 Fixed Effects in Panel Data

Fixed effects are used extensively in panel data models. They are used to control for time-invariant, unit-level heterogeneity and flexible, aggregate time trends. Consider, the model

$$Y_{it} = \beta Z_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

As before, this notation is used as a shorthand. The full set of unit fixed effects are given by,

[2] I deliberately changed the parameters denoting the county fixed effects from δ_k to $\tilde{\delta}_k$ since the true model (data generating process) may contain both country and county fixed effects. However, we cannot separately identify these. We can identify $\tilde{\delta}_k$: a combination of both sets of parameters.

$$\alpha_i = \sum_{j=1}^{N} \alpha_j \mathbf{1}\{i = j\}$$

and the time fixed effects are given by,

$$\delta_t = \sum_{j=1}^{T} \delta_k \mathbf{1}\{t = k\}$$

What if we wanted to include group-level controls in the model; e.g., an indicator for treatment group status? If group membership is stable over time, then for the reasons discussed above, group membership is perfectly co-linear with the unit fixed effects. The dummy variables for treated units (a subset of the unit fixed effects) add up to the dummy variable for the treated group. You will see this again in EC338.

A final point on this. What if you need to include a variable that identifies treatment-group status in the model for *identification.* For example, in a simple two-group-two-period difference-in-differences model you have,

$$Y_{it} = \alpha + \psi D_i + \delta T_t + \beta D_i \cdot T_t + \varepsilon_{it}$$

where $D_i = \mathbf{1}\{treated\}$ and $T_t = \mathbf{1}\{t \geq t_0\}$ where t_0 is the period of treatment. The coefficient on the interaction term is the parameter of interest. If we include unit fixed effects in this model,

$$Y_{it} = \alpha_i + \delta T_t + \beta D_i \cdot T_t + \upsilon_{it}$$

then we drop the treatment-group indicator dummy variable: D_i. And if we include include time fixed effects,

$$Y_{it} = \alpha_i + \delta_t + \beta D_i \cdot T_t + \upsilon_{it}$$

we drop T_t. Is this a problem for identification? No, the group indicator, D_i, is "implied" by unit fixed effects. Because the group dummy is perfectly co-linear with the unit fixed effects, you could always have included it by excluding one of the unit fixed effects. In fact, *with a balanced panel*, replacing the group indicator with unit fixed effects will not actually change the estimated coefficient on the interaction term. It will, however, reduce the standard error of estimator.

For this reason, the most common notation used for (static) difference-in-differences is,

$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + \upsilon_{it}$$

where $D_{it} = \mathbf{1}\{treated\} \cdot \mathbf{1}\{t \geq t_0\}$. This notation is particularly useful because it does not need to be adapted when you add more time periods. However, it can only be used in applications with longitudinal data. If you are using repeated cross-sections then you cannot include unit fixed effects and must retain a treatment group indicator. We will discuss this further in EC338.

6 Compendium

6.1 Linear independence

Consider a set of k n-dimensional vectors $\{X_1, X_2, ..., X_k\}$. These vectors are,

> [**definition**] *linearly dependent* if there exists a set of scalars $\{a_1, a_2, \ldots, a_k\}$ such that
>
> $$a_1 X_1 + a_2 X_2 + ... + a_k X_k = 0$$
>
> where at least one $a_i \neq 0$.

Alternatively, they are,

> [**definition**] *linearly independent* if the only set of scalars $\{a_1, a_2, \ldots, a_k\}$ that satisfies the above condition is $a_1, a_2, \ldots, a_k = 0$.

If we collect these k column-vectors in a matrix, $X = [X_1 \ X_2 \ldots X_k]$, then the linear dependence condition can be written as,

$$a_1 X_1 + a_2 X_2 + ... + a_k X_k = \begin{bmatrix} X_1 \ X_2 \ldots X_k \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = Xa = 0$$

Given any $n \times k$ matrix X, its columns are,

> [**definition**] *linearly dependent* if there exists a vector $a \in \mathbb{R}^k$ such that $a \neq 0$ and $Xa = 0$;

or

> [**definition**] *linearly independent* if the only vector $a \in \mathbb{R}^k$ such that $Xa = 0$ is $a = 0$.

For any matrix there may be more than one vector $a \in \mathbb{R}^k$ such that $Xa = 0$. Indeed, if both $a_1, a_2 \in \mathbb{R}^k$

satisfy this condition and $a_1 \neq a_2$ then you can show that any linear combination of $\{a_1, a_2\}$ satisfies the

condition $X(a_1 b_1 + a_2 b_2) = 0$ for $b_1, b_2 \in \mathbb{R}$. Thus, there exists an entire set of vectors which satisfy this condition. This set is referred to as the,

> [**definition**] *null space* of X,

$$\mathcal{N}(X) = \{a \in \mathbb{R}^k : \ Xa = 0\}$$

It should be evident from the definition that if the columns of X are linearly independent then $\mathcal{N}(X) = \{0\}$, a singleton. That is, it just includes the 0-vector.

6.2 Vector spaces, bases, and spans

Here, we concern ourselves only with real vectors from $\mathbb{R}^n$.

> [**definition**] A *vector space*, denoted $\mathcal{V}$, refers to a set of vectors which is closed under finite addition and scalar multiplication.
>
> [**definition**] A set of k linearly independent vectors, $\{X_1, X_2, \dots, X_k\}$, forms a basis for vector space $\mathcal{V}$ if $\forall\ y \in \mathcal{V}$ there exists a set of k scalars such that,

$$y = X_1 b_1 + X_2 b_2 + \dots + X_k b_k$$

Based on these definitions, it is evident that for the Euclidean space, $\mathbb{E}^n$, any n linearly independent vectors from $\mathbb{R}^n$ is a basis. For example, any point in $\mathbb{E}^2$ can be defined as a multiple of,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Consider again the $n \times k$ matrix X, where $k < n$. Then we define the,

> [**definition**] *column space* (or *span*) of X, denoted $\mathcal{S}(X)$, as the vector space generate by the k columns of X. Formally,

$$\mathcal{S}(X) = \{y \in \mathbb{R}^n : \ y = Xb \quad \text{for some } b \in \mathbb{R}^k\}$$

A property to note about the span or column space X is,

> [**result**] $\mathcal{S}(X) = \mathcal{S}(XX')$

where XX' is a $n \times n$ matrix.

Finally, we can define the,

> [**definition**] *orthogonal column space* (or *orthogonal span*) of X as,

$$\mathcal{S}^{\perp}(X) = \{y \in \mathbb{R}^k : \; y'x = 0 \quad \forall x \in \mathcal{S}(X)\}$$

6.3 Rank

Consider a $n \times k$ matrix X, the

> [**definition**] *row rank* of X is the maximum number of linearly independent rows:

$$rowrank(X) \leq n$$

We say that matrix X has *full* row rank if $rowrank(X) = n$.

The,

> [**definition**] *column rank* of X is the maximum number of linearly independent columns:

$$colrank(X) \leq k$$

We say that matrix X has *full* column rank if $colrank(X) = k$.

An important result is,

> [**result**] the rank of X:

$$r(X) = rowrank(X) = colrank(X) \Rightarrow r(X) \leq min\{n, k\}$$

In addition, since the $r(X)$ depends on the number of linearly independent columns, we can say that,

> [**result**] the dimension of $\mathcal{S}(X)$, $dim(\mathcal{S}(X))$, is given by the $r(X)$.

Here are a few additional results,

[result] $r(X) = r(X')$

[result] $r(XY) \leq min\{r(X), r(Y)\}$

[result] $r(XY) = r(X)$ if Y is square and full rank

[result] $r(X+Y) \leq r(X) + r(Y)$

6.4 Properties of square matrices

Consider the case of a square, $n \times n$, matrix A. We say that,

[definition] A is *singular* if the $r(A) < n$,

or that,

[definition] A is *non-singular* if the $r(A) = n$.

The singularity of a square matrix is important as it determines the invertibility of a matrix, which typically relates the existence of a unique solution in systems of linear equations. Here are a few key results,

[result] There exists a matrix $B = A^{-1}$, such that $AB = I_n$ (where I_n is the identity matrix), if and only if A is non-singular.

[result] A is non-singular if and only if the determinant of A is non-zero: $det(A) \neq 0$.[1]

[result] Likewise, A is singular if and only if $det(A) = 0$.

[result] $AA^{-1} = A^{-1}A = I$

[result] $(A')^{-1} = (A^{-1})'$

[result] If their respective inverses exist, then $(AB)^{-1} = B^{-1}A^{-1}$.

[result] $det(AB) = det(A)det(B)$

[result] $det(A^{-1}) = det(A)^{-1}$

For any square matrix A,

[definition] the *trace* of A is the sum of all diagonal elements:

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

[1] These notes do not cover how to calculate the determinant of a square matrix. You should be able to find a definition easily online.

Regarding the trace of a square matrix, here are a few important results:

[result] $tr(A + B) = tr(A) + tr(B)$

[result] $tr(\lambda A) = \lambda tr(A)$ where λ is a scalar

[result] $tr(A) = tr(A')$

[result] $tr(AB) = tr(BA)$ where AB and BA are both square, but need not be of the same order.

[result] $||A|| = (tr(A'A))^{1/2}$

6.5 Properties of symmetric matrices

A symmetric matrix has the property that $A = A'$. Therefore, A must be square.

Here are a few important results concerning symmetric matrices.

[result] A^{-1} exists if $det(A) \neq 0$ and $r(A) = n$

[result] A is *diagonalizable.*[2]

[result] The eigenvector decomposition of a square matrix gives you $A = C\Lambda C^{-1}$ where Λ is a diagonal matrix of eigenvalues and \$C\$ a matrix of the corresponding eigenvectors. The symmetry of A gives you that $C^{-1} = C' \Rightarrow A = C\Lambda C'$ with $C'C = CC' = I_n$.[3]

A key definition concerning symmetric matrices is their positive definiteness:

[definition] A is *positive semi-definite* if for any $x \in \mathbb{R}^n,\ x'Ax \geq 0$.

Given the eigenvector decomposition of a symmetric matrix, *positive semi-definiteness* implies Λ is *positive semi-definite*: $\lambda_i \geq 0 \quad \forall i$. Likewise,

[definition] A is *positive definite* if for any $x \in \mathbb{R}^n,\ x'Ax > 0$.

Again, based on the egeinvector decomposition, *positive semi-definiteness* implies Λ is *positive definite*: $\lambda_i > 0 \quad \forall i$.

A few more results are:

[2] A matrix is diagonalizable if it is *similar* to some other diagonal matrix. Matrices B and C are similar if $C = PBP^{-1}$. A square matrix which is not diagonalizable is *defective.* This property relates closely to eigenvector decomposition.

[3] Recall, an eigenvalue and eigenvector pair, (λ, c), of matrix A satisfy:

$$Ac = \lambda c \Rightarrow (A - \lambda I_n)c = 0$$

[result] $tr(A) = \sum_{i=1}^{n} \lambda_i$

[result] $r(A) = r(\Lambda)$

[result] $det(A) = \prod_{i=1}^{n} \lambda_i$

This last result can be used to prove that any positive definite matrix is non-singular and therefore has an inverse.

Any full-rank, positive semi-definite, symmetric matrix B has the additional properties:

[result] $B = C\Lambda C'$ and $B^{-1} = C\Lambda^{-1}C'$

[result] We can define the square-root of B as $B^{1/2} = C\Lambda^{1/2}C'$. Similarly, $B^{-1/2} = C\Lambda^{-1/2}C'$.

6.6 Properties of idempotent matrices

An idempotent matrix has the property that $D = DD$. Therefore, D must be square.

Here are a few important results concerning idempotent matrices.

[result] D is positive definite

[result] D is diagonalizable

[result] $(I_n - D)$ is also an idempotent matrix

[result] With the exception of I_n, all idempotent matrices are singular.

[result] $r(D) = tr(D) = \sum_{i=1}^{n} \lambda_i$

[result] $\lambda_i \in \{0, 1\} \quad \forall\ i$

Projection matrices are idempotent, but need not be symmetric. However, for the purposes of this module we will deal exclusively with symmetric idempotent projection matrices.

6.7 Vector Differentiation

Here we will look at the derivatives of scalar with respect to (W.r.t.) a vector. You can also define other derivatives, such as the derivative of a vector w.r.t. a vector and the derivative of a scalar with respect to a matrix. However, these are not needed for these notes.

7 General case

Suppose $f(x) \in R$ (i.e. a scalar) and $x \in R^n$ (i.e. a $n \times 1$ vector). Then we can define the partial derivative of $f(x)$ w.r.t. x as,

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

8 Linear: scalar case

A special case is when $f(x)$ is linear in x,

$$f(x) = a'x = \sum_{i=1}^{n} a_i x_i$$

for $a \in R^n$. The derivative of $a'x$ with respect to the **vector** x can be defined as,

$$\begin{aligned} \frac{\partial a'x}{\partial x} &= \begin{bmatrix} \frac{\partial a'x}{\partial x_1} \\ \frac{\partial a'x}{\partial x_2} \\ \vdots \\ \frac{\partial a'x}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \\ &= a \end{aligned}$$

since the the partial derivate of $a'x = \sum_{i=1}^{n} a_i x_i$ w.r.t. x_i is just the scalar a_i.

The derivative of a scalar w.r.t. to a vector yields a vector of partial derivatives.

Since $a'x$ is a scalar, it is by definition symmetric: $a'x = x'a$. Thus,

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

# Linear: vector case

Suppose $f(x)$ is a linear transformation of x,

$$f(x) = A'x$$

for any $m \times n$ matrix A,

$$A = \begin{bmatrix} a'_1 \\ a'_2 \\ \vdots \\ a'_m \end{bmatrix}$$

where $a_i \in R^n \; \forall i = 1, \ldots, m$ and,

$$Ax = \begin{bmatrix} a'_1 x \\ a'_2 x \\ \vdots \\ a'_m x \end{bmatrix}$$

Note, $f(x) = Ax \in R^m$, a $m \times 1$ vector. We can then define,

$$\begin{aligned} \frac{\partial Ax}{\partial x'} &= \begin{bmatrix} \frac{\partial a'_1 x}{\partial x_1} & \frac{\partial a'_1 x}{\partial x_2} & \ldots & \frac{\partial a'_1 x}{\partial x_n} \\ \frac{\partial a'_2 x}{\partial x_1} & \frac{\partial a'_2 x}{\partial x_2} & \ldots & \frac{\partial a'_2 x}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a'_m x}{\partial x_1} & \frac{\partial a'_m x}{\partial x_2} & \ldots & \frac{\partial a'_m x}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \\ &= A \end{aligned}$$

Since Ax is $m \times 1$ column vector, we take the derivative w.r.t. x' a row vector and not the column vector x. This results in a matrix of partial derivatives.

9 Quadratic form

A second special case is where the function takes on the quadaratic form,

$$f(x) = x'Ax = \sum_{i=1}^{N} \sum_{j=1}^{n} a_{ij} x_i x_j$$

for $n \times n$ (square) matrix A. As in the first linear case, $f(x)$ is scalar.

Define $c = Ax$, the $x'Ax = x'c$. From the linear case, we know that,

$$\frac{\partial x'c}{\partial x} = c$$

Similarly, if we define $d = A'x$ then $x'Ax = d'x$. From the linear case, we know that,

$$\frac{\partial d'x}{\partial x} = d$$

We can define the total derivative as the sum of the partial derivatives w.r.t. to the first and second x. Combining these two results, we have that,

$$\frac{\partial x'Ax}{\partial x} = Ax + A'x$$

And **if** A is symmetric, this result simplifies to $2Ax$.