The Geometry and Statistical Properties of Ordinary Least Squares

Contents

1 Introduction

In these notes, we will go through the geometry and statistical properties of ordinary least squares (OLS), hopefully highlighting the trigonometric intuition behind OLS as a methodology. My own experience is that when one first learns about OLS[1] it may seem rather vague and hard to pin down as a concept, and this leads to trouble when one is trying to understand finite sample properties of OLS, test statistics, or other method of moments estimation procedures. Not because the intuition behind OLS is directly related to the understanding of statistical distributions and sampling, but because to many students OLS is sort of like a blackbox or "don't worry, it just works!" method. Without knowing the fundamentals, it becomes difficult to build further knowledge.

Before proceeding, I want to make clear that nothing in these notes is particularly innovative.

[1] And hence, by implication, GLS, IV, and GMM.

They're based on the fabulous texts by Davidson and MacKinnon (2004) and Hayashi (2000). I'm a fan of the notation in these texts, so I recommend avid students to consult them for further reading.

2 The Algebra of OLS

To get familiar with some notation, I will write the similar linear regression model as:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 X_1 + u_1, \\ y_2 &= \beta_1 + \beta_2 X_2 + u_2, \\ &\vdots \\ y_n &= \beta_1 + \beta_2 X_n + u_n, \ t = 1, ..., n, \\ \Leftrightarrow \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \end{aligned} \tag{1}$$

where $\mathbf{X}\boldsymbol{\beta}$ has typical element $\beta_1 + \beta_2 X_t$. Using this matrix notation, we can compactly use (1) for the case where we have k regressors, one of which may or may not correspond to a constant, and the others to a number of explanatory variables. Then, the matrix $\mathbf{X}$ becomes

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix},$$

where X_{ti} denotes the t-th observation on the i-th regressor, and the vector $\boldsymbol{\beta}$ now has k elements, β_1 through β_k.

Now, to estimate the simple linear regression model, let's begin by writing the error term for the t-th observation as

$$u_t = y_t - \beta_1 - \beta_2 X_t,$$

and since we have a sample size of n (and thus n error terms), we can consider the sample mean of

the error terms:

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta X_t).$$

We would like to set this sample mean equal to zero.

It's quite easy to see that if we assume that $\beta_2 = 0$, then given the assumption that $\frac{1}{n}\sum_{t=1}^{n} u_t = 0$, we can estimate β_1 to be

$$\hat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

But if we assume $\beta_2 \neq 0$, and assume $\frac{1}{n}\sum_{t=1}^{n} u_t = 0$, then we have

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta_2 X_t) = 0, \tag{2}$$

which is just one equation in two unknowns. How do we get a second equation? We use our assumption that the mean of u_t is zero conditional on the explanatory variable X_t. The conditional mean assumption implies that not only is $\mathbb{E}[u_t] = 0$, but that $\mathbb{E}[X_t u_t] = 0$ as well, since by the law of iterated expectations (LIE),

$$\mathbb{E}[X_t u_t] = \mathbb{E}\left[\mathbb{E}[X_t u_t | X_t]\right] = \mathbb{E}\left[X_t \mathbb{E}[u_t | X_t]\right] = 0. \tag{3}$$

So we can replace the population mean in (3) by the corresponding sample mean, and by making a clever substitution:

$$\frac{1}{n}\sum_{t=1}^{n} X_t(y_t - \beta_1 - \beta_2 X_t) = 0. \tag{4}$$

Thus the Equations (2) and (4) are two linear equations in two unknowns, β_1 and β_2.

Since β_1 and β_2 do not depend on t, (2) and (4) can be written as

$$\begin{aligned} \beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} X_t\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} y_t, \\ \left(\frac{1}{n}\sum_{t=1}^{n} X_t\right)\beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} X_t^2\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} X_t y_t. \end{aligned}$$

Multiplying both equations by n and using the rules of matrix multiplication, we can also write these

equations as

$$\begin{bmatrix} n & \sum_{t=1}^{n} X_t \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} X_t y_t \end{bmatrix}. \tag{5}$$

But notice something: we can use the notation in (1) and apply it to this case where we have $\mathbf{X}$ as being an $n \times 2$ matrix, $\begin{bmatrix} \boldsymbol{\iota} & \mathbf{x} \end{bmatrix}$, where $\boldsymbol{\iota}$ is a column of ones (the unit vector) and $\mathbf{x}$ denotes a column with typical element X_t. Thus, we can write

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} X_t y_t \end{bmatrix},$$

and

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} n & \sum_{t=1}^{n} X_t \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \end{bmatrix}.$$

These are the principle quantities that appear in the Equations (5)! Thus, it is clear that we can rewrite those equations using matrix notation as

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}.$$

To find the estimator $\hat{\boldsymbol{\beta}}$, we simply do some matrix algebra (and assume that $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists) to get the famous formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{6}$$

The estimator $\hat{\boldsymbol{\beta}}$ given by this formula is generally called the ordinary least squares (OLS) estimator for the linear regression model.

2.1 Estimating the multiple linear regression model

The formula (6) gives us the OLS, and method-of-moments (MM), estimator for the simple linear regression model (1). But it also gives us the MM estimator for the multiple linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, since each of the explanatory variables is required to be in the information set of the

data generating process, we have, for $i = 1, ..., k$,

$$\mathbb{E}[X_{ti}u_t] = 0;$$

which, in the corresponding sample mean form, yields

$$\frac{1}{n}\sum_{t=1}^{n} X_{ti}(y_t - \mathbf{X}_t\boldsymbol{\beta}) = 0.$$

As i varies from 1 to k, this equation yields k equations for the k unknown components of $\boldsymbol{\beta}$! It should be apparent now how handy matrix notation is (and it will become even more powerful and convenient going forward). In most cases, there will be a constant, which we may take to be the first regressor. If so $X_{t1} = 1$, and the first of these equations simply says that the sample mean of the error terms is 0.

In matrix form, after multiplying them by n, the k equations above can be written as

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{7}$$

The notation $\mathbf{0}$ is used to signify a zero vector, here a k-vector, each element of which is zero.

There are some clear parallels between (5) and (7), and it is easy to see that the OLS estimator (6) depends on $\mathbf{y}$ and $\mathbf{X}$ exclusively through a number of scalar products. Each column $\mathbf{x}_i$ of the matrix $\mathbf{X}$ corresponds to one of the regressors, as does each row $\mathbf{x}_i^\top$ of the transposed matrix $\mathbf{X}^\top$. Thus, we can write $\mathbf{X}^\top\mathbf{y}$ as

$$\mathbf{X}^\top\mathbf{y} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_k^\top \end{bmatrix} \mathbf{y} = \begin{bmatrix} \mathbf{x}_1^\top\mathbf{y} \\ \mathbf{x}_2^\top\mathbf{y} \\ \vdots \\ \mathbf{x}_k^\top\mathbf{y} \end{bmatrix}.$$

The elements of the rightmost expression here are just the scalar products of the regressors $\mathbf{x}_i$ with

the regressand $\mathbf{y}$. Similarly, we can write $\mathbf{X}^\top\mathbf{X}$ as

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_k^\top \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \cdots & \mathbf{x}_1^\top\mathbf{x}_k \\ \mathbf{x}_2^\top\mathbf{x}_1 & \mathbf{x}_2^\top\mathbf{x}_2 & \cdots & \mathbf{x}_2^\top\mathbf{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_k^\top\mathbf{x}_1 & \mathbf{x}_k^\top\mathbf{x}_2 & \cdots & \mathbf{x}_k^\top\mathbf{x}_k \end{bmatrix}.$$

Thus, multiplying out (7) and doing some matrix algebra (and, again, assuming that $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists) will give us an estimate for $\boldsymbol{\beta}$, akin to (6) for the case with k elements.

2.2 OLS minimises the sum of squared residuals

So far, we have derived the estimator (6) by using the method of moments. Deriving it this way has at least two major advantages. First, the method of moments is a very general and very powerful principle in estimation, one that we will encounter again and again throughout econometrics. Second, by using the method of moments, we are able to obtain (6) without making any use of [matrix] calculus. However, as mentioned, (6) is generally referred to as the OLS estimator, not the MM estimator. We will show why this is so.

For the multiple linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{8}$$

the expression $y_t - \mathbf{X}_t\boldsymbol{\beta}$ is equal to the error term for the t-th observation, but only if the correct value of the parameter vector $\boldsymbol{\beta}$ is used. Although we do not observe the error term, we can calculate the value implied by a hypothetical value, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ as

$$y_t - \mathbf{X}_t\tilde{\boldsymbol{\beta}}.$$

This is called the residual for observation t. From this, form the sum of squared residuals (SSR):

$$\begin{aligned}\text{SSR}(\tilde{\boldsymbol{\beta}}) &\equiv \sum_{t=1}^{n}(y_t - \mathbf{X}_t\tilde{\boldsymbol{\beta}})^2 \\ &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}).\end{aligned}$$

The SSR (or residual sum of squares) is a function of $\tilde{\boldsymbol{\beta}}$ because the residual depends on it. The OLS estimate, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ is the $\tilde{\boldsymbol{\beta}}$ that minimises this function:

$$\hat{\boldsymbol{\beta}} \equiv \arg\min_{\tilde{\boldsymbol{\beta}}} \text{SSR}(\tilde{\boldsymbol{\beta}}). \tag{9}$$

Since $\hat{\boldsymbol{\beta}}$ depends on the sample $(\mathbf{y}, \mathbf{X})$, the OLS estimate is in general different from the true value $\boldsymbol{\beta}$; if $\hat{\boldsymbol{\beta}}$ equals $\boldsymbol{\beta}$, it is by sheer accident.

By having squared residuals in the objective function, this method imposes a heavy penalty on large residuals; the OLS residuals is chosen to prevent large residuals for a few observations at the expense of tolerating relatively small residuals for many other observations.

A surefire way of solving he minimisation problem is to derive the FOCs by setting the partial derivatives equal to zero. To this end we seek a k-dimensional vector of partial derivatives, $\partial\text{SSR}(\tilde{\boldsymbol{\beta}})/\partial\tilde{\boldsymbol{\beta}}$.The task is facilitated by writing

$$\begin{aligned}\text{SSR}(\tilde{\boldsymbol{\beta}}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= (\mathbf{y}^\top - \tilde{\boldsymbol{\beta}}^\top\mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\mathbf{X}^\top\mathbf{y} + \tilde{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\tilde{\boldsymbol{\beta}} \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\tilde{\boldsymbol{\beta}} \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{a}^\top\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}^\top\mathbf{A}\tilde{\boldsymbol{\beta}},\end{aligned} \tag{10}$$

where $\mathbf{a} = \mathbf{X}^\top\mathbf{y}$ and $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$. The term $\mathbf{y}^\top\mathbf{y}$ does no depend on $\tilde{\boldsymbol{\beta}}$ and so can be ignored in the

differentiation of SSR($\tilde{\boldsymbol{\beta}}$). Recalling from matrix algebra that

$$\frac{\partial \mathbf{a}^\top \tilde{\boldsymbol{\beta}}}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{a},$$
$$\frac{\partial \tilde{\boldsymbol{\beta}}^\top \mathbf{A} \tilde{\boldsymbol{\beta}}}{\partial \tilde{\boldsymbol{\beta}}} = 2\mathbf{A}\tilde{\boldsymbol{\beta}},$$

for symmetric $\mathbf{A}$, the k-dimensional vector of partial derivatives is

$$\frac{\partial \text{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{a} + 2\mathbf{A}\tilde{\boldsymbol{\beta}}.$$

The FOCs are obtained by setting this equal to zero. Using our definitions of $\mathbf{a}$ and $\mathbf{A}$, and doing some rearranging, we can write the FOCs as

$$\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}, \tag{11}$$

Here, we have replaced $\tilde{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}$ because the OLS estimate $\hat{\boldsymbol{\beta}}$ is the $\tilde{\boldsymbol{\beta}}$ that satisfies the FOCs. These k equations are called the normal equations.

The vector of residuals evaluated at $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$,

$$\hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{12}$$

is called the vector of OLS residuals. Its t-th element is $\hat{u}_t \equiv y_t - \mathbf{X}_t^\top \hat{\boldsymbol{\beta}}$. Also note that rearranging (11) gives

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \Leftrightarrow \mathbf{X}^\top \hat{\mathbf{u}} = \mathbf{0}$$
$$\Leftrightarrow \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t \hat{u}_t = \mathbf{0} \Leftrightarrow \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t (y_t - \mathbf{X}_t^\top \hat{\boldsymbol{\beta}}) = \mathbf{0},$$

which shows that the normal equations can be interpreted as the sample analogue of the orthogonality conditions (7).

To be sure the FOCs are just a necessary condition for minimisation, and we have to check the

second-order condition to make sure that $\hat{\boldsymbol{\beta}}$ achieves the minimum, not the maximum. Those who are familiar with the Hessian of a function of several variables can immediately recognise that the second-order condition is satisfied because $\mathbf{X}^\top\mathbf{X}$ is positive definite. There is, however, a more direct way to show that $\hat{\boldsymbol{\beta}}$ indeed achieves a minimum. It utilises the "add-and-subtract" strategy, which is effective when the objective function is quadratic, as we have here.

In addition, consider the OLS estimate of σ^2 (the variance of the error term), denoted s^2, as being the SSR divided by $n-k$,

$$s^2 \equiv \frac{\text{SSR}}{n-k} = \frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{n-k}. \tag{13}$$

This definition presumes that $n > k$, otherwise s^2 is not well defined. As will be shown later, dividing the SSR by $n-k$ (called the degrees of freedom) rather than by n makes this estimate unbiased for σ^2. The intuitive reason is that k parameters have to be estimated before obtaining the residual vector $\hat{\mathbf{u}}$ used to calculate s^2. More specifically, $\hat{\mathbf{u}}$ has to satisfy the k normal equations (11), which limits the variability of the residual. The square root of s^2, s, is called the standard error of the regression. It is an estimate of the standard deviation of the error term.

2.2.1 Two expressions for the OLS estimator

Thus, we have obtained a system of k linear simultaneous equations in k unknowns in $\hat{\boldsymbol{\beta}}$, (11). By assuming no multicollinearity, the coefficient matrix $\mathbf{X}^\top\mathbf{X}$ is positive definite and hence nonsingular, such that $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists. So, we can premultiply (11) by $(\mathbf{X}^\top\mathbf{X})^{-1}$ to get

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{14}$$

which is nothing but the OLS estimator, which we found in (6) using the method of moments.

There are a myriad of ways and notations to get the OLS estimator. We could've set up (9) by first defining a criterion function such as

$$\phi(\boldsymbol{\beta}) = \sum_{t=1}^{n} u_t^2(\boldsymbol{\beta}) = \mathbf{u}^\top\mathbf{u},$$

and then using the chain rule of matrix calculus[2] would give us the FOC:

$$\begin{aligned}\frac{\partial \phi(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \mathbf{u}^\top \mathbf{u}}{\partial \boldsymbol{\beta}} \\ \mathbf{0} &= \frac{\partial \mathbf{u}}{\partial \boldsymbol{\beta}} \frac{\partial \mathbf{u}^\top \mathbf{u}}{\partial \mathbf{u}} \\ \mathbf{0} &= -\mathbf{X}^\top 2\mathbf{u} \\ \mathbf{0} &= -2\mathbf{X}^\top \mathbf{u}.\end{aligned}$$

Substituting in our value for $\mathbf{u}$ gives:

$$\begin{aligned}\mathbf{0} &= -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^\top\mathbf{y} \\ \implies \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.\end{aligned}$$

There's also another way to write the OLS estimator. We can make use of the inverse term/operator in (6) and write as

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{y} = \mathbf{S}^{-1}_{\mathbf{X}^\top\mathbf{X}}\mathbf{S}_{\mathbf{X}^\top\mathbf{y}}, \tag{15}$$

where

$$\begin{aligned}\mathbf{S}_{\mathbf{X}^\top\mathbf{X}} &= \frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^\top\mathbf{X}_t, \\ \mathbf{S}_{\mathbf{X}^\top\mathbf{y}} &= \frac{1}{n}\mathbf{X}^\top\mathbf{y} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^\top y_t.\end{aligned}$$

The data matrix form (14) is more convenient for developing the finite-sample results, while the sample average form (15) is the form to be utilised for when we look at large-sample theory.

[2] Turkington (2007, 2013) provides a good treatment of matrix calculus for those interested.

3 Geometry of Linear Regression

With the algebra and derivation of OLS out of the way, now we can focus on the geometry of OLS or the numerical properties of the OLS estimates.

Just to reiterate, we have n observations of a linear regression model with k regressors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{16}$$

and $\mathbf{y}$ and $\mathbf{u}$ are n-vectors and, $\mathbf{X}$ is an $n \times k$ matrix, one column of which can be the unit vector (or any constant, for that matter), and $\boldsymbol{\beta}$ is a k-vector. Using either the method-of-moments or calculus, we can get OLS estimates of the vector $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{17}$$

In order to understand the numerical properties of OLS estimates, it is useful to look at them from the perspective of Euclidean geometry. This geometrical interpretation is remarkably simple and powerful (especially we want to construct test statistics). Essentially, it involves using Pythagoras' Theorem and a little bit of high-school trigonometry in the context of finite-dimensional vector spaces.

For those that are not interested in the definitions and explanation of the geometry of vector spaces, you can skip ahead to Section 3.2.

3.1 Geometry of vector spaces

The elements of the vectors we will focus on are real numbers. The usual notation for the real line is $\mathbb{R}$, and it is therefore natural to denote the set of n-vectors as $\mathbb{R}^n$. However, in order to use the insights of Euclidean geometry to enhance our understanding of the algebra of vectors and matrices, it is desirable to introduce the notion of a Euclidean space in n dimensions, which we will denote as E^n. The difference between $\mathbb{R}^n$ and E^n is not that they consist of different sorts of vectors, but rather that a wider set of operations is defined on E^n. A shorthand way of saying that a vector $\mathbf{x}$ belongs to an n-dimensional Euclidean space is to write $\mathbf{x} \in E^n$.

Addition and subtraction of vectors in E^n is no different from the addition and subtraction of $n \times 1$ matrices/vectors we've looked at before. The same thing is true of multiplication by a scalar in E^n. The final operation essential to E^n is that of the scalar or inner product. For any two vectors $\mathbf{x}, \mathbf{y} \in E^n$, their scalar product is

$$\langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{x}^\top \mathbf{y}.$$

This notation should be somewhat familiar. Note that $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$, since $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$. Thus the scalar product is commutative.

The scalar product is what allows us to make a close connection between n-vectors considered as matrices and considered as geometrical objects. It allows us to define the length of any vector in E^n. The length, or norm, of a vector $\mathbf{x}$ is simply

$$\|x\| \equiv (\mathbf{x}^\top \mathbf{x})^{1/2}.$$

This is just the square root of the inner product of $\mathbf{x}$ with itself. In scalar terms, it is

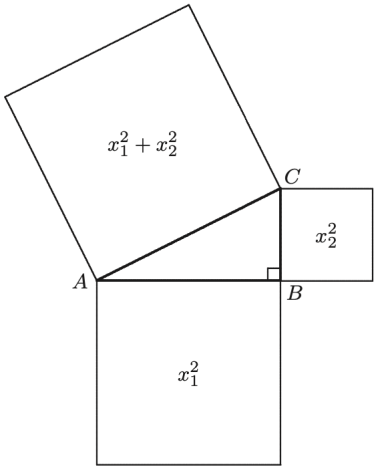
$$\|\mathbf{x}\| \equiv \left(\sum_{t=1}^{n} x_t^2 \right)^{1/2}. \tag{18}$$

Those that are familiar with set theory and real analysis in microeconomics, should be fairly comfortable with this.

3.1.1 Pythagoras' Theorem

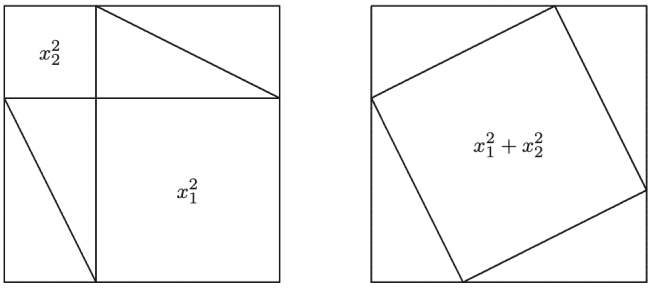
The definition (18) is inspired by Pythagoras' Theorem, which says that the square on the longest side of a right-angled triangle is equal to the sum of the squares on the other two sides. This longest side is called a hypotenuse. Pythagoras' Theorem is illustrated in Figure 1. This figure shows a right-angled triangle, ABC, with hypotenuse AC, and two other sides, AB and BC, of lengths x_1 and x_2, respectively. The squares on each of the three sides of the triangle are drawn, and the area of the square on the hypotenuse is shown as $x_1^2 + x_2^2$, in accordance with the theorem.

Figure 1: Pythagoras' Theorem



A beautiful proof of Pythagoras' Theorem, not often found in high school geometry and trigonometry texts, is shown in Figure 2. Two squares of equal area are drawn. Each square contains four copies of the same right-angled triangle. The square on the left also contains the squares on the two shorter sides of the triangle, while the square on the right contains the square on the hypotenuse. The theorem follows at once.

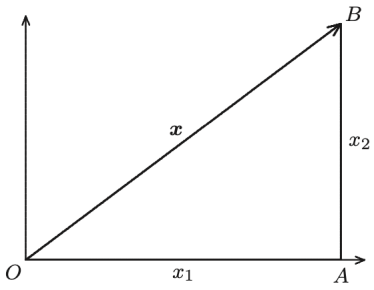
Figure 2: Proof of Pythagoras' Theorem



Any vector $\mathbf{x} \in E^2$ has two components, usually denoted as x_1 and x_2. These two components can be interpreted as the Cartesian coordinates of the vector in the plane. The situation is illustrated in Figure 3. With O as the origin of the coordinates, a right-angled triangle is formed by the lines OA, AB, and OB. The length of the horizontal side of the triangle, OA, is the horizontal coordinate x_1.

The length of the vertical side, AB, is the vertical coordinate x_2. Thus the point B has Cartesian coordinates (x_1, x_2). The vector $\mathbf{x}$ itself is usually represented as the hypotenuse of the triangle, OB, that is, the directed line (depicted as an arrow) joining the origin to the point B, with coordinates (x_1, x_2). By Pythagoras' Theorem, the length of the vector $\mathbf{x}$, the hypotenuse of the triangle, is $(x_1^2 + x_2^2)^{1/2}$. This is what (18) becomes for the special case when $n = 2$.

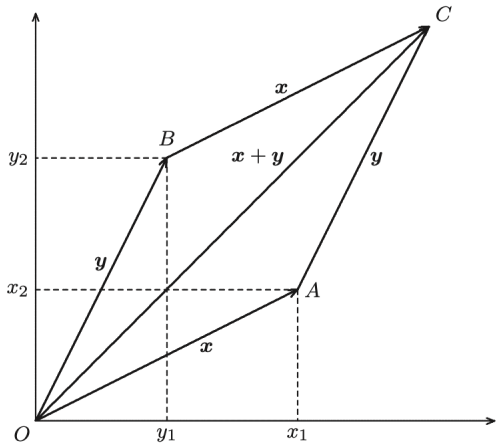
Figure 3: A Vector $\mathbf{x}$ in E^2



3.1.2 Vector geometry in two dimensions

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors E^2, with components (x_1, x_2) and (y_1, y_2), respectively. Then, by the rules of matrix addition, the components $\mathbf{x} + \mathbf{y}$ are $(x_1 + y_1, x_2 + y_2)$. Figure shows how the addition of $\mathbf{x}$ and $\mathbf{y}$ can be performed geometrically in two different ways. The vector $\mathbf{x}$ is drawn as the directed line segment, or arrow, from the origin O to the point A with coordinates (x_1, x_2). The vector $\mathbf{y}$ can be drawn similarly and represented by the arrow OB However, we could also draw $\mathbf{y}$ starting, not at O, but at the point reached after drawing, namely A. The arrow AC has the same length and direction of OB, and we will see in general that arrows with the same length and direction can be taken to represent the same vector. It is clear by construction that the coordinates of C are $(x_1 + y_1, x_2 + y_2)$, that is, the coordinates of $\mathbf{x} + \mathbf{y}$. Thus the sum $\mathbf{x} + \mathbf{y}$ is represented geometrically by the arrow OC.

Figure 4: Addition of Vectors



The classical way of adding vectors geometrically is to form a parallelogram using the line segments OA and OB that represent the two vectors as adjacent sides of the parallelogram. The sum of the two vectors is then the diagonal through O of the resulting parallelogram. It is easy to see that this classical method also gives the result that the sum of the two vectors is represented by the arrow OC, since the figure $OACB$ is just the parallelogram required by the construction, and OC is its diagonal through O. The parallelogram construction also shows clearly that vector addition is commutative, since $\mathbf{y} + \mathbf{x}$ is represented by OB, for $\mathbf{y}$, followed by BC, for $\mathbf{x}$. The end result is once again OC.

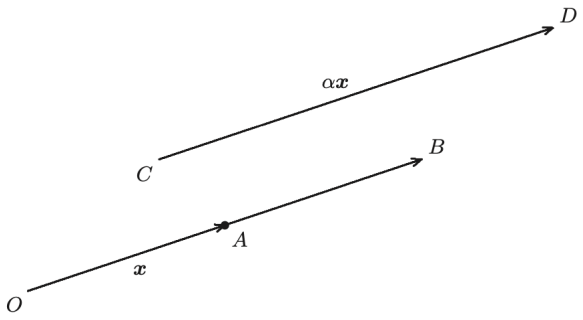
Multiplying a vector by a scalar is also very easy to represent graphically. If a vector $\mathbf{x}$ is multiplied by a scalar α, then $\alpha\mathbf{x}$ has components $(\alpha x_1, \alpha x_2)$. This is depicted in Figure 5, where $\alpha = 2$. The line segments OA and OB represent $\mathbf{x}$ and $\alpha\mathbf{x}$, respectively. It is clear that even if we move $\alpha\mathbf{x}$ so that it starts somewhere other than O, as with CD in the figure, the vectors $\mathbf{x}$ and $\alpha\mathbf{x}$ are always parallel. If α were negative, then $\alpha\mathbf{x}$ would simply point in the opposite direction. Thus, for $\alpha = -2$, $\alpha\mathbf{x}$ would be represented by DC, rather than CD.

Another property of multiplication by a scalar is clear from Figure 5. By direct calculation,

$$\|\alpha\mathbf{x}\| = \langle \alpha\mathbf{x}, \alpha\mathbf{x} \rangle^{1/2} = |\alpha|(\mathbf{x}^\top\mathbf{x})^{1/2} = |\alpha|\|\mathbf{x}\|. \tag{19}$$

Since $\alpha = 2$, OB and CD in the figure are twice as long as OA.

Figure 5: Multiplication by a Scalar



3.1.3 The geometry of scalar products

The scalar product of two vectors $\mathbf{x}$ and $\mathbf{y}$, whether in E^2 or E^n, can be expressed geometrically in terms of the lengths of the two vectors and the angle between them, and this result will turn out to be very useful. In the case of E^2, it is natural to think of the angle between two vectors as the angle between the two line segments that represent them.

If the angle between two vectors is 0, they must be parallel. The vector $\mathbf{y}$ is parallel to the vector $\mathbf{x}$ if $\mathbf{y} = \alpha\mathbf{x}$ for some suitable α. In that event,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \alpha\mathbf{x} \rangle = \alpha \mathbf{x}^\top \mathbf{x} = \alpha \|\mathbf{x}\|^2.$$

From (4), we know that $\|\mathbf{y}\| = |\alpha| \|\mathbf{x}\|$, and so, if $\alpha > 0$, it follows that

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\|. \tag{20}$$

Of course, this result is only true of $\mathbf{x}$ and $\mathbf{y}$ are parallel and point in the same direction (rather than in opposite directions).

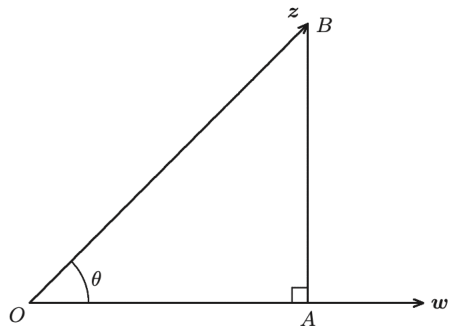
For simplicity, consider initially two vectors, $\mathbf{w}$ and $\mathbf{z}$, both of length 1, and let θ denote the angle between them. This is illustrated in Figure 6. Suppose that the first vector, $\mathbf{w}$, has coordinates $(1, 0)$. It is therefore represented by a horizontal line of length 1 in the figure. Then suppose that $\mathbf{z}$ also has length 1, so $\|\mathbf{z}\| = 1$. Then, by elementary trigonometry, the coordinates of $\mathbf{z}$ must be $(\cos\theta, \sin\theta)$.

To show this, note first that, if so

$$\|\mathbf{z}\|^2 = \cos^2\theta + \sin^2\theta = 1, \tag{21}$$

as required. Next, consider the right-angled triangle OAB, in which the hypotenuse OB represents $\mathbf{z}$ and is of length 1, by (21). The length of the side AB opposite of O is $\sin\theta$, the vertical coordinate of $\mathbf{z}$.[3] Then the sine of the angle BOA is given, by the usual trigonometric rule, by the ratio of the length of the opposite side AB to that of the hypotenuse OB. This ratio is $\sin\theta/1 = \sin\theta$, and so the angle BOA is indeed equal to θ.

Figure 6: The Angle Between Two Vectors



Now, let us compute the scalar product of $\mathbf{w}$ and $\mathbf{z}$. It is

$$\langle \mathbf{w}, \mathbf{z} \rangle = \mathbf{w}^\top \mathbf{z} = w_1 z_1 + w_2 z_2 = z_1 = \cos\theta,$$

because $w_1 = 1$ and $w_2 = 0$. This result holds for vectors $\mathbf{w}$ and $\mathbf{z}$ of length 1. More generally, let $\mathbf{x} = \alpha\mathbf{w}$ and $\mathbf{y} = \gamma\mathbf{z}$, for positive scalars α and γ. Then $\|\mathbf{x}\| = \alpha$ and $\|\mathbf{y}\| = \gamma$. Thus we have

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \alpha\gamma \mathbf{w}^\top \mathbf{z} = \alpha\gamma \langle \mathbf{w}, \mathbf{z} \rangle .$$

Because $\mathbf{x}$ is parallel to $\mathbf{w}$, and $\mathbf{y}$ is parallel to $\mathbf{z}$, the angle between $\mathbf{x}$ and $\mathbf{y}$ is the same as that

[3] Now would be a good time to brush up on high school trigonometry and recall the SOHCAHTOA rule.

between $\mathbf{w}$ and $\mathbf{z}$, namely θ. Therefore,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta \tag{22}$$

This is the general expression, in geometrical terms, for the scalar product of two vectors. It is true in E^n just as it is in E^2, although we have not proven this. In fact, we have not quite proved (22) even for the two dimensional case, because we made the simplifying assumption that the direction of $\mathbf{x}$ and $\mathbf{w}$ is horizontal.

The cosine of the angle between two vectors provides a natural way to measure how close two vectors are in terms of their directions. Recall that $\cos\theta$ varies between -1 and 1; if we measure angles in radians, $\cos 0 = 1$, $\cos \pi/2 = 0$, and $\cos \pi = -1$. Thus, $\cos\theta$ will be 1 for vectors that are parallel, 0 for vectors that are right angles to each other, and -1 for vectors that point in directly opposite directions. If the angle θ between the vectors $\mathbf{x}$ and $\mathbf{y}$ is a right angle, its cosine is 0, and so, from (22), the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle$ is 0. Conversely, if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ then $\cos\theta = 0$ unless $\mathbf{x}$ or $\mathbf{y}$ is a zero vector. If $\cos\theta = 0$, it follows that $\theta = \pi/2$. Thus, if two nonzero vectors have a zero scalar product, they are at right angles. Such vectors are often said to be orthogonal. This definition implies that the zero vector is orthogonal to everything.

Since the cosine function can take on values only between -1 and 1, a consequence of (22) is that

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \tag{23}$$

This result is called the Cauchy-Schwartz inequality, and it says that the inner product of $\mathbf{x}$ and $\mathbf{y}$ can never be greater than the length of the vector $\mathbf{x}$ times the length of the vector $\mathbf{y}$. Only if $\mathbf{x}$ and $\mathbf{y}$ are parallel does the Cauchy-Schwartz inequality (23) become the quality (20).

3.1.4 Subspaces of Euclidean space

Just one more concept to nail down before we can talk about linear regression in more direct terms – though the definitional concept of orthogonality above should've caught your attention.

We need the concept of a subspace of Euclidean space E^n. Why? Because we want to represent

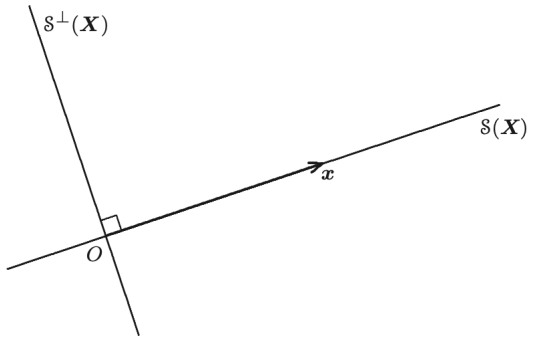
(16) without needing to draw all n dimensions, and to compactly represent using just the vectors $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{u}$.

Normally such a subspace will have a dimension lower than n. The easiest way to define a subspace of E^n is in terms of a set of basis vectors. A subspace that is of particular interest to us is the one for which the columns of $\mathbf{X}$ provide the basis vectors. We may denote the k columns of $\mathbf{X}$ as $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$. Then the subspace associated with these k basis vectors will be denoted by $\mathcal{S}(\mathbf{X}\}$ or $\mathcal{S}(\mathbf{x}_1, ..., \mathbf{x}_k)$. The basis vectors are said to span this subspace, which will in general be a k-dimensional subspace.

The subspace $\mathcal{S}(\mathbf{x}_1, ..., \mathbf{x}_k)$ consists of every vector that can be formed as a linear combination of the $\mathbf{x}_i,\ i = 1, ..., k$. Formally, it is defined as

$$\mathcal{S}(\mathbf{x}_1, ..., \mathbf{x}_k) \equiv \left\{ \mathbf{z} \in E^n | \mathbf{z} = \sum_{i=1}^{k} b_i \mathbf{x}_i,\ b_i \in \mathbb{R} \right\}. \tag{24}$$

Figure 7: The Spaces $\mathcal{S}(\mathbf{X})$ and $\mathcal{S}^{\perp}(\mathbf{X})$



The subspace defined in (24) is called the subspace spanned by the $\mathbf{x}_i$, or the column space of $\mathbf{X}$; less formally, it may simply be referred to as the span of $\mathbf{X}$, or the span of the $\mathbf{x}_i$.

The orthogonal complement of $\mathcal{S}(\mathbf{X})$ in E^n, which is denoted $\mathcal{S}^{\perp}(\mathbf{X})$, is the set of all vectors $\mathbf{w}$ in E^n that are orthogonal to everything in $\mathcal{S}(\mathbf{X})$. This means that, for every $\mathbf{z}$, in $\mathcal{S}(\mathbf{X})$, $\langle \mathbf{w}, \mathbf{z} \rangle = \mathbf{w}^{\top}\mathbf{z} = 0$. Formally,

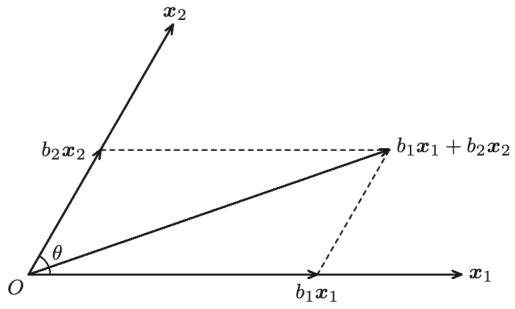
$$\mathcal{S}^{\perp}(\mathbf{X}) \equiv \left\{ \mathbf{w} \in E^n | \mathbf{w}^{\top}\mathbf{z} = 0,\ \forall \mathbf{z} \in \mathcal{S}(\mathbf{X}) \right\}.$$

If the dimension of $\mathcal{S}(\mathbf{X})$ is k, then the dimension of $\mathcal{S}^{\perp}(\mathbf{X})$ is $n - k$.

Figure 7 illustrates the concepts of a subspace and its orthogonal complement for the simplest cast, in which $n = 2$ and $k = 1$. The matrix $\mathbf{X}$ has only one column in this case, and it is therefore represented in the figure by a single vector, denoted $\mathbf{x}$. As a consequence, $\mathcal{S}(\mathbf{X})$ is one-dimensional, and, since $n = 2$, $\mathcal{S}^{\perp}(\mathbf{X})$ is also one-dimensional. Notice that $\mathcal{S}(\mathbf{X})$ and $\mathcal{S}^{\top}(\mathbf{X})$ would be the same if $\mathbf{x}$ were any vector, except for the origin, parallel to the straight line that represents $\mathcal{S}(\mathbf{X})$.

Now let us return E^n. Suppose, to being with, that $k = 2$. We have two vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, which span a subspace of, at most, two dimensions. It is always possible to represent vectors in a two-dimensional space on a piece of paper, whether that space is E^2 itself or, as in this case, the two-dimensional subspace of E^n spanned by the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. To represent the first vector, $\mathbf{x}_1$, we choose an origin and a direction, both of which are entirely arbitrary, and draw an arrow of length $\|\mathbf{x}_1\|$ in that direction. Suppose that the origin is the point O in Figure 8, and that the direction is the horizontal direction in the plane of the page. Then, an arrow to represent $\mathbf{x}_1$ can be drawn as shown in the figure. For $\mathbf{x}_2$, we compute its length, $\|\mathbf{x}_2\|$, and the angle, θ, that it makes with $\mathbf{x}_1$. Suppose for now that $\theta \neq 0$. Then we choose as our second dimension the vertical direction in the plane of the page, with the result that we can draw an arrow for $\mathbf{x}_2$, as shown.

Figure 8: A 2-Dimensional Subspace

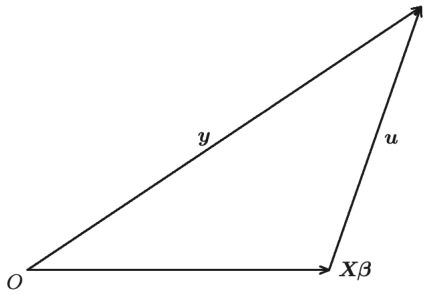


Any vector in $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$ can be drawn in the plane of Figure 8. Consider, the linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$ given by the expression $\mathbf{z} \equiv b_1\mathbf{x}_1 + b_2\mathbf{x}_2$. We could draw the vector $\mathbf{z}$ by computing its length and the angle that it makes with $\mathbf{x}_1$. Alternatively, we could apply the rules for adding vectors geometrically that were illustrated in Figure 4 to the vectors $b_1\mathbf{x}_1$ and $b_2\mathbf{x}_2$. This is illustrated in the

figure for the case in which $b_1 = 2/3$ and $b_2 = 1/2$.

We can finally represent the regression model (16) geometrically. This is done in Figure 9. The horizontal direction is chosen for the vector $\mathbf{X}\boldsymbol{\beta}$, and then the other two vectors $\mathbf{y}$ and $\mathbf{u}$ are shown in the plane of the page. It is clear that, by construction, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. Notice that $\mathbf{u}$ is not orthogonal to $\mathbf{X}\boldsymbol{\beta}$. The figure contains no reference to any system of axes, because there would be n of them, and we would not be able to avoid needing n dimensions to treat them all.

Figure 9: The Geometry of the Linear Regression Model



3.1.5 Linear independence

In order to define the OLS estimator by the formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \tag{25}$$

it is necessary to assume that the $k \times k$ square matrix $\mathbf{X}^\top \mathbf{X}$ is invertible, i.e., that it is nonsingular or that it has full rank. This condition is equivalent to the condition that the columns of $\mathbf{X}$ should linearly independent. This is a very important concept for econometrics. Note that the meaning of linear independence is quite different from the meaning of statistical independence, which we covered before. It is important not to confuse these two concepts.

The vectors $\mathbf{x}_1$ through $\mathbf{x}_k$ are said to be linearly dependent if we can write one of them as a linear combination of the others. In other words, there is a vector $\mathbf{x}_j$, $1 \leq j \leq k$, and coefficients c_i such

that

$$\mathbf{x}_j = \sum_{i \neq j} c_i \mathbf{x}_i. \tag{26}$$

Another, equivalent, definition is that there exist coefficients b_i, at least one of which is nonzero, such that

$$\sum_{i=1}^{k} b_i \mathbf{x}_i = \mathbf{0}. \tag{27}$$

It is clear from this definition that, if any of the $\mathbf{x}_i$ is itself equal to the null vector, then the $\mathbf{x}_i$ are linearly dependent. If $\mathbf{x}_j = \mathbf{0}$, for example, then (27) will be satisfied if we make b_j nonzero and set $b_i = 0,\ \forall i \neq j$.

If the vectors $\mathbf{x}_i,\ i = 1, ..., k$, are the columns of an $n \times k$ matrix $\mathbf{X}$, then another way of writing (27) is

$$\mathbf{X}\mathbf{b} = \mathbf{0}, \tag{28}$$

where $\mathbf{b}$ is a k-vector with typical element b_i. The set of vectors $\mathbf{x}_i,\ i = 1, ..., k$, is linearly independent if it is not linearly dependent, that is, if there are no coefficients c_i such that (26) is true, or no coefficients b_i such that (27) is true, or no vector $\mathbf{b}$ such that (28) is true (again, these are equivalent statements).

It is easy to show that if the columns of $\mathbf{X}$ are linearly dependent, the matrix $\mathbf{X}^\top\mathbf{X}$ is not invertible. Premultiplying (28) by $\mathbf{X}^\top$ yields

$$\mathbf{X}^\top\mathbf{X}\mathbf{b} = \mathbf{0}. \tag{29}$$

Now suppose that the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible. If so, there exists a matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ such that $(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}) = \mathbf{I}$. Thus, equation (29) implies that

$$\begin{aligned} \mathbf{b} &= \mathbf{I}\mathbf{b} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})\mathbf{b} \\ &= \mathbf{0}. \end{aligned}$$

But this is a contradiction, since we have assumed that $\mathbf{b} \neq \mathbf{0}$. Therefore, we conclude that the matrix

$(\mathbf{X}^\top\mathbf{X})^{-1}$ cannot exist when the columns of $\mathbf{X}$ are linearly dependent. Thus, a necessary condition for the existence of $(\mathbf{X}^\top\mathbf{X})^{-1}$ is that the columns of $\mathbf{X}$ should be linearly independent. With a little more work, it can be shown that this condition is also sufficient, and so, if the regressors $\mathbf{x}_1, ..., \mathbf{x}_k$ are linearly independent, $\mathbf{X}^\top\mathbf{X}$ is invertible.

If the k columns of $\mathbf{X}$ are not linearly independent, then they will span a subspace of dimension less than k, say k', where k' is the largest number of columns of $\mathbf{X}$ that are linearly independent of each other. Then number k' is called the rank of $\mathbf{X}$. Look again at Figure 8, and imagine that the angle θ between $\mathbf{x}_1$ and $\mathbf{x}_2$ tends to zero. If $\theta = 0$, then $\mathbf{x}_1$ and $\mathbf{x}_2$ are parallel, and we can write $\mathbf{x}_1 = \alpha\mathbf{x}_2$, for some scalar α. But this means that

$$\mathbf{x}_1 - \alpha\mathbf{x}_2 = \mathbf{0},$$

and so a relation of the form (27) holds between $\mathbf{x}_1$ and $\mathbf{x}_2$, which they are therefore linearly dependent. In the figure, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are parallel, then only one dimension is used, and there is no need for the second dimension in the plane of the page. Thus, in this case, $k = 2$ and $k' = 1$.

When the dimension of $\mathcal{S}(\mathbf{X})$ is $k' < k$, $\mathcal{S}(\mathbf{X})$ will be identical to $\mathcal{S}(\mathbf{X}')$, where $\mathbf{X}'$ is an $n \times k'$ matrix consisting of any k' linearly independent columns of $\mathbf{X}$. For example, consider the following $\mathbf{X}$ matrix, which is 5×3:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{30}$$

The columns of this matrix are clearly not linearly independent, since

$$\mathbf{x}_1 = \frac{1}{4}\mathbf{x}_2 + \mathbf{x}_3.$$

However, any two of the columns are linearly independent, and so

$$\mathcal{S}(\mathbf{X}) = \mathcal{S}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{S}(\mathbf{x}_1, \mathbf{x}_3) = \mathcal{S}(\mathbf{x}_2, \mathbf{x}_3).$$

For the remainder of these notes, unless stated otherwise, we will assume that all the columns of any regressor matrix $\mathbf{X}$ is linearly independent.

3.2 Geometry of OLS estimation

The geometrical interpretation of OLS estimation, that is, MM estimation of linear regression models, is simple and intuitive. In many cases, it entirely does away with the need for algebraic proofs.

As we see, any point in a subspace $\mathcal{S}(\mathbf{X})$, where $\mathbf{X}$ is an $n \times k$ matrix, can be represented as a linear combination of the columns of $\mathbf{X}$. We can partition $\mathbf{X}$ in terms of its columns explicitly as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{bmatrix}.$$

In order to compute the matrix product $\mathbf{X}\boldsymbol{\beta}$ in terms of this partitioning, we need to partition the vector $\boldsymbol{\beta}$ by its rows. Since $\boldsymbol{\beta}$ only has one column, the elements of the partitioned vector are just the individual elements of $\boldsymbol{\beta}$. Thus, we find that

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} &= \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \\ &= \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + ... + \mathbf{x}_k\beta_k \\ &= \sum_{i=1}^{k} \beta_i \mathbf{x}_i, \end{aligned}$$

which is just a linear combination of the columns of $\mathbf{X}$. In fact, it is clear from the definition (24) that any linear combination of the columns of $\mathbf{X}$, and thus any element of the subspace $\mathcal{S}(\mathbf{X})$, can be written

as $\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$. The specific linear combination (24) is constructed by using $\boldsymbol{\beta} = (b_1, ..., b_k)^\top$. Thus, every n-vector $\mathbf{X}\boldsymbol{\beta}$ belongs to $\mathcal{S}(\mathbf{X})$, which is, in general, a k-dimensional subspace of E^n. In particular, the vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ constructed using the OLS estimator $\hat{\boldsymbol{\beta}}$ belongs to this subspace.

The estimator $\hat{\boldsymbol{\beta}}$ was obtained by solving (7):

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}, \tag{31}$$

and these equations have a simple geometrical interpretation. Note first that each element of the LHS of (31) is a scalar product. By the rule for selecting a single row of a matrix product, the i-th element is

$$\mathbf{x}_i^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \left\langle \mathbf{x}_i, \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right\rangle, \tag{32}$$

since $\mathbf{x}_i$, the i-th column of $\mathbf{X}$, is the transpose of the i-th row of $\mathbf{X}^\top$. By (31), the scalar product in (32) is zero, and so the vector $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is orthogonal to all of the regressors, that is, all of the vectors $\mathbf{x}_i$ that represent the explanatory variables in this regression. For this reason, equations like (31) are often referred to as orthogonality conditions.

Recall that the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, treated as a function of $\boldsymbol{\beta}$, is called the vector of residuals. This vector may be written as $\mathbf{u}(\boldsymbol{\beta})$. We are interested in $\mathbf{u}(\hat{\boldsymbol{\beta}})$, the vector of residuals evaluated at $\hat{\boldsymbol{\beta}}$, which is often called the vector of least squares residuals or error vector and is usually written as $\hat{\mathbf{u}}$. We have just seen, in(32), that $\hat{\mathbf{u}}$ is orthogonal to all the regressors. This implies that $\hat{\mathbf{u}}$ is in fact orthogonal to every vector in $\mathcal{S}(\mathbf{X})$, the span of the regressors. To see this, remember that any element of $\mathcal{S}(\mathbf{X})$ can be written as $\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, with the result that, by (31),

$$\langle \mathbf{X}\boldsymbol{\beta}, \hat{\mathbf{u}} \rangle = (\mathbf{X}\boldsymbol{\beta})^\top \hat{\mathbf{u}} = \boldsymbol{\beta}^\top \mathbf{X}^\top \hat{\mathbf{u}} = \mathbf{0}.$$

The vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ is referred to as the vector of fitted values. Clearly, it lies in $\mathcal{S}(\mathbf{X})$, and, consequently, it must be orthogonal to $\hat{\mathbf{u}}$.

Figure 10: Residuals and Fitted Values

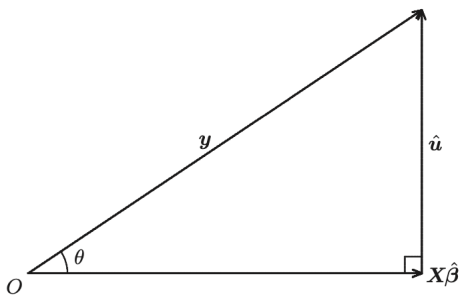
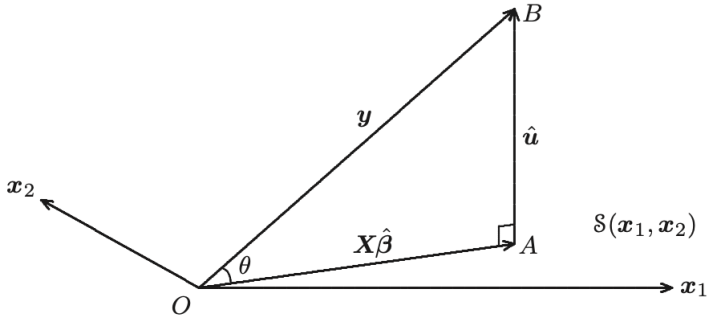


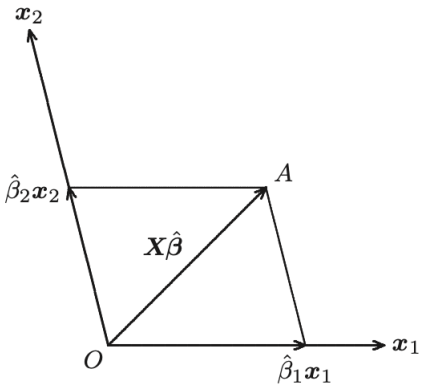
Figure 10 is similar to Figure 9, but it shows the vector of least squares residuals $\hat{\mathbf{u}}$ and the vector of fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$ instead of $\mathbf{u}$ and $\mathbf{X}\boldsymbol{\beta}$. The key feature of this figure, which is a consequence of the orthogonality conditions (31), is that the vector $\hat{\mathbf{u}}$ makes it a right angle with the vector $\mathbf{X}\hat{\boldsymbol{\beta}}$. Now, this is where the magic happens.

Figure 11: Linear Regression in Three Dimensions

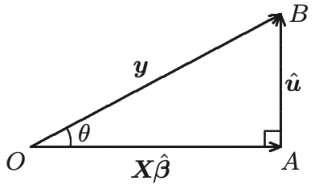
(a) $\mathbf{y}$ Projected on Two Regressors



(b) The Span $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$ of the Regressors



(c) The Vertical Plane through $\mathbf{y}$



Some things about the orthogonality conditions (31) are clearer if we add a third dimension to the picture. Accordingly, in Figure 11a we consider the case of two regressors $\mathbf{x}_1$ and $\mathbf{x}_2$, which together span the horizontal plane labelled $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$, seen in perspective from slightly above the plane. Although the perspective rendering of the figure does not make it clear, both the lengths of $\mathbf{x}_1$ and $\mathbf{x}_2$ and the angle between them are totally arbitrary, since they do not affect $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$ at all. The vector $\mathbf{y}$ is intended to be viewed as rising up out of the plane spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

In the three dimensional setup, it is clear that, if $\hat{\mathbf{u}}$ is to be orthogonal to the horizontal plane, it must itself be vertical. Thus it is obtained by "dropping a perpendicular" from $\mathbf{y}$ to the horizontal plane. The least squares interpretation of the MM estimator $\hat{\boldsymbol{\beta}}$ can now be seen to be a consequence of simple geometry.

The shortest distance from $\mathbf{y}$ to the horizontal plane is obtained by descending vertically to it, and the point in the horizontal plane vertically below $\mathbf{y}$, labelled A in the figure, is the closest point in the plane to $\mathbf{y}$. Thus $\|\hat{\mathbf{u}}\|$ minimises $\|\mathbf{u}(\boldsymbol{\beta})\|$, the norm of $\mathbf{u}(\boldsymbol{\beta})$, with respect to $\boldsymbol{\beta}$.

The squared norm, $\|\mathbf{u}(\boldsymbol{\beta})\|^2$, is just the sum of squared residuals, $\text{SSR}(\boldsymbol{\beta})$. Since minimising the norm of $\mathbf{u}(\boldsymbol{\beta})$ is the same thing as minimising the squared norm, it follows that $\hat{\boldsymbol{\beta}}$ is the OLS estimator.

Figure 11b shows the horizontal plane $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$ as a straightforward two-dimensional picture, seen from directly above. The point A is the point directly underneath $\mathbf{y}$, and so, since $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}$ by definition, the vector represented by the line segment OA is the vector of fitted values, $\mathbf{X}\hat{\boldsymbol{\beta}}$. Geometrically, it is much simpler to represent $\mathbf{X}\hat{\boldsymbol{\beta}}$ than to represent just the vector $\hat{\boldsymbol{\beta}}$, because the latter lies in $\mathbb{R}^k$, a difference space from the space E^n that contains the variables and all linear combinations of them. However, it is easy to see that the information in Figure 11b does indeed determine $\hat{\boldsymbol{\beta}}$. Plainly, $\mathbf{X}\hat{\boldsymbol{\beta}}$ can be decomposed in just one way as a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, as shown. The numerical value of $\hat{\beta}_1$ can be computed as the ratio of the length of vector $\hat{\beta}_1\mathbf{x}_1$ to that of $\mathbf{x}_1$, and similarly for $\hat{\beta}_2$.

In Figure 11c, we show the right-angled triangle that corresponds to dropping a perpendicular from $\mathbf{y}$, labelled in the same way as in Figure 11a. This triangle lies in the vertical plane that contains the vector $\mathbf{y}$. We can see that $\mathbf{y}$ is the hypotenuse of the triangle, the other two sides being $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$. Thus, this figure corresponds to what we saw already in Figure 10. Since we have a right-angled triangle, we can apply Pythagoras' Theorem, which gives

$$\|\mathbf{y}\|^2 = \|\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\hat{\mathbf{u}}\|^2. \tag{33}$$

If we write out the squared norms as scalar products, this becomes

$$\underbrace{\mathbf{y}^\top\mathbf{y}}_{\text{TSS}} = \underbrace{\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}}_{\text{ESS}} + \underbrace{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}_{\text{SSR}}. \tag{34}$$

In words, the total sum of squares (TSS) is equal to the explained sum of squares (ESS) plus the sum of squared residuals (SSR)! This is a fundamental property of OLS estimates, and it will prove

to be very useful in many contexts. Intuitively, it lets us break down the total variation (TSS) of the dependent variable into the explain variation (ESS) and the unexplained variation (SSR), unexpected because the residuals represent the aspects of $\mathbf{y}$ about which we remain in ignorance.

3.3 Orthogonal projections

Before going onto discuss some of the statistical properties of OLS estimation, we should briefly cover orthogonal projects and the Frisch-Waugh-Lovell (FWL) Theorem. Though the notes in the previous section were close to direct quotations from Davidson and MacKinnon (2004), I'll keep this section brief and will be skipping quite a bit material in order to keep things moving along.

OLS estimation maps the regressand $\mathbf{y}$ into a vector of fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Geometrically, these mappings are examples of orthogonal projections. A projection is a mapping that takes each point of E^n into a point in a subspace of E^n, while leaving all points in that subspace unchanged. Because of this, the subspace is called the invariant subspace of the projection. An orthogonal projection maps any point into the point of the subspace that is closest to it. If a point is already in the invariant subspace, it is mapped into itself.

The concept of an orthogonal projection formalises the notion of "dropping a perpendicular" that we mentioned previously. Algebraically, an orthogonal projection onto a given subspace can be performed by premultiplying the vector to be projected by a suitable projection matrix. In the case of OLS, the two main projection matrices that yield the vector of fitted values and the vector of residuals, respectively, are the projection matrix,

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top, \tag{35}$$

and the elimination/annihilator matrix,

$$\mathbf{M}_{\mathbf{X}} = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top, \tag{36}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Recall the formula for the OLS estimates of $\boldsymbol{\beta}$ (6)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

From this, we see that

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{P}_{\mathbf{X}} \mathbf{y}. \tag{37}$$

Therefore, the projection matrix $\mathbf{P}_{\mathbf{X}}$ projects onto $\mathcal{S}(\mathbf{X})$. For any n-vector $\mathbf{y}$, $\mathbf{P}_{\mathbf{X}}\mathbf{y}$ always lies in $\mathcal{S}(\mathbf{X})$, because

$$\mathbf{P}_{\mathbf{X}} \mathbf{y} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

takes the form $\mathbf{Xb}$ for $\mathbf{b} = \hat{\boldsymbol{\beta}}$, it is a linear combination of the columns of $\mathbf{X}$, and hence it belongs to $\mathcal{S}(\mathbf{X})$.

From (35) it is easy to show that $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$. Since any vector in $\mathcal{S}(\mathbf{X})$ can be written as $\mathbf{Xb}$ for some $\mathbf{b} \in \mathbb{R}^k$, we see that

$$\mathbf{P}_{\mathbf{X}} \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{b}. \tag{38}$$

We saw from (37) that the result of acting on any vector $\mathbf{y} \in E^n$ with $\mathbf{P}_{\mathbf{X}}$ is a vector in $\mathcal{S}(\mathbf{X})$. Thus the invariant subspace of the projection $\mathbf{P}_{\mathbf{X}}$ must be contained in $\mathcal{S}(\mathbf{X})$. But by (38), every vector in $\mathcal{S}(\mathbf{X})$ is mapped into itself by $\mathbf{P}_{\mathbf{X}}$. Therefore, the image of $\mathbf{P}_{\mathbf{X}}$, aka its invariance subspace, is precisely $\mathcal{S}(\mathbf{X})$.

When the elimination matrix $\mathbf{M}_{\mathbf{X}}$ is applied to $\mathbf{y}$, it yields the vector of residuals:

$$\begin{aligned}
\mathbf{M}_{\mathbf{X}} \mathbf{y} &= (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y} \\
&= \mathbf{y} - \mathbf{P}_{\mathbf{X}} \mathbf{y} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \hat{\mathbf{u}}.
\end{aligned}$$

The image of $\mathbf{M_X}$ is $\mathcal{S}^{\perp}(\mathbf{X})$, the orthogonal complement of the image of $\mathbf{P_X}$. To see this, consider any vector $\mathbf{w} \in \mathcal{S}^{\perp}(\mathbf{X})$. It must satisfy the defining condition $\mathbf{X}^{\top}\mathbf{w} = \mathbf{0}$. By definition this implies that $\mathbf{P_X w} = \mathbf{0}$, the zero vector. Since $\mathbf{M_X} = \mathbf{I} - \mathbf{P_X}$, we find that $\mathbf{M_X w} = \mathbf{w}$. Thus, $\mathcal{S}^{\perp}(\mathbf{X})$ must be contained in the image of $\mathbf{M_X}$. Next, consider any vector in the image of $\mathbf{M_X}$. It must take the form $\mathbf{M_X y}$, where $\mathbf{y}$ is some vector in E^n. From this, it will follow that $\mathbf{M_X y}$ belongs to $\mathcal{S}^{\perp}(\mathbf{X})$. Observe that

$$(\mathbf{M_X y})^{\top}\mathbf{X} = \mathbf{y}^{\top}\mathbf{M_X X}, \tag{39}$$

an equality that relies on the symmetry of $\mathbf{M_X}$. Then, we have

$$\begin{aligned} \mathbf{M_X X} &= (\mathbf{I} - \mathbf{P_X})\mathbf{X} \\ &= \mathbf{X} - \mathbf{X} \\ &= \mathbf{O}, \end{aligned} \tag{40}$$

where $\mathbf{O}$ denotes a zero matrix (not a vector), which is $n \times k$ in this instance. The result (39) says that any vector $\mathbf{M_X y}$ in the image of $\mathbf{M_X}$ is orthogonal to $\mathbf{X}$, and thus belongs to $\mathcal{S}^{\perp}(\mathbf{X})$. We saw above that $\mathcal{S}^{\perp}(\mathbf{X})$ was contained in the image of $\mathbf{M_X}$, and so this image must coincide with $\mathcal{S}^{\perp}(\mathbf{X})$. For obvious reasons, $\mathbf{M_X}$ is sometimes called the projection **off** $\mathcal{S}(\mathbf{X})$.

A couple things regarding projection matrices. First, as we just saw above, they are symmetric, so

$$\begin{aligned} \mathbf{P_X} &= \mathbf{P_X}^{\top}, \\ \mathbf{M_X} &= \mathbf{M_X}^{\top}. \end{aligned}$$

Secondly, they are idempotent:

$$\begin{aligned} \mathbf{P_X P_X} &= \mathbf{P_X}, \\ \mathbf{M_X M_X} &= \mathbf{M_X}. \end{aligned}$$

The intuition should be somewhat obvious. If you take a point and project it onto, say, $\mathcal{S}(\mathbf{X})$, and then

project it again onto $\mathcal{S}(\mathbf{X})$ again, the second projection doesn't do anything as the point is already in $\mathcal{S}(\mathbf{X})$. Third, the projection and elimination matrices are complementary projections:

$$\mathbf{P_X} + \mathbf{M_X} = \mathbf{I}.$$

This can be proven algebraically very easily, and the intuition is pretty straightforward too. If you apply both projection and elimination matrices to a vector, then the sum product of those projections is the original vector itself. Finally, the fact that $\mathcal{S}(\mathbf{X})$ and $\mathcal{S}^{\perp}(\mathbf{X})$ are orthogonal subspaces implies that the projection and elimination matrices define what is called an orthogonal decomposition of E^n, so we have:

$$\mathbf{P_X M_X} = \mathbf{O}.$$

So, the elimination matrix $\mathbf{M_X}$ annihilates all points that lie in $\mathcal{S}(\mathbf{X})$, and $\mathbf{P_X}$ likewise annihilates all points that lie in $\mathcal{S}^{\perp}(\mathbf{X})$. Consider Figure 7. If we project any point in $\mathcal{S}(\mathbf{X})$ onto $\mathcal{S}^{\perp}(\mathbf{X})$, we end up at the origin, likewise if we project any point in $\mathcal{S}^{\perp}(\mathbf{X})$ onto $\mathcal{S}(\mathbf{X})$.

Provided that $\mathbf{X}$ has full rank, the subspace $\mathcal{S}(\mathbf{X})$ is k-dimensional, and so the first term in the decomposition $\mathbf{y} = \mathbf{P_X y} + \mathbf{M_X y}$ belongs to a k-dimensional space. Since $\mathbf{y}$ itself belongs to E^n, which has n dimensions, it follows that the complementary subspace $\mathcal{S}^{\perp}(\mathbf{X})$ must have $n - k$ dimensions. The number $n - k$ is called the codimension of $\mathbf{X}$ in E^n.

Geometrically, an orthogonal decomposition $\mathbf{y} = \mathbf{P_X y} + \mathbf{M_X y}$ can be represented by a right-angled triangle, with $\mathbf{y}$ as the hypotenuse and $\mathbf{P_X y}$ and $\mathbf{M_X y}$ as the other two sides. In terms of projections, equation (33), which is really just Pythagoras' Theorem, can be rewritten as

$$\|\mathbf{y}\|^2 = \|\mathbf{P_X y}\|^2 + \|\mathbf{M_X y}\|^2.$$

Since every term is nonnegative, we obtain the useful result that, for any orthogonal projection matrix $\mathbf{P_X}$ and any vector $\mathbf{y} \in E^n$,

$$\|\mathbf{P_X y}\| \leq \|\mathbf{y}\|.$$

In effect, this just says that the hypotenuse is longer than either of the other sides of a right-angled triangle.

In general, we will use $\mathbf{P}$ and $\mathbf{M}$ subscripted by matrix expressions to denote the matrices that, respectively, project onto and off subspaces spanned by the columns of those matrix expressions. This will become extremely useful when we have to construct things like test statistics, especially for generalised least squares (GLS) or instrumental variables (IV) estimation.

3.4 The Frisch-Waugh-Lovell (FWL) Theorem

Suppose we have the following regression equation

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \tag{41}$$

where $\mathbf{X}_1$ is $n \times k_1$, $\mathbf{X}_2$ is $n \times k_2$, and $\mathbf{X}$ may be written as a partitioned matrix $\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$, with $k = k_1 + k_2$. We begin by assuming all the regressors in $\mathbf{X}_1$ are orthogonal to all the regressors in $\mathbf{X}_2$, so that $\mathbf{X}_2\mathbf{X}_1 = \mathbf{O}$. Under this assumption, the vector of least squares estimates $\hat{\boldsymbol{\beta}}_1$ is the same the one obtained from the regression

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1, \tag{42}$$

and $\hat{\boldsymbol{\beta}}_2$ from (41) is likewise the same as the vector of estimates obtained from the regression $\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$. In other words, when $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, we can drop either set of regressors from (41) without affecting the coefficients of the other set.

The vector of fitted values from (41) is $\mathbf{P}_\mathbf{X}\mathbf{y}$, while that from (42) is $\mathbf{P}_1\mathbf{y}$, where

$$\mathbf{P}_1 \equiv \mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top.$$

We also have

$$\mathbf{P}_1\mathbf{P}_\mathbf{X} = \mathbf{P}_\mathbf{X}\mathbf{P}_1 = \mathbf{P}_1, \tag{43}$$

which is true whether or not $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal. Thus,

$$\begin{aligned}\mathbf{P}_1\mathbf{y} &= \mathbf{P}_1\mathbf{P}_\mathbf{X}\mathbf{y} \\ &= \mathbf{P}_1(\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) \\ &= \mathbf{P}_1\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 \\ &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1.\end{aligned}$$

We can prove (43), as it's quite straightforward:

$$\begin{aligned}\mathbf{P}_\mathbf{X}\mathbf{P}_1 &= \mathbf{P}_\mathbf{X}\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \\ &= \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \\ &= \mathbf{P}_1.\end{aligned}$$

The middle equality follows by noting that $\mathbf{P}_\mathbf{X}\mathbf{X}_1 = \mathbf{X}_1$, because all the columns of $\mathbf{X}_1$ are in $\mathcal{S}(\mathbf{X})$, and so are left unchanged by $\mathbf{P}_\mathbf{X}$. As for using the property in (43), $\mathbf{P}_1\mathbf{P}_\mathbf{X} = \mathbf{P}_1$, that's obtained directly by transposing $\mathbf{P}_\mathbf{X}\mathbf{P}_1 = \mathbf{P}_1$ and using the symmetry of $\mathbf{P}_\mathbf{X}$ and $\mathbf{P}_1$.

But, how to go about retrieving $\hat{\boldsymbol{\beta}}_1$ or $\hat{\boldsymbol{\beta}}_2$ in the case that we have (41)? This is when we use the elimination matrix, $\mathbf{M}_1 \equiv \mathbf{I} - \mathbf{P}_1$:

$$\begin{aligned}\mathbf{M}_1\mathbf{y} &= \mathbf{M}_1\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\mathbf{u} \\ &= \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\mathbf{u}, \qquad (44) \\ \implies \hat{\boldsymbol{\beta}}_2 &= (\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2^\top)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{y},\end{aligned}$$

which follows many of the properties we just established for the projection matrix $\mathbf{P}_1$. We could apply an analogous operation to attain $\hat{\boldsymbol{\beta}}_1$. This will yield the same vector of OLS estimates $\hat{\boldsymbol{\beta}}_2$ as (41), and also the same vector of residuals. This regression is sometimes called the FWL regression. One thing to take note of is that the difference between $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ is not the same thing as the vector $\mathbf{u}$ in (41). We can now formally state the FWL Theorem.

Theorem (Frisch-Waugh-Lovell):

1. The OLS estimates from regressions (41) and (44) are numerically identical.
2. The residuals from regressions (41) and (44) are numerically identical.

The proof is as follows (and was basically covered above). By the standard formula (25), the estimate from (44) is

$$(\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2^\top)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}.$$

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ denote the two vectors of OLS estimates from (41). Then

$$\mathbf{y} = \mathbf{P_x y} + \mathbf{M_X y} = \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 + \mathbf{M_X y}. \tag{45}$$

Premultiplying by $\mathbf{X}_2^\top \mathbf{M}_1$, we obtain

$$\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} = \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2.$$

This is because $\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 = \mathbf{0}$ and $\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{M_X y} = \mathbf{X}_2^\top \mathbf{M_X y} = \mathbf{0}$. We can these use simple matrix algebra to show

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2^\top)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y},$$

which is what we showed before. The proves the first part of the theorem.

If we had premultiplied (45) by $\mathbf{M}_1$ instead of $\mathbf{X}_2^\top \mathbf{M}_1$, we would have obtained

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \hat{\beta} + \mathbf{M_X y}, \tag{46}$$

where the last term is unchanged from (45) because $\mathbf{M_1 M_X} = \mathbf{M_X}$. The regressand in (46) is the regressand from regression (44). Because $\hat{\boldsymbol{\beta}}_2$ is the estimate of $\boldsymbol{\beta}_2$ from (44), by the first part of the theorem, the first term on the RHS of (46) is the vector of fitted values from that regression. Thus, the second term must be the vector of residuals from regression (44). But $\mathbf{M_X y}$ is also the vector of residuals from regression (41), and this therefore proves the second part of the theorem. ■

3.4.1 Applications of FWL Theorem: R^2

We showed before that the TSS in the regression model can be expressed as the sum of the ESS and SSR. This was just an application of Pythagoras' Theorem. In terms of the orthogonal projection matrices $\mathbf{P_X}$ and $\mathbf{M_X}$, the relation between TSS, ESS, and SSR can be written as

$$\text{TSS} = \|\mathbf{y}\|^2 = \|\mathbf{P_X y}\|^2 + \|\mathbf{M_X y}\|^2 = \text{ESS} + \text{SSR}.$$

This allows us to write the goodness of fit, or the uncentered R^2 as

$$R_u^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\|\mathbf{P_X y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\|\mathbf{M_X y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\text{SSR}}{\text{TSS}} = \cos^2\theta,$$

where θ is the angle between $\mathbf{y}$ and $\mathbf{P_X y}$ (see Figure 10). For any angle θ, we know that $-1 \leq \cos\theta \leq 1$, and consequently $0 \leq R_u^2 \leq 1$. If the angle θ were zero, $\mathbf{y}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$ would coincide, the residual vector $\hat{\mathbf{u}}$ would vanish, and we would have a perfect fit, $R_u^2 = 1$. At the other extreme, $R_u^2 = 0$, the fitted value vector would vanish, and $\mathbf{y}$ would coincide with the residual vector $\hat{\mathbf{u}}$.

We also have the centred R^2, R_c^2, which we use because R_u^2 varies due to changes in units which affect the angle θ. An example of such a change is given by the conversion between Celsius and Fahrenheit scales of temperature, where a constant is involved. TO see this, let us consider a very simple change of measuring units, whereby a constant α, analogous to the constant 32 used in converting from Celsius to Fahrenheit, is added to each element of $\mathbf{y}$. In terms of these new units, the regression of $\mathbf{y}$ on a regressor matrix $\mathbf{X}$ becomes

$$\mathbf{y} + \alpha\boldsymbol{\iota} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

If we assume that the matrix $\mathbf{X}$ includes a constant, it follows that $\mathbf{P_X}\boldsymbol{\iota} = \boldsymbol{\iota}$ and $\mathbf{M_X}\boldsymbol{\iota} = \mathbf{0}$, and so we find that

$$\mathbf{y} + \alpha\boldsymbol{\iota} = \mathbf{P_X}(\mathbf{y} - \alpha\boldsymbol{\iota}) + \mathbf{M_X}(\mathbf{y} + \alpha\boldsymbol{\iota}) = \mathbf{P_X y} - \alpha\boldsymbol{\iota} + \mathbf{M_X y}.$$

This allows us to compute R_u^2 as

$$R_u^2 = \frac{\|\mathbf{P_X y} + \alpha\boldsymbol{\iota}\|^2}{\|\mathbf{y} + \alpha\boldsymbol{\iota}\|^2},$$

which is clearly different from the expression for R_u^2 we had before. By choosing α sufficient large, we can in fact make R_u^2 as close as we wish to 1, because, for a very large α, the term $\alpha\boldsymbol{\iota}$ will completely dominate the terms $\mathbf{P_X y}$ and $\mathbf{y}$ in the numerator and denominator, respectively. But a large R_u^2 in that case would be entirely misleading.

But we can get around this problem (for regressions which include a constant term) fairly easily. An elementary consequence of the FWL Theorem is that we can express all variables as deviations from their means, bu the operation of the elimination matrix $\mathbf{M}_\iota$. The ordinary R^2 from the regression that uses centred variables is

$$R_c^2 \equiv \frac{\|\mathbf{P_X M}_\iota \mathbf{y}\|^2}{\|\mathbf{M}_\iota \mathbf{y}\|^2} = 1 - \frac{\|\mathbf{M_X y}\|^2}{\|\mathbf{M}_\iota \mathbf{y}\|^2}.$$

This is but a small preview of the applications of the FWL Theorem. We will use it again when we look at the derivation of test statistics.

4 Statistical Properties of OLS

In this section we now review the statistical properties of the OLS estimator. We should make an important distinction upfront though: we need to separate the discussion between exact/finite-sample and asymptotic/large-sample statistical properties. What does this mean? Suppose are studying the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{47}$$

we may wish to assume that the data were actually generated by the DGP

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}, \quad \mathbf{u} \sim \text{NID}(\mathbf{0}, \sigma_0^2\mathbf{I}). \tag{48}$$

Notice in (47) we use the abbreviation "IID", which means "independently and identically distributed." The notation $\text{IID}(\mathbf{0}, \sigma^2\mathbf{I})$ means that the u_t are statistically independent and all follow the same

distribution, with mean 0 and variance σ^2. Similarly, in the DGP (48), the notation NID($\mathbf{0}, \sigma_0^2\mathbf{I}$) means that the u_t are normally, independently, and identically distributed, with mean 0 and variance σ_0^2. In both cases, it is being assumed that the distribution $\mathbf{u}$ is no way dependent on $\mathbf{X}$. In (48) we must provide specific values for the parameters $\boldsymbol{\beta}$ and σ^2 (hence we use the zero subscripts), and we must specify from what distribution the error terms are to be drawn (here, the normal distribution). In the linear regression model (47), $\boldsymbol{\beta} \in \mathbb{R}^k$, $\sigma^2 \in \mathbb{R}_{++}$, and $\mathbf{u}$ follows a distribution over all possible distributions that have mean 0 and variance σ^2. The classical normal linear model (48) evidently belongs to this set, but the restrictions placed on it are more restrictive than in (47).

As we will soon see, for finite-sample properties, we will be working with the classical normal linear regression model (48), whereby with a limited sample of size n we place some strict assumptions on the distribution of $\mathbf{u}$. With sufficiently strict restrictions, we can then derive finite-sample distributions of the OLS estimator itself, as well as test statistics used for hypothesis testing. But, as you can imagine, rarely in economics can such strict model assumptions lead to accurate inferences based on the sample data collected. As such, we need to move beyond exact/finite-sample inference and rely on asymptotic/large-sample theory.

We first begin with finite-sample properties of the OLS estimator.

4.1 Finite-sample properties

Having derived the OLS estimator in Section 2, we now examine its finite-sample properties, namely the characteristics of the distribution of the estimator that are valid for any given sample size n.

Proposition (finite-sample properties of the OLS estimator):

1. The OLS estimator is unbiased: Assuming that the model is linear, that $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ (strict exogeneity, i.e., $\mathbb{E}[u_i|\mathbf{X}] = 0, \ \forall i$), and that $\mathbf{X}$ is of full rank (no multicollinearity), we have that $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ because we assume that the model is correctly specified.

    - Proof: $\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] = \mathbf{0}$ if and only if $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$. We prove the former. By the expression

for the sampling error, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - \boldsymbol{\beta} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u} - \boldsymbol{\beta} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u}, \end{aligned} \tag{49}$$

but then

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u}|\mathbf{X}] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\underbrace{\mathbb{E}[\mathbf{u}|\mathbf{X}]}_{=\mathbf{0}} \\ &= \mathbf{0}. \end{aligned}$$

Here the second equality holds by the linearity of conditional expectations; $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a function of $\mathbf{X}$ (obviously) and so it can be treated as if nonrandom. ■

2. The OLS estimator has variance $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$: Assuming that the model is linear, that $\mathbf{X}$ is strictly exogenous, that $\mathbf{X}$ has no multicollinearity, and that the errors are homoskedastic in their variance, $\text{Var}(\mathbf{u}|\mathbf{X}) = \mathbb{E}[\mathbf{u}\mathbf{u}^\top|\mathbf{X}] = \sigma^2\mathbf{I}$, or equivalently, $\mathbb{E}[u_t^2|\mathbf{X}_t] = \sigma^2 > 0,\ \forall t = 1, ..., n$.

   - Proof:

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \text{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}) \\ &= \text{Var}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u}|\mathbf{X}) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Var}(\mathbf{u}|\mathbf{X})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}. \end{aligned} \tag{50}$$

The first line makes use of the fact that $\boldsymbol{\beta}$ is a constant, the second line is simply using our definition of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ from our proof of unbiasedness, the third makes use of the fact that the term $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a function of $\mathbf{X}$ and we have to pre- and post-multiply because we take the term out of the Var operator, and then the fourth and fifth lines follow pretty intuitively. ■

3. **Gauss-Markov Theorem**: Assuming that the model is linear, that $\mathbf{X}$ is strictly exogenous, that $\mathbf{X}$ has no multicollinearity, and that the errors are conditionally homoskedastic in their variance, the OLS estimator is efficient in the class of linear unbiased estimators. That is, for any unbiased estimator $\tilde{\boldsymbol{\beta}}$ that is linear in $\mathbf{y}$, $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) \geq \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ in the matrix sense.[4]

   - Proof: Since $\tilde{\boldsymbol{\beta}}$ is linear in $\mathbf{y}$, it can be written as $\tilde{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ for some matrix $\mathbf{C}$, which is possibly a function of $\mathbf{X}$. Let $\mathbf{D} \equiv \mathbf{C} - \mathbf{A}$ or $\mathbf{C} = \mathbf{D} + \mathbf{A}$ where $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Then

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \mathbf{C}\mathbf{y} = (\mathbf{D} + \mathbf{A})\mathbf{y} \\
&= \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y} \\
&= \mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) + \hat{\boldsymbol{\beta}} \\
&= \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \hat{\boldsymbol{\beta}},
\end{aligned}$$

     where the third line follows from the fact that the correctly specified model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ and $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Taking the conditional expectations of both sides yields

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}] &= \mathbb{E}[\mathbf{D}\mathbf{X}\boldsymbol{\beta}|\mathbf{X}] + \mathbb{E}[\mathbf{D}\mathbf{u}|\mathbf{X}] + \mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] \\
&= \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbb{E}[\mathbf{D}\mathbf{u}|\mathbf{X}] + \mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] \\
&= \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbb{E}[\mathbf{u}|\mathbf{X}] + \boldsymbol{\beta} \\
&= \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}.
\end{aligned}$$

     But we assumed that $\tilde{\boldsymbol{\beta}}$ is unbiased, so somehow we need $\mathbf{D}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, which can only hold if

[4] Let $\mathbf{A}$ and $\mathbf{B}$ be two square matrices of the same size. We say that $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semidefinite. A $k \times k$ matrix $\mathbf{C}$ is said to be positive semidefinite (or nonnegative definite) if $\mathbf{x}^\top\mathbf{C}\mathbf{x} \geq 0$ for all k-dimensional vectors $\mathbf{x}$.

$\mathbf{DX} = \mathbf{0}$ since we know $\boldsymbol{\beta} \neq \mathbf{0}$. We will come back to this, but for now we can state:[5]

$$\tilde{\boldsymbol{\beta}} = \mathbf{Du} + \hat{\boldsymbol{\beta}},$$

and then if we subtract $\boldsymbol{\beta}$ from the both the LHS and RHS:

$$\begin{aligned}\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \mathbf{Du} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ &= \mathbf{Du} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u} \\ &= (\mathbf{D} + \mathbf{A})\mathbf{u}.\end{aligned}$$

So it then follows that

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \text{Var}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}) \\ &= \text{Var}\left((\mathbf{D} + \mathbf{A})\mathbf{u}|\mathbf{X}\right) \\ &= (\mathbf{D} + \mathbf{A})\text{Var}(\mathbf{u}|\mathbf{X})(\mathbf{D} + \mathbf{A})^\top \\ &= \sigma^2(\mathbf{D} + \mathbf{A})(\mathbf{D} + \mathbf{A})^\top \\ &= \sigma^2(\mathbf{DD}^\top + \mathbf{DA}^\top + \mathbf{AD}^\top + \mathbf{AA}^\top),\end{aligned}$$

where we can pull $(\mathbf{D} + \mathbf{A})$ out of the conditional $\text{Var}(\cdot)$ operator since both $\mathbf{D}$ and $\mathbf{A}$ are functions of $\mathbf{X}$. But notice that we have $\mathbf{DA}^\top = \mathbf{DX}(\mathbf{X}^\top\mathbf{X})^{-1}$ and we conjectured that $\mathbf{DX} = \mathbf{0}$. So presumably $\mathbf{DA}^\top = \mathbf{0}$. Furthermore, we have $\mathbf{AA}^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1}$. So,

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{DD}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}),$$

[5]Note that

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbf{Du} + \hat{\boldsymbol{\beta}} \\ &= \mathbf{Du} + \mathbf{Ay} \\ &= \mathbf{Du} + \mathbf{AX}\boldsymbol{\beta} + \mathbf{Au} \\ &= (\mathbf{D} + \mathbf{A})\mathbf{u} + \boldsymbol{\beta}.\end{aligned}$$

where, recall, $\mathbf{D}\mathbf{D}^\top$ is a positive semidefinite matrix. Thus,

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}) \geq \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}). \quad \blacksquare$$

- The implication here is that $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semidefinite. So

$$\mathbf{a}^\top \left[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})\right] \mathbf{a} \geq 0,$$

for any k dimensional vector $\mathbf{a}$. In particular, consider a special vector whose elements are all 0 except for the j-th element, which is 1. For this particular $\mathbf{a}$, the quadratic form $\mathbf{a}^\top\mathbf{A}\mathbf{a}$ picks up the (j,j) element of $\mathbf{A}$. But the (j,j) element of $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$, for example, is $\text{Var}(\tilde{\beta}_j|\mathbf{X})$ where $\tilde{\beta}_j$ is the j-th element of $\tilde{\boldsymbol{\beta}}$. Thus, the matrix inequality in our proof implies

$$\text{Var}(\tilde{\beta}_j|\mathbf{X}) \geq \text{Var}(\hat{\beta}_j|\mathbf{X}), \ j = 1, ..., k.$$

That is, for any regression coefficient, the variance of the OLS estimator is no larger than that of any other linear unbiased estimator.

4. Assuming that the model is linear, that $\mathbf{X}$ is strictly exogenous, that $\mathbf{X}$ has no multicollinearity, and that the errors are conditionally homoskedastic in their variance, $\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}|\mathbf{X}) = \mathbf{0}$.

- Proof: By definition, we have

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}|\mathbf{X}) &\equiv \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right)\left(\hat{\mathbf{u}} - \mathbb{E}[\hat{\mathbf{u}}|\mathbf{X}]\right)^\top |\mathbf{X}\right] \\ &= \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\mathbf{u}} - \mathbb{E}[\hat{\mathbf{u}}|\mathbf{X}]\right)^\top |\mathbf{X}\right],\end{aligned}$$

now, before proceeding, note that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}$ and

$$\begin{aligned}
\hat{\mathbf{u}} - \mathbb{E}[\hat{\mathbf{u}}|\mathbf{X}] &= \mathbf{M}_{\mathbf{X}} \mathbf{y} - \mathbb{E}[\mathbf{M}_{\mathbf{X}} \mathbf{y}|\mathbf{X}] \\
&= \mathbf{M}_{\mathbf{X}} \mathbf{y} - \mathbf{M}_{\mathbf{X}} \mathbb{E}[\mathbf{y}|\mathbf{X}] \\
&= \mathbf{M}_{\mathbf{X}} (\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{X}]) \\
&= \mathbf{M}_{\mathbf{X}} (\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}|\mathbf{X}]) \\
&= \mathbf{M}_{\mathbf{X}} (\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{M}_{\mathbf{X}} \mathbf{u},
\end{aligned}$$

so we can write

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}}|\mathbf{X}) &= \mathbb{E}\left[\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \right) (\mathbf{M}_{\mathbf{X}} \mathbf{u})^\top |\mathbf{X} \right] \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}\left[\mathbf{u}\mathbf{u}^\top |\mathbf{X} \right] \mathbf{M}_{\mathbf{X}} \\
&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_{\mathbf{X}} \\
&= \mathbf{0},
\end{aligned}$$

where we pull $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ and $\mathbf{M}_{\mathbf{X}}$ out of the conditional $\mathbb{E}[\cdot]$ operator as they're both functions of $\mathbf{X}$, and we make use of the properties of elimination matrix $\mathbf{M}_{\mathbf{X}}$. ■

As is clear from (25), the OLS estimator is linear in $\mathbf{y}$. There are many other estimators of $\boldsymbol{\beta}$ that are linear and unbiased. The Gauss-Markov Theorem says that the OLS estimator is efficient in the sense that its conditional variance-covariance matrix is smallest among linear unbiased estimators. For this reason, the OLS estimator is called the best linear unbiased estimator (BLUE).

The OLS estimator $\hat{\boldsymbol{\beta}}$ is a function of the sample $(\mathbf{y}, \mathbf{X})$. Since $(\mathbf{y}, \mathbf{X})$ are random, so is $\hat{\boldsymbol{\beta}}$. Now imagine that we fix $\mathbf{X}$ at some given value, calculate $\hat{\boldsymbol{\beta}}$ for all samples corresponding to all possible realisations of $\mathbf{y}$, and take the average of $\hat{\boldsymbol{\beta}}$ (think of a Monte Carlo exercise). This average is the (population) conditional $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]$. Our first proposition (1) says that this average equals the true value $\boldsymbol{\beta}$.

There is another notion of unbiasedness that is weaker than the unbiasedness in part (1). By the LIE, we have $\mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]] = \mathbb{E}[\hat{\boldsymbol{\beta}}]$. So (1) implies

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

This says: if we calculated $\hat{\boldsymbol{\beta}}$ for all possible different samples, differing not only in $\mathbf{y}$ but also in $\mathbf{X}$, the average would be the true value. This unconditional statement is probably more relevant in economics because samples do differ in both $\mathbf{y}$ and $\mathbf{X}$. The importance of the conditional statement (1) is that it implies the unconditional statement, which is more relevant. The same holds for the conditional statement (3) about the variance.

Proposition: Assuming that the model is linear, that $\mathbf{X}$ is strictly exogenous, that $\mathbf{X}$ has no multicollinearity, and that the errors are homoskedastic in their variance, $\mathbb{E}[s^2|\mathbf{X}] = \sigma^2$ (and hence $\mathbb{E}[s^2] = \sigma^2$), provided that $n > k$. In other words, the OLS estimator of σ^2, s^2, is unbiased.

Proof: We can prove this proposition easily by the use of the trace operator.[6] Since

$$s^2 = \frac{\hat{\mathbf{u}}^\top \hat{\mathbf{u}}}{n-k}, \tag{51}$$

the proof amounts to showing that $\mathbb{E}[\hat{\mathbf{u}}^\top \hat{\mathbf{u}}|\mathbf{X}] = (n-k)\sigma^2$. As we showed

$$\hat{\mathbf{u}} = \mathbf{M}_{\mathbf{X}}\mathbf{u},$$

so we have

$$\text{SSR} = \hat{\mathbf{u}}^\top \hat{\mathbf{u}} = \mathbf{u}^\top \mathbf{M}_{\mathbf{X}} \mathbf{u}.$$

The proof consists of proving two properties: i) $\mathbb{E}[\mathbf{u}^\top \mathbf{M}_{\mathbf{X}} \mathbf{u}|\mathbf{X}] = \sigma^2 \cdot \text{trace}(\mathbf{M}_{\mathbf{X}})$, and ii) $\text{trace}(\mathbf{M}_{\mathbf{X}}) = n - k$:

- Since

$$\mathbf{u}\mathbf{M}_{\mathbf{X}}\mathbf{u} = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} u_i u_j,$$

[6] The trace of a square matrix $\mathbf{A}$ is the sum of the diagonal elements of $\mathbf{A}$: $\text{trace}(\mathbf{A}) = \sum_i a_{ii}$.

we have

$$\begin{aligned}\mathbb{E}[\mathbf{u}^\top \mathbf{M}_\mathbf{X} \mathbf{u}|\mathbf{X}] &= \sum_{i=1}^n \sum_{j=1}^n m_{ij} \mathbb{E}[u_i u_j|\mathbf{X}] \\ &= \sum_{i=1}^n m_{ii}\sigma^2 \\ &= \sigma^2 \sum_{i=1}^n m_{ii} \\ &= \sigma^2 \cdot \text{trace}(\mathbf{M}_\mathbf{X}).\end{aligned}$$

In the first line, we moved m_{ij} out of the conditional $\mathbb{E}[\cdot]$ operator as its a function of $\mathbf{X}$, and in the second line we use the homogeneous error assumption ($\mathbb{E}[u_i u_j|\mathbf{X}] = 0$ for $i \neq j$).

- Next,

$$\begin{aligned}\text{trace}(\mathbf{M}_\mathbf{X}) &= \text{trace}(\mathbf{I}_n - \mathbf{P}_\mathbf{X}) \\ &= \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{P}_\mathbf{X}) \\ &= n - \text{trace}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top) \\ &= n - \text{trace}((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}) \\ &= n - \text{trace}(\mathbf{I}_k) \\ &= n - k,\end{aligned}$$

where the third line comes from the fact that $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$. ■

Thus, if s^2 is the estimate of σ^2, a natural estimate of $\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}|\mathbf{X}) \equiv s^2(\mathbf{X}^\top \mathbf{X})^{-1}. \tag{52}$$

4.1.1 Hypothesis testing under normality

Suppose we have some theory which suggests $\beta_2 = 1$. We of course assume that due to unbiasedness, on average, $\hat{\beta}_2 = 1$ if the restriction is true. If we collect sample data, then of course it may be the case that based on our particular example $\hat{\beta}_2 \neq 1$. This obviously doesn't mean that we can rule out our restriction or assumption that $\beta_2 = 1$.

In order for us to decide whether the sampling error $\hat{\beta}_2 - \beta_2$ is "too large" for the restriction to be true, we need to construct from the sampling error some test statistic whose probability distribution is known given the truth of the hypothesis. It might appear that doing so requires one to specify the joint distribution of $(\mathbf{X}, \mathbf{u})$ because, as is clear from (49), the sampling error is a function of $\mathbf{X}$ and $\mathbf{u}$. A surprising fact about the theory of hypothesis testing, which will discuss, is that the distribution can be derived without specifying the joint distribution when the distribution of $\mathbf{u}$ conditional on $\mathbf{X}$ is normal; there is no need to specify the distribution of $\mathbf{X}$.

In the language of hypothesis testing, the restriction to be tested (such as $\beta_2 = 1$) is called the null hypothesis. It is a restriction on "the model", a set of assumptions which, combined with the null, produces some test statistic with a known distribution. For the present case of testing the null of regression coefficients, only the normality assumption about the conditional distribution of $\mathbf{u}$ needs to be added to the classical regression model.[7] We say that the model is correctly specified if the maintained hypothesis is true.

The Central Limit Theorem (CLT) suggests that the error term is normally distributed, which leads us to the normality assumption:

Assumption: We assume normality of the error term, i.e., $\mathbf{u}|\mathbf{X} \sim \text{NID}(\mathbf{0}, \sigma^2\mathbf{I})$. Recall from probability theory that the normal distribution has several convenient properties:

- The distribution only depends on the mean and variance. But more importantly, if the conditional distribution on $\mathbf{X}$ is normal, the mean and the variance can depend on $\mathbf{X}$. It follows that, if the distribution condition on $\mathbf{X}$ is normal and if neither the conditional mean nor the conditional variance depends on $\mathbf{X}$, then the marginal (i.e., unconditional) distribution is the same normal

[7] The assumptions that the model is linear, that $\mathbf{X}$ is strictly exogenous, that $\mathbf{X}$ has no multicollinearity, that the errors are homoskedastic in their variance, and that (48) correctly specifies the model.

distribution.

- In general, if two random variables are independent, then they are uncorrelated, but the converse is not true. However, if two random variables are joint normal, the converse is also true, so that independence and a lack of correlation are equivalent. This carries over to conditional distributions: if two random variables are joint normal and uncorrelated conditional on $\mathbf{X}$, then they are independent on $\mathbf{X}$.
- A linear function of random variables that are jointly normally distributed is itself normally distributed. This also carries over to conditional distributions. If the distribution of $\mathbf{u}$ conditional on $\mathbf{X}$ is normal, then $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{u}$ is normal conditional on $\mathbf{X}$.

Thus, since the distribution of $\mathbf{u}$ conditional on $\mathbf{X}$ does not depend on $\mathbf{X}$, it then follows that $\mathbf{u}$ and $\mathbf{X}$ are independent. Therefore, in particular, the marginal or unconditional distribution of $\mathbf{u}$ is

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{53}$$

We know that the sampling error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is linear in $\mathbf{u}$ given $\mathbf{X}$. Since $\mathbf{u}$ is normal given $\mathbf{X}$, so is the sampling error. We already derived its mean and variance, and so under our assumptions we can write:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|\mathbf{X} \sim N\left(\mathbf{0}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right). \tag{54}$$

4.1.2 Testing a single restriction with known variance

Let's work through the following example. Suppose that our null is:

$$H_0 : \beta_k = \beta_0,$$

where β_0 is some specified value of β_k. We wish to test this null against the alternative hypothesis $H_1 : \beta_k \neq \beta_0$, at a significance level of α. Looking at the k-th component of $\hat{\boldsymbol{\beta}}$ and imposing the

restriction of the null, we obtain

$$(\hat{\beta}_k - \beta_0)|\mathbf{X} \sim N\left(0, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}\right),$$

where $(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}$ is the (k,k) element of the inverse of the precision matrix, $(\mathbf{X}^\top\mathbf{X})^{-1}$. So if we define the ratio z_k by dividing $\hat{\beta}_k - \beta_0$ by its standard deviation, we can then write

$$z_k \equiv \frac{\hat{\beta}_k - \beta_0}{\sqrt{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}}} \sim N(0,1) \tag{55}$$

which basically says that the test statistic z_k follows the standard normal distribution.

This test statistic is really neat, but somewhat of a fantasy. First, it assumes that σ^2 is known, which a big assumption. Second, its distribution conditional on $\mathbf{X}$ does not depend on $\mathbf{X}$ (which should not be confused with the fact that the value of z_k depends on $\mathbf{X}$). So z_k and $\mathbf{X}$ are independently distributed, and, regardless of the value of $\mathbf{X}$, the distribution of z_k is the same as its unconditional distribution. Third, the distribution is known and it does not depend on unknown parameters (such as $\boldsymbol{\beta}$).[8]

4.1.3 Testing a single restriction with unknown variance

If we do not know the true value of σ^2, a natural idea is to replace the nuisance parameter, σ^2, by its OLS estimate, s^2. The statistic after the substitution of s^2 for σ^2 is called the t-statistic or the t-value. The denominator of this statistic is called the standard error of the OLS estimate of β_k and is sometimes written as $\text{SE}(\beta_k)$:

$$\text{SE}(\beta_k) \equiv \sqrt{s^2(\mathbf{X}^\top\mathbf{X})^{-1}_{kk}} = \sqrt{\widehat{\text{Var}}(\boldsymbol{\beta}|\mathbf{X})_{kk}}.$$

[8] If the distribution of a statistic depends on unknown parameters, those parameters are called nuisance parameters.

If our model assumptions continue to hold, then under the null hypothesis, the t-statistic is

$$t_k \equiv \frac{\hat{\beta}_k - \beta_0}{\text{SE}(\hat{\beta}_k)} = \frac{\hat{\beta}_k - \beta_0}{\sqrt{s^2(\mathbf{X}^\top \mathbf{X})^{-1}_{kk}}} \sim t(n-k), \tag{56}$$

where $t(n-k)$ is the Student t-distribution with $n-k$ degrees of freedom. I won't cover the proof here, but I have some notes elsewhere based on Davidson and MacKinnon (2004) that do cover the proof.

The decision rule or process for a t-test is as follows:

1. Given the hypothesised value, β_0, of β_k, form the t-statistic as in (56). Too large a deviation of t_k from 0 is a sign of the failure of the null hypothesis. The next step specifies how large is too large.

2. Go to a t-table and look up the entry for $n-k$ degrees of freedom. Find the critical value, $t_{\alpha/2}(n-k)$, such that the area in the t-distribution to the right of $t_{\alpha/2}(n-k)$ is $\alpha/2$. If $n-k=30$ and $\alpha=0.05$, for example, $t_{\alpha/2}(n-k)=2.042$. Then, since the t distribution is symmetric around 0, one can construct a $(1-\alpha)\%$ exact confidence interval for the test:

$$\Pr\left(t_{\alpha/2}(n-k) \leq t_{\beta_i} \leq t_{1-\alpha/2}(n-k)\right) = 1-\alpha.$$

Substitute our test statistic, and our significance level, $\alpha = 0.05$, say, and rearrange:

$$\begin{aligned}
\Pr\left(t_{0.025} \leq \frac{\hat{\beta}_k - \beta_k}{SE_k} \leq t_{0.975}\right) &= 0.95 \\
\Pr\left(t_{0.025} SE_k \leq \hat{\beta}_k - \beta_k \leq t_{0.975} SE_k\right) &= 0.95 \\
\Pr\left(t_{0.025} SE_k - \hat{\beta}_k \leq -\beta_k \leq t_{0.975} SE_k - \hat{\beta}_k\right) &= 0.95 \\
\Pr\left(\hat{\beta}_k - t_{0.025} SE_k \leq \beta_k \leq \hat{\beta}_k - t_{0.975} SE_k\right) &= 0.95.
\end{aligned}$$

which gives the – in this case – 95% confidence interval,

$$\left[\hat{\beta}_k - t_{0.025}SE_k, \hat{\beta}_k - t_{0.975}SE_k\right].$$

3. Accept H_0 if $-t_{\alpha/2}(n-k) < t_k < t_{\alpha/2}(n-k)$. That is, if $|t_k| < t_{\alpha/2}(n-k)$. Reject H_0 otherwise. Since $t_k \sim t(n-k)$ under H_0, the probability of rejecting H_0 when H_0 is true is α. So the size (significance level) of the test is indeed α.

A convenient feature of the t-test is that the critical value does not depend on $\mathbf{X}$, so there is no need to calculate the critical values for each sample.

4.1.4 Tests of several restrictions

Here is where things get a bit trick in terms of notation. I'll try show both a "conventional" approach to setting up an F-test, as well as an approach which makes use of orthogonal projections. I personally think the projection method is more intuitive, but it isn't as common in textbooks.

Suppose we want to test multiple restrictions, r, where $r \leq k$, since there cannot be more equality restrictions than there are parameters in the unrestricted model. Let's assume that the unrestricted model, based on the alternative hypothesis, is

$$H_1 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},\ \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{57}$$

where $\mathbf{X}_1$ is an $n \times k_1$ matrix, $\mathbf{X}_2$ is an $n \times k_2$ matrix, $\boldsymbol{\beta}_1$ is a k_1-vector, $\boldsymbol{\beta}_2$ is a k_2-vector, and $k = k_1 + k_2$. Now suppose that the restriction we wish to apply is that $\boldsymbol{\beta}_2 = \mathbf{0}$, which means that $r = k_2$, leading us to write the null as:

$$H_0 : \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u},\ \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{58}$$

$$\Leftrightarrow H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \tag{59}$$

where $\mathbf{R}$ is $r \times k$ and $\mathbf{r}$ is an r-vector, and whose values are known and specified by the hypothesis.

The two ways of writing the null are equivalent. One thing to note when setting up the null is to make sure that there are no redundant equations and that the equations are consistent with each other. This means we require that $\text{rank}(\mathbf{R}) = r$, i.e., $\mathbf{R}$ is of full row rank and its rank equaling the number of rows.

Let's work through a quick example just show highlight that the two ways of writing the null are indeed the same. Suppose we have the following model:

$$y_t = \beta_1 + X_{2,t}\beta_2 + X_{3,t}\beta_3 + X_{4,t}\beta_4 + X_{5,t}\beta_5 + u_t,$$

and suppose that we want to test two restrictions (so $r = 2$):

$$\beta_3 = \beta_4,$$
$$\beta_5 = 0.$$

In the format of (58), we could simply lump the constant term and $X_{2,t}$ together and then stack to form $\mathbf{X}_1 = \begin{bmatrix} \boldsymbol{\iota} & \mathbf{x}_2 \end{bmatrix}$, which means that $\mathbf{X}_2$ consists of the remaining regressors. Clearly, we would then have $\boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_3 & \beta_4 & \beta_5 \end{bmatrix}$. But in the format of (59), we would have:

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \ \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Just a quick example of a redundant equation, causing a failure of the rank condition: suppose we additionally required

$$\beta_3 - \beta_4 = \beta_5.$$

This is redundant because it holds whenever the first two equations do. With these three equations,

$r = 3$, and

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & -1 \end{bmatrix}, \ \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Since the third row of $\mathbf{R}$ is the difference between the first two, $\mathbf{R}$ is not of full row rank. The consequence of adding redundant equations is that $\mathbf{R}$ no longer meets the full row rank condition.

Now, let's derive the F-test statistic. I will first follow the "conventional" notation of (59).

Proposition: If we assume that the model is linear, $\mathbf{X}$ is strictly exogenous, $\mathbf{X}$ has no multicollinearity, the errors are homoskedastic in their variance, and model is correctly specified (normal distribution of errors), then under the null $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, the F-test statistic is defined as

$$\begin{aligned} F &\equiv \frac{\dfrac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{r}}{\dfrac{s^2}{}} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\mathbf{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}|\mathbf{X})\mathbf{R}^\top\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{r} \sim F(r, n-k), \end{aligned} \tag{60}$$

where $F(r, n-k)$ denotes the F distribution with $(r, n-k)$ degrees of freedom. As in our proposition for the t-statistic, it suffices to show that the distribution conditional on $\mathbf{X}$ is $F(r, n-k)$; because the F distribution does not depend on $\mathbf{X}$, it is also the unconditional distribution of the statistic. Again, I won't cover the proof for this here, but Hayashi (2000) has a really good concise proof.

Now, let's derive the F-statistic using orthogonal projections and the format of (58). Let us state that USSR denotes the unrestricted sum of squared residuals, and RSSR denotes the restricted sum of squared residuals, so we can write

$$F_{\boldsymbol{\beta}_2} \equiv \frac{\dfrac{\text{RSSR} - \text{USSR}}{r}}{\dfrac{\text{USSR}}{n-k}} \sim F(r, n-k). \tag{61}$$

The RSSR is $\mathbf{y}^\top\mathbf{M}_1\mathbf{y}$ and the USSR $\mathbf{y}^\top\mathbf{M}_\mathbf{X}\mathbf{y}$. We can use the FWL Theorem to obtain a convenient expression for the difference between these two expressions. By this theorem, the USSR is the SSR

from the FWL regression

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta} + \text{residuals},$$

for which the TSS is $\mathbf{y}^\top\mathbf{M}_1\mathbf{y}$. The ESS can be expressed in terms of the orthogonal projection onto the r-dimensional subspace $\mathcal{S}(\mathbf{M}_1\mathbf{X}_2)$, and so the difference is:

$$\text{USSR} = \mathbf{y}^\top\mathbf{M}_1\mathbf{y} - \mathbf{y}^\top\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{y}.$$

Therefore,

$$\text{RSSR} - \text{USSR} = \mathbf{y}^\top\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{y},$$

and the F-statistic (61) can be written as

$$F_{\boldsymbol{\beta}_2} = \frac{\dfrac{\mathbf{y}^\top\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{y}}{r}}{\dfrac{\mathbf{y}^\top\mathbf{M}_\mathbf{X}\mathbf{y}}{n-k}}.$$

Under the null hypothesis, $\mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{M}_\mathbf{X}\mathbf{u}$ and $\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{u}$. Thus, under the null, the F-statistic reduces to

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{\epsilon}^\top\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon}/r}{\boldsymbol{\epsilon}^\top\mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}/(n-k)}, \tag{62}$$

where $\boldsymbol{\epsilon} = \mathbf{u}/\sigma$. I won't get into it here, but the F-statistic is actually the ratio of two quadratic quantities that follow a χ^2 distribution. The numerator is distributed as $\chi^2(n-k)$, and since the numerator can be written as $\boldsymbol{\epsilon}^\top\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\boldsymbol{\epsilon}$, it is distributed as $\chi^2(r)$ (this applies to the "conventional" expression for the F-statistic). Moreover, the random variables in the numerator and denominator are independent, because $\mathbf{M}_\mathbf{X}$ and $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}$ project onto mutually orthogonal subspaces:$\mathbf{M}_\mathbf{X}\mathbf{M}_1\mathbf{X}_2 = \mathbf{M}_\mathbf{X}(\mathbf{X}_2 - \mathbf{P}_1\mathbf{X}_2) = \mathbf{O}$.

Finally, we should note that (61) is analogous to how the likelihood-ratio statistic is derived in maximum likelihood estimation as the difference in log likelihood with and without the restriction of the null hypothesis. For this reason, this derivation of the F-statistic is said to be by the Likelihood-Ratio principle.

4.2 Asymptotic theory

So we just went through the exact- or finite-sample properties of the OLS estimator and its associated test statistics. However, not very often in economics are the assumptions of the exact distribution satisfied. The finite-sample theory breaks down if one of the following three assumptions is violated:

1. The exogeneity of regressors;
2. The normality of the error term; and
3. The linearity of the regression equation.

We now develop an approach in which we disregard the second assumption.[9] The approach is called asymptotic or large-sample theory, and it derives an approximation to the distribution of the estimator and its associated statistics assuming that the sample is sufficiently large.

Asymptotic theory relies on the limiting behaviour of a sequence of random variables, so we first cover some limit theorems and the idea of convergence. In the rest of these notes, a sequence of random variables $(z_1, z_2, ...)$ will be denoted by $\{z_n\}$.

4.2.1 Various modes of convergence

- Convergence in Probability
  - A sequence of random scalars $\{z_n\}$ converges in probability to a constant (non-random) α if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|z_n - \alpha| > \epsilon) = 0. \tag{63}$$

    The constant α is called the probability limit of z_n and is written as $\text{plim}_{n\to\infty} z_n = \alpha$ or $z_n \xrightarrow{p} \alpha$. Evidently,

$$z_n \xrightarrow{p} \alpha \Leftrightarrow z_n - \alpha \xrightarrow{p} 0.$$

    This definition of convergence in probability is extended to a sequence of random vectors or random matrices by requiring element-by-element convergence in probability. That is, a

[9] If we additionally disregarded the first assumption, then we would be looking at asymptotic theory and instrumental variable estimation. As such, there maybe some references and examples where we ignore the first assumption too.

sequence of k-dimensional random vectors $\{\mathbf{z}_n\}$ converges in probability to a k-dimensional vector of constants $\boldsymbol{\alpha}$ if, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|z_{nl} - \alpha_l| > \epsilon) = 0, \ \forall l = 1, ..., k, \tag{64}$$
$$\Leftrightarrow \operatorname*{plim}_{n\to\infty} \mathbf{z}_n = \boldsymbol{\alpha},$$

where z_{nl} is the l-th element of $\mathbf{z}_n$ and α_l the l-th element of $\boldsymbol{\alpha}$.

- Almost Sure Convergence
  - A sequence of random scalars $\{z_n\}$ converges almost surely to a constant α if

$$\Pr\left(\lim_{n\to\infty} z_n = \alpha\right) = 1. \tag{65}$$

    We write this as $z_n \overset{a.s.}{\to} \alpha$. The extension to random vectors is analogous to that for convergence in probability. This concept of convergence is stronger than convergence in probability; that is, if a sequence converges almost surely, then it converges in probability. The concept involved in (65) is harder to grasp because the probability is about an event concerning an infinite sequence $(z_1, z_2, ...)$. For our purposes, however, all that matters is that almost sure convergence is stronger than convergence in probability. If we can show that a sequence converges almost surely, that is one way to prove the sequence converges in probability.

- Convergence in Mean Square
  - A sequence of random scalars $\{z_n\}$ converges in mean square (or in quadratic mean) to α (written as $z_n \overset{m.s.}{\to} \alpha$) if

$$\lim_{n\to\infty} \mathbb{E}[(z_n - \alpha)^2] = 0. \tag{66}$$

    The extension to a random vector is analogous to that for convergence in probability: $\mathbf{z}_n \overset{m.s.}{\to} \boldsymbol{\alpha}$ if each element of $\mathbf{z}_n$ converges in mean square to the corresponding component of $\boldsymbol{\alpha}$.

- Convergence to a Random Variable
  - In these definitions of convergence, the limit is a constant (i.e., a real number). The limit can be a random variable. We say that a sequence of k-dimensional random variables $\{\mathbf{z}_n\}$ converges to a k-dimensional random variable $\mathbf{z}$ and write $\mathbf{z}_n \xrightarrow{p} \mathbf{z}$ if $\{\mathbf{z}_n - \mathbf{z}\}$ converges to $\mathbf{0}$:

$$\mathbf{z}_n \xrightarrow{p} \mathbf{z} \Leftrightarrow \mathbf{z}_n - \mathbf{z} \xrightarrow{p} \mathbf{0}. \tag{67}$$

    Similarly,

$$\mathbf{z}_n \xrightarrow{a.s.} \mathbf{z} \Leftrightarrow \mathbf{z}_n - \mathbf{z} \xrightarrow{a.s.} \mathbf{0}, \tag{68}$$

$$\mathbf{z}_n \xrightarrow{m.s.} \mathbf{z} \Leftrightarrow \mathbf{z}_n - \mathbf{z} \xrightarrow{m.s.} \mathbf{0}. \tag{69}$$

- Convergence in Distribution
  - Let $\{z_n\}$ denote a sequence of random scalars and F_n be the cumulative distribution function (CDF) of z_n. We say that $\{z_n\}$ converges in distribution to a random scalar z if the CDF F_n of z_n converges to the CDF F of z at every continuity point of F. We write $z_n \xrightarrow{d} z$ and call F the asymptotic (or limit or limiting) distribution of z_n. Sometimes we write $z_n \xrightarrow{d} F$ when the distribution F is well-known. For example, $z_n \xrightarrow{d} N(0,1)$ should read "$z_n \xrightarrow{d} z$ and the distribution of z is $N(0,1)$, the standard normal distribution mean 0 and variance 1." It can be shown from the definition that convergence in probability is stronger than convergence in distribution, that is,

$$z_n \xrightarrow{p} z \implies z_n \xrightarrow{d} z. \tag{70}$$

  - A special case of convergence in distribution is that z is a constant (a trivial random variable).
  - The extension to a sequence of random vectors is immediate: $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$ if the joint CDF F_n of the random vector $\mathbf{z}_n$ converges to the joint CDF F of $\mathbf{z}$ at every continuity point

of F. Note, however, that, unlike the other concepts of convergence, for convergence in distribution, element-by-element convergence does not necessarily mean convergence for the vector sequence. That is, "each element $\mathbf{z}_n \xrightarrow{d}$ corresponding element of $\mathbf{z}$" does not necessarily imply $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$. A common way to establish the connection between scalar convergence and vector convergence in distribution is

* **Theorem (Multivariate Convergence in Distribution)** (Rao 1973): Let $\{\mathbf{z}_n\}$ be a sequence of k-dimensional random vectors. Then

$$\mathbf{z}_n \xrightarrow{d} \mathbf{z} \Leftrightarrow \boldsymbol{\lambda}^\top \mathbf{z}_n \xrightarrow{d} \boldsymbol{\lambda}^\top \mathbf{z},$$

for any k-dimensional vector of real numbers $\boldsymbol{\lambda}$.

- Convergence in Distribution vs Convergence in Moments
  - It is worth emphasising that the moments of the limit distribution of z_n are not necessarily equal to the limits of the moments of z_n. For example, $z_n \xrightarrow{d} z$ does not necessarily imply $\lim_{n\to\infty} \mathbb{E}[z_n] = \mathbb{E}[z]$. However,
    * **Lemma (Convergence in Distribution and in Moments)**: Let α_{sn} be the s-th moment of z_n and $\lim_{n\to\infty} \alpha_{sn} = \alpha_s$, where α_s is finite (i.e., a real number). Suppose that for some $\delta > 0$, $\mathbb{E}[|z_n|^{s+\delta}] < M < \infty, \forall n$. Then: $z_n \xrightarrow{d} z$ implies that α_s is the s-th moment of z.
  - Thus, for example, if the variance of a sequence of random variables converging in distribution converges to some finite number, then that number is the variance of the limiting distribution.
- Relation among Modes of Convergence
  - Some modes of convergence are weaker than others. The following theorem establishes the relationship between the four modes of convergence.

∗ **Lemma (Relationship among the Four Modes of Convergence)**:

$$\mathbf{z}_n \xrightarrow{m.s.} \boldsymbol{\alpha} \implies \mathbf{z}_n \xrightarrow{p} \boldsymbol{\alpha}, \ \therefore \mathbf{z}_n \xrightarrow{m.s.} \mathbf{z} \implies \mathbf{z}_n \xrightarrow{p} \mathbf{z}, \tag{71}$$

$$\mathbf{z}_n \xrightarrow{a.s.} \boldsymbol{\alpha} \implies \mathbf{z}_n \xrightarrow{p} \boldsymbol{\alpha}, \ \therefore \mathbf{z}_n \xrightarrow{a.s.} \mathbf{z} \implies \mathbf{z}_n \xrightarrow{p} \mathbf{z}, \tag{72}$$

$$\mathbf{z}_n \xrightarrow{p} \boldsymbol{\alpha} \Leftrightarrow \mathbf{z} \xrightarrow{d} \boldsymbol{\alpha}, \tag{73}$$

where the third line states that if the limiting random variable is a constant (is a trivial random variable), convergence in distribution is the same as convergence in probability.

Having defined the modes of convergence, we can state three results essential for developing asymptotic theory.

1. **Lemma (Preservation of Convergence for Continuous Transformation)**: Suppose $\mathbf{a}(\cdot)$ is a vector-valued continuous function that does not depend on n.

   (a) $\mathbf{z}_n \xrightarrow{p} \boldsymbol{\alpha} \implies \mathbf{a}(\mathbf{z}_n) \xrightarrow{p} \mathbf{a}(\boldsymbol{\alpha})$, where $\mathbf{a}(\cdot)$ is continuous at $\boldsymbol{\alpha}$. Stated differently,

   $$\operatorname*{plim}_{n\to\infty} \mathbf{a}(\mathbf{z}_n) = \mathbf{a} \operatorname*{plim}_{n\to\infty} \mathbf{z}_n,$$

   provided the plim exists. An implication of this is that the usual arithmetic operations preserve convergence in probability. For example:

   $$\begin{aligned} x_n \xrightarrow{p} \beta, y_n \xrightarrow{p} \gamma &\implies x_n + y_n \xrightarrow{p} \beta + \gamma, \\ &\implies x_n y_n \xrightarrow{p} \beta\gamma, \\ &\implies x_n / y_n \xrightarrow{p} \beta/\gamma, \ \gamma \neq 0, \\ \mathbf{Y}_n \xrightarrow{p} \boldsymbol{\Gamma} &\implies \mathbf{Y}_n^{-1} \xrightarrow{p} \boldsymbol{\Gamma}^{-1}, \ \boldsymbol{\Gamma}^{-1} \exists. \end{aligned}$$

   (b) $\mathbf{z}_n \xrightarrow{d} \mathbf{z} \implies \mathbf{a}(\mathbf{z}_n) \xrightarrow{d} \mathbf{a}(\mathbf{z})$, where $\mathbf{a}(\cdot)$ is continuous everywhere.

2. **Lemma (Slutsky's Theorem)**: Technically, (a) and (c) below are called Slutsky's Theorem:

   (a) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{y}_n \xrightarrow{p} \boldsymbol{\alpha} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x} + \boldsymbol{\alpha}$.

(b) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{y}_n \xrightarrow{p} \mathbf{0} \implies \mathbf{y}_n^\top \mathbf{x}_n \xrightarrow{d} 0$.

(c) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{A}_n \mathbf{x}_n \xrightarrow{d} \mathbf{A}\mathbf{x}$, provided that $\mathbf{A}_n$ and $\mathbf{x}_n$ are conformable. In particular, if $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\mathbf{A}_n \mathbf{x}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

(d) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{x}_n^\top \mathbf{A}_n^{-1} \mathbf{x}_n \xrightarrow{d} \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x}$, provided that $\mathbf{A}_n$ and $\mathbf{x}_n$ are conformable and that $\mathbf{A}$ is nonsingular ($\mathbf{A}^{-1} \exists$).

- By setting $\boldsymbol{\alpha} = \mathbf{0}$, part (a) implies:

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{y}_n \xrightarrow{p} \mathbf{0} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x}. \tag{74}$$

  That is, if $\mathbf{z}_n = \mathbf{x}_n + \mathbf{y}_n$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{0}$ (i.e., $\mathbf{z}_n - \mathbf{x}_n \xrightarrow{p} \mathbf{0}$), then the asymptotic distribution of $\mathbf{z}_n$ is the same as that of $\mathbf{x}_n$. When $\mathbf{z}_n - \mathbf{x}_n \xrightarrow{p} \mathbf{0}$, we sometimes say that the two sequences are asymptotically equivalent and write it as

$$\mathbf{z}_n \overset{a}{\sim} \mathbf{x}_n \Leftrightarrow \mathbf{z}_n = \mathbf{x}_n + o_p,$$

  where o_p is some suitable random variable ($\mathbf{y}_n$ here) that converges to zero in probability.

- A standard trick in deriving the asymptotic distribution of a sequence of random variables is to find an asymptotically equivalent sequence whose asymptotic distribution is easier to derive. In particular, by replacing $\mathbf{y}_n$ by $\mathbf{y}_n - \boldsymbol{\alpha}$ in part (b) of the lemma, we obtain

$$\begin{aligned} \mathbf{x}_n \xrightarrow{d} \mathbf{x}, \mathbf{y}_n \xrightarrow{p} \boldsymbol{\alpha} &\implies \mathbf{y}_n^\top \mathbf{x}_n \sim \boldsymbol{\alpha}^\top \mathbf{x}_n, \\ &\Leftrightarrow \mathbf{y}_n^\top \mathbf{x}_n = \boldsymbol{\alpha} \mathbf{x}_n + o_p. \end{aligned} \tag{75}$$

  The o_p here is $(\mathbf{y}_n - \boldsymbol{\alpha})\mathbf{x}_n$. Therefore, replacing $\mathbf{y}_n$ by its probability limit does not change the asymptotic distribution of $\mathbf{y}_n^\top \mathbf{x}_n$, so long as $\mathbf{x}_n$ converges in distribution to some random variable.

3. **Lemma (The Delta Method)**: Suppose $\{\mathbf{x}_n\}$ is a sequence of k-dimensional random vectors

such that $\mathbf{x}_n \xrightarrow{p} \boldsymbol{\beta}$ and

$$\sqrt{n}(\mathbf{x}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{z},$$

and suppose $\mathbf{a}(\cdot) : \mathbb{R}^k \to \mathbb{R}^r$ has continuous first derivatives with $\mathbf{A}(\boldsymbol{\beta})$ denoting the $r \times k$ matrix of first derivatives evaluated at $\boldsymbol{\beta}$:

$$\underset{r\times k}{\mathbf{A}(\boldsymbol{\beta})} \equiv \frac{\partial \mathbf{a}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}.$$

Then

$$\sqrt{n}\left[\mathbf{a}(\mathbf{x}_n) - \mathbf{a}(\boldsymbol{\beta})\right] \xrightarrow{d} \mathbf{A}(\boldsymbol{\beta})\mathbf{z}.$$

In particular:

$$\sqrt{n}(\mathbf{x}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}) \implies \sqrt{n}\left[\mathbf{a}(\mathbf{x}_n) - \mathbf{a}(\boldsymbol{\beta})\right] \xrightarrow{d} N\left(\mathbf{0}, \mathbf{A}(\boldsymbol{\beta})\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\beta})^\top\right).$$

Proving the Delta Method is a good way to recap the lemmas we've covered so far.

- Proof: By the Mean Value Theorem from calculus,[10] there exists a k-dimension vector $\mathbf{y}_n$ between $\mathbf{x}_n$ and $\boldsymbol{\beta}$ such that

  $$\mathbf{a}(\mathbf{x}_n) - \mathbf{a}(\boldsymbol{\beta}) = \underset{r\times k}{\mathbf{A}(\mathbf{y}_n)}\underset{k\times 1}{(\mathbf{x}_n - \boldsymbol{\beta})}.$$

  Multiplying both sides by $\sqrt{n}$, we obtain

  $$\sqrt{n}\left[\mathbf{a}(\mathbf{x}_n) - \mathbf{a}(\boldsymbol{\beta})\right] = \mathbf{A}(\mathbf{y}_n)\sqrt{n}(\mathbf{x}_n - \boldsymbol{\beta}).$$

  Since $\mathbf{y}_n$ is between $\mathbf{x}_n$ and $\boldsymbol{\beta}$ and since $\mathbf{x}_n \xrightarrow{p} \boldsymbol{\beta}$, we know that $\mathbf{y}_n \xrightarrow{p} \boldsymbol{\beta}$. Moreover, the first derivative $\mathbf{A}(\cdot)$ is continuous by assumption. So by using the lemma of preservation of

[10] **Theorem (Mean Value)**: Let $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^q$ be continuously differentiable. Then $\mathbf{h}(\mathbf{x})$ admits the mean value expansion

$$\underset{q\times 1}{\mathbf{h}(\mathbf{x})} = \underset{q\times 1}{\mathbf{h}(\mathbf{x}_0)} + \underset{q\times p}{\frac{\partial \mathbf{h}(\bar{\mathbf{x}})}{\partial \mathbf{x}^\top}}\underset{p\times 1}{(\mathbf{x} - \mathbf{x}_0)},$$

where $\bar{\mathbf{x}}$ is a mean value lying between $\mathbf{x}$ and $\mathbf{x}_0$. The Mean Value Theorem only applies to individual elements of $\mathbf{h}$, so that $\bar{\mathbf{x}}$ actually differs from element to element of the vector equation. This complication does not affect our discussion.

convergence for continuous transformations, we have

$$\mathbf{A}(\mathbf{y}_n) \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}).$$

By Slutsky's Theorem, this and the hypothesis that $\sqrt{n}(\mathbf{x}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{z}$ imply that

$$\mathbf{A}(\mathbf{y}_n)\sqrt{n}(\mathbf{x}_n - \boldsymbol{\beta}) = \sqrt{n}\,[\mathbf{a}(\mathbf{x}_n) - \mathbf{a}(\boldsymbol{\beta})] \xrightarrow{d} \mathbf{A}(\boldsymbol{\beta})\mathbf{z}. \quad \blacksquare$$

4.2.2 Root-n consistency

Let $\hat{\boldsymbol{\theta}}_n$ be an estimator of a parameter vector $\boldsymbol{\theta}$ based on a sample of size n. The sequence $\{\hat{\boldsymbol{\theta}}_n\}$ is an example of a sequence of random variables, so the concepts introduced above for sequences of random variables are applicable to $\{\hat{\boldsymbol{\theta}}_n\}$. We say that an estimator $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ if

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta} \Leftrightarrow \hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}.$$

The asymptotic bias of $\hat{\boldsymbol{\theta}}_n$ is defined as $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$. So if the estimator is consistent, its asymptotic bias is zero. A consistent estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically normal if

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Such an estimator is called $\sqrt{n}$-consistent. The variance matrix $\boldsymbol{\Sigma}$ is called the asymptotic variance and is denoted $\operatorname{AVar}(\hat{\boldsymbol{\theta}}_n)$.[11]

4.2.3 Laws of large numbers and Central Limit Theorem

For a sequence of random scalars $\{z_i\}$, the sample mean $\bar{z}_n$ is defined as

$$\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i.$$

[11] Some books and lecturers use the notation $\operatorname{AVar}(\hat{\boldsymbol{\theta}}_n)$ to mean $\boldsymbol{\Sigma}/n$ (which is zero in the limit). I'm sticking to the notation in Davidson and MacKinnon (2004) and Hayashi (2000), where $\operatorname{AVar}(\hat{\boldsymbol{\theta}}_n)$ refers to the variance of the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$.

Consider the sequence $\{\bar{z}_n\}$. Laws of large numbers (LLNs) concern conditions under which $\{\bar{z}_n\}$ converges either in probability or almost surely. An LLN is called strong if the convergence is almost surely and weak if the convergence is in probability. We can derive the following weak LLN easily from (71).

Definition (Chebychev's Weak LLN):

$$\lim_{n\to\infty} \mathbb{E}[\bar{z}_n] = \mu, \lim_{n\to\infty} \text{Var}(\bar{z}_n) = 0 \implies \bar{z}_n \xrightarrow{p} \mu.$$

This holds because, under the condition specified, it is easy to prove that $\bar{z}_n \xrightarrow{m.s.} \mu$. The following strong LLN assumes that $\{z_i\}$ is IID, but the variance does not need to be finite.

Definition (Kolmogorov's Second Strong LLN): Let $\{z_i\}$ be IID with $\mathbb{E}[z_i] = \mu$, then

$$\bar{z}_n \xrightarrow{a.s.} \mu$$

These LLNs extend readily to random vectors by requiring element-by-element convergence.

Central Limit Theorems (CLTs) are about the limiting behaviour of the difference between $\bar{z}_n$ and $\mathbb{E}[\bar{z}_n]$ (which equals $\mathbb{E}[z_i]$ if $\{z_i\}$ is IID) blown up by $\sqrt{n}$. The only CLT we need for the case of IID sequences is:

Definition (Lindeberg-Levy CLT): Let $\{\mathbf{z}_i\}$ be IID with $\mathbb{E}[\mathbf{z}_i] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{z}_i) = \boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\bar{\mathbf{z}}_n - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{z}_i - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

This reads: a sequence of random vectors $\{\sqrt{n}(\bar{\mathbf{z}}_i - \boldsymbol{\mu})\}$ converges in distribution tp a random vector whose distribution is $N(\mathbf{0}, \boldsymbol{\Sigma})$. Technically, the Lindeberg-Levy CLT is for a sequence of scalar random variables. The vector version is derived from the scalar version which is as follows. Let $\{\mathbf{z}_i\}$ be IID with $\mathbb{E}[\mathbf{z}_i] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{z}_i) = \boldsymbol{\Sigma}$, and let $\boldsymbol{\lambda}$ be any vector of real numbers of the same dimension. Then $\{\boldsymbol{\lambda}^\top \mathbf{x}_n\}$ is a sequence of scalar random variables with $\mathbb{E}[\boldsymbol{\lambda}^\top \mathbf{x}_n] = \boldsymbol{\lambda}^\top \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{\lambda}^\top \mathbf{x}_n) = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}$. The

scalar version of Lindeberg-Levy then implies that

$$\sqrt{n}(\boldsymbol{\lambda}^\top \bar{\mathbf{z}}_n - \boldsymbol{\lambda}^\top \boldsymbol{\mu}) = \boldsymbol{\lambda}^\top \sqrt{n}(\bar{\boldsymbol{z}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(0, \boldsymbol{\lambda}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}).$$

But this limit distribution is the distribution of $\boldsymbol{\lambda}^\top \mathbf{x}$ where $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. So by the Multivariate Convergence in Distribution Theorem, $\{\sqrt{n}(\bar{\mathbf{z}}_n - \boldsymbol{\mu})\} \xrightarrow{d} \mathbf{x}$, which is the claim of the vector version of Lindeberg-Levy.

4.2.4 The model and its assumptions

With our theorems and lemmas established for asymptotic theory, we now return back to the model. As previously mentioned, we drop the distribution assumption of the error term (no longer normally distributed). Further, the requirement in finite-sample theory that the regressors be strictly exogenous or "fixed" is replaced by the much weaker requirement that they be "predetermined."

We will proceed with the following assumptions.

1. Linearity: The model defined in (47) is linear, where $\mathbf{X}$ is a k-column matrix of regressors, $\boldsymbol{\beta}$ is a k-dimensional coefficient vector, and $\mathbf{u}$ is the unobservable error term:[12]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

2. Ergodic stationarity: The $(k+1)$-dimensional vector stochastic process $\{\mathbf{y}, \mathbf{X}\}$ is jointly stationary and ergodic.[13]

   - A trivial but important special case of ergodic stationarity is that $\{y_t, \mathbf{X}_t\}$ is IID, that is, the sample is a random sample. Most existing microdata on households are random samples, with observations randomly drawn from a population of a nation's households. Thus, we are in no way ruling out models that use cross-section data.

[12] Notice that unlike (47), we do not assume that $\mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I})$! This has some important implications, which will be discussed later.

[13] We won't cover ergodicity here, nor Martingale difference sequences, as it's more relevant in time-series analysis. But I am introducing some base concepts here as they're good to be aware of.

3. Predetermined regressors: All the regressors are predetermined in the sense that they are orthogonal to the contemporaneous error term: $\mathbb{E}[u_t|\mathbf{X}_t] = 0$ and $\mathbb{E}[u_t^2|\mathbf{X}_t] = \sigma_0^2$. Alternatively, this can be written as:

$$\begin{aligned}\mathbb{E}[\mathbf{X}_t^\top(y_t - \mathbf{X}_t\boldsymbol{\beta})] &= \mathbf{0} \\ \Leftrightarrow \mathbb{E}[\mathbf{g}_i] &= \mathbf{0},\end{aligned}$$

where $\mathbf{g}_t \equiv \mathbf{X}_t^\top u_t$.

- The model accommodates conditional heteroskedasticity. If $\{y_t, \mathbf{X}_t\}$ is stationary, then the error term $u_t = y_t - \mathbf{X}_t\boldsymbol{\beta}$ is also stationary. Thus, this assumption implies that the unconditional second moment $\mathbb{E}[u_t^2]$, if it exists and is finite, is constant across t. That is, the error term is unconditionally homoskedastic. Yet the error can be conditionally heteroskedastic in that the conditional second moment, $\mathbb{E}[u_t^2|\mathbf{X}_t]$, can depend on $\mathbf{X}_t$. A treatment where the error is conditionally homoskedastic is discussed in the next section.
- $\mathbb{E}[\mathbf{X}_t^\top u_t] = \mathbf{0}$ vs $\mathbb{E}[u_t|\mathbf{X}_t] = 0$: Sometimes, instead of the orthogonality conditions, $\mathbb{E}[\mathbf{X}_t^\top u_t] = \mathbf{0}$, it is assumed that the error is unrelated in the sense that $\mathbb{E}[u_t|\mathbf{X}_t] = 0$. This is stronger than the orthogonality condition because it implies that, for any measurable function f of $\mathbf{X}_t$, $f(\mathbf{X}_t)$ is orthogonal to u_t:

$$\begin{aligned}\mathbb{E}[f(\mathbf{X}_t)] &= \mathbb{E}\left[\mathbb{E}[f(\mathbf{X}_t)u_t|\mathbf{X}_t]\right] \\ &= \mathbb{E}\left[f(\mathbf{X}_t)\mathbb{E}[u_t|\mathbf{X}_t]\right] \\ &= 0.\end{aligned}$$

  In rational expectations models, this stronger condition is satisfied, but for the purpose of developing asymptotic theory, we need only the weaker assumption of the orthogonality conditions.
- Predetermined vs strictly exogenous regressors: The regressors are not required to be strictly exogenous. When we assumed strict exogeneity, we ruled out the possibility that the current

error term, u_t, is correlated with future regressors, $\mathbf{X}_{t+j}, j \geq 1$. The assumption we make in this section restricts only the contemporaneous relationship between the error term and the regressors. For example, the AR(1) process, which does not satisfy the exogeneity assumption of the classical regression model, can be accommodated in the model.

4. Rank condition: The $k \times k$ matrix $\mathbb{E}[\mathbf{X}_t^\top \mathbf{X}_t]$ is nonsingular (and hence finite). We denote this matrix by $\mathbf{\Sigma}_{\mathbf{X}^\top \mathbf{X}}$.

   - There is no multicollinearity in the limit. Since $\mathbb{E}[\mathbf{X}_t^\top \mathbf{X}_t]$ is finite by the assumption of $\lim_{n\to\infty} \mathbf{S}_{\mathbf{X}^\top \mathbf{X}} = \mathbf{\Sigma}_{\mathbf{X}^\top \mathbf{X}}$, with probability 1 by the Ergodic Theorem,[14] where, recall, $\mathbf{S}_{\mathbf{X}^\top \mathbf{X}} = \frac{1}{n} \sum_{t=1}^n \mathbf{X}_t^\top \mathbf{X}_t$. So, for n sufficiently large, the sample cross moment of the regressors, $\mathbf{S}_{\mathbf{X}^\top \mathbf{X}}$, is nonsingular. Since $\mathbf{S}_{\mathbf{X}^\top \mathbf{X}}$ is nonsingular if and only if $\text{rank}(\mathbf{X}) = k$, the assumption of no multicollinearity is satisfied with probability 1 for sufficiently large n.

5. $\mathbf{g}_t$ is a Martingale difference sequence (MDS) with finite second moments: $\{\mathbf{g}_t\}$ is a MDS (so fortiori $\mathbb{E}[\mathbf{g}_t] = \mathbf{0}$). The $k \times k$ matrix of cross moments,

$$\begin{aligned} \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top] &= \mathbb{E}\left[\mathbf{X}_t^\top u_t u_t \mathbf{X}_t\right] \\ &= \sigma_0^2 \mathbb{E}\left[\mathbf{X}_t^\top \mathbf{X}_t\right], \end{aligned}$$

   is nonsingular. We use $\mathbf{S}$ for $\text{AVar}(\bar{\mathbf{g}})$, the variance of the asymptotic distribution of $\sqrt{n}\bar{\mathbf{g}}$, where $\bar{\mathbf{g}} = n^{-1} \sum_{t=1}^n \mathbf{g}_t$, and $\mathbf{S} = \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top]$.

   - Since an MDS is zero mean by definition, this assumption is stronger than Assumption (3). We will need this assumption to prove the asymptotic normality of the OLS estimator. The assumption, about the product of the regressors and the error term, may be hard to interpret. A sufficient condition that is easier to interpret is

$$\mathbb{E}[u_t | u_{t-1}, u_{t-2}, ..., u_1, \mathbf{X}_t, \mathbf{X}_{t-1}, ..., \mathbf{X}_1] = 0. \tag{76}$$

[14] Again, not going to cover this here. But it's something to be aware of for time-series analysis.

Note that the current as well as lagged regressors is included in the information set. This condition implies that the error term is serially uncorrelated and also is uncorrelated with the current and past regressors. That (76) is sufficient for $\{\mathbf{g}_t\}$ to be an MDS can be seen as follows. We have

$$\mathbb{E}[\mathbf{g}_t|\mathbf{g}_{t-1},...,\mathbf{g}_1] = \mathbb{E}\left[\mathbb{E}[\mathbf{g}_t|u_{t-1},u_{t-2},...,u_1,\mathbf{X}_t,\mathbf{X}_{t-1},...,\mathbf{X}_1|\mathbf{g}_{t-1},...,\mathbf{g}_1\right].$$

This holds by the LIE because there is more information in the "inside" of the information set $(u_{t-1},u_{t-2},...,u_1,\mathbf{X}_t,\mathbf{X}_{t-1},...,\mathbf{X}_1)$ than in the "outside" information set $(\mathbf{g}_{t-1},...,\mathbf{g}_1)$. Therefore,

$$\begin{aligned}\mathbb{E}[\mathbf{g}_t|\mathbf{g}_{t-1},...,\mathbf{g}_1] &= \mathbb{E}\left[\mathbf{X}_t^\top\mathbb{E}[u_t|u_{t-1},u_{t-2},...,u_1,\mathbf{X}_t,\mathbf{X}_{t-1},...,\mathbf{X}_1|\mathbf{g}_{t-1},...,\mathbf{g}_1\right]\\ &= \mathbf{0}.\end{aligned}$$

- When the regressors include a constant: In virtually all applications, the regressors include a constant. If the regressors include a constant so that $x_{t,1} = 1, \forall t \implies \mathbf{x}_1 = \boldsymbol{\iota}$, then Assumption (3) of predetermined regressors can be stated in more familiar terms: the mean of the error term is zero (which is implied by $\mathbb{E}[x_{t,1}u_t] = 0$), and the contemporaneous correlation between the error term and the regressors is zero (which is implied by $\mathbb{E}[x_{t,i}u_t] = 0$ for $i \neq 1$ and $\mathbb{E}[u_t] = 0$). Also, since the first element of the k-dimensional vector $\mathbf{g}_t$ is ϵ_t, Assumption (5) implies

$$\mathbb{E}[u_t|\mathbf{g}_{t-1},\mathbf{g}_{t-2},...,\mathbf{g}_1] = 0.$$

Then, by the LIE, $\{u_t\}$ is a scalar MDS:

$$\mathbb{E}[u_t|u_{t-1},u_{t-2},...,u_1] = 0.$$

Therefore, Assumption (5) implies that the error term itself is an MDS and hence is serially uncorrelated.

- $\mathbf{S}$ is a matrix of fourth moments. Since $\mathbf{g}_t \equiv \mathbf{X}_t^\top u_t$, the $\mathbf{S}$ in Assumption (5) can be written as $\mathbb{E}[u_t^2 \mathbf{X}_t^\top \mathbf{X}_t]$. Its (i,j) element is $\mathbb{E}[u_t^2 x_{t,i} x_{t,j}]$. So $\mathbf{S}$ is a matrix of fourth moments (the expectation of products of four different variables). Consistent estimation of $\mathbf{S}$ will require an additional assumption (which we will cover later).
- $\mathbf{S}$ will take a different expression without Assumption (5). Thanks to the assumption that $\{\mathbf{g}_t\}$ is an MDS, $\mathbf{S} = \text{AVar}(\bar{\mathbf{g}}) = \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top]$. Without the assumption, the expression for $\mathbf{S}$ is more complicated and involves autocovariances of $\mathbf{g}_t$.

4.2.5 Asymptotic distribution of the OLS estimator

We now prove that the OLS estimator is consistent and asymptotically normal. Where convenient, we may presume that there is available some consistent estimator, $\hat{\mathbf{S}}$, of $\mathbf{S} = \text{AVar}(\bar{\mathbf{g}}) = \sigma_0^2 \mathbb{E}[\mathbf{X}_t^\top \mathbf{X}_t]$. The issue of estimating $\mathbf{S}$ consistently will be taken up later.

Proposition: The asymptotic distribution of the OLS estimator is as follows:

1. Consistency of $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$: Under Assumption (1)-(4),

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0,$$

so Assumption (5) is not needed for consistency.

- Proof: Begin by writing the sample error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \\
&= \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \mathbf{u} \\
&= \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \mathbf{u},
\end{aligned}$$

where all we've done is some simple algebra. Now, note that:

$$\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1} \xrightarrow{p} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1},$$
$$\frac{1}{n}\mathbf{X}^\top\mathbf{u} = \bar{\mathbf{g}} \xrightarrow{p} = \mathbb{E}[\mathbf{X}^\top\mathbf{u}] = \mathbf{0},$$

so we can write

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\bar{\mathbf{g}}. \tag{77}$$

Now, we know that $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}^{-1}$ and $\bar{\mathbf{g}} \xrightarrow{p} \mathbb{E}[\mathbf{g}_t] = \mathbf{0}$, so

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}^{-1}\mathbf{0}$$
$$\implies \operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0. \quad \blacksquare$$

2. Asymptotic normality of $\hat{\boldsymbol{\beta}}$: If Assumption (3) is strengthened as Assumption (5), then:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0}, \text{AVar}(\hat{\boldsymbol{\beta}})\right),$$

where

$$\text{AVar}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}^{-1}.$$

- Proof: There's actually two ways to go about proving this. The reason for the difference is down to the implication brought about by assuming conditional heteroskedasticity and unconditional homoskedasticity, as well as Assumption (5). I will first outline the methodology outlined in Hayashi (2000) as it's quite short. I will then discuss an alternative approach by Davidson and MacKinnon (2004). Begin by multiplying the sample error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ from (77) and then multiply by the factor $\sqrt{n}$:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\sqrt{n}\bar{\mathbf{g}}.$$

Now, we know that due to Assumption (5): $\sqrt{n}\bar{\mathbf{g}} = \sqrt{n}\frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^\top u_t = n^{-1/2}\mathbf{X}^\top\mathbf{u} \xrightarrow{d} N(\mathbf{0},\mathbf{S})$. So by Slutsky's Theorem, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ converges to a normal distribution with mean $\mathbf{0}$, and since $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}$ is symmetric, with variance $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\mathbf{S}\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}$. Also, by Assumption (2), $\mathbf{X}_t^\top\mathbf{X}_t$ is ergodic stationary, so $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}} \xrightarrow{a.s.} \boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}$. Since $\boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}$ is invertible by Assumption (4), $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1} \xrightarrow{a.s.} \boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{X}}^{-1}$. Thus, our proof is complete. ■

3. Consistent estimate of $\text{AVar}(\hat{\boldsymbol{\beta}})$: Suppose there is available a consistent estimator, $\hat{\mathbf{S}}$, of $\mathbf{S}$. Then, under Assumption (2), $\text{AVar}(\hat{\boldsymbol{\beta}})$ is consistently estimated by

$$\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}}) = \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}. \tag{78}$$

- A proof really isn't needed here, as, in the words of Hayashi (2000), it's an immediate implication of ergodic stationarity.

The proofs used above are a showcase of all the standard tricks in asymptotic theory. To prove Proposition (1) and (2): write the object in question in terms of sample means, apply the relevant LLN and CLT to sample means, and then use Slutsky's Theorem to derive the asymptotic distribution.

4.2.6 Consistency of s^2

Proposition: The OLS estimator of the error variance is consistent. Let $\hat{u}_t \equiv y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}}$ be the OLS residual for observation t. If we assume that the model is linear, $\mathbf{X}$ is strictly exogenous, $\mathbf{X}$ has no multicollinearity, and the errors are homoskedastic in their variance, we have

$$s^2 \equiv \frac{1}{n-k}\sum_{t=1}^{n}\hat{u}_t^2 \xrightarrow{p} \mathbb{E}[u_t^2],$$

provided that $\mathbb{E}[\hat{u}_t^2]$ exists and is finite.

If we could observe the error term, u_t, then the obvious estimator for σ_0^2 would be the sample mean of u_t^2. It is consistent by ergodic stationarity. The proposition we just made implies that the substitution of the OLS residual, $\hat{u}_t$, for the true error term, u_t, does not impair consistency.

Proof: Since

$$s^2 = \frac{n}{n-k}\left(\frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2\right),$$

it suffices to prove that the sample mean of $\hat{u}_t^2$, $n^{-1}\sum_t \hat{u}_t^2$, converges in probability to $\mathbb{E}[u_t^2]$. Begin by writing the probability limit

$$\operatorname*{plim}_{n\to\infty} s^2 = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2,$$

and apply a LLN to the RHS:

$$\operatorname*{plim}_{n\to\infty} s^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[\hat{u}_t^2].$$

The easiest way to calculate the variance of $\hat{u}_t$ is to calculate the covariance matrix of the entire vector $\hat{\mathbf{u}}$:

$$\begin{aligned}
\operatorname{Var}(\hat{\mathbf{u}}) &= \operatorname{Var}(\mathbf{M_X u}) \\
&= \mathbb{E}[\mathbf{M_X u u}^\top \mathbf{M_X}] \\
&= \mathbf{M_X}\mathbb{E}[\mathbf{u u}^\top]\mathbf{M_X} \\
&= \mathbf{M_X}\operatorname{Var}(\mathbf{u})\mathbf{M_X} \\
&= \mathbf{M_X}\sigma_0^2\mathbf{I}\mathbf{M_X} \\
&= \sigma_0^2\mathbf{M_X}\mathbf{M_X} \\
&= \sigma_0^2\mathbf{M_X}.
\end{aligned}$$

We immediately see that, in general, $\mathbb{E}[\hat{u}_t\hat{u}_s] \neq 0$ for $t \neq s$. Thus, even though the original error terms are assumed to be uncorrelated, the residuals are not uncorrelated. It can also be seen that the residuals do not have a constant variance, and that the variance of every residual must always be smaller than σ_0^2. Let h_t denote the t-th diagonal element of the projection matrix $\mathbf{P_X}$. Thus, a typical diagonal element of $\mathbf{M_X}$ is $1 - h_t$. Therefore, it follows that

$$\operatorname{Var}(\hat{u}_t) = \mathbb{E}[\hat{u}_t^2] = (1-h_t)\sigma_0^2.$$

Since $0 \leq h_t \leq 1$, this equation implies that $\mathbb{E}[\hat{u}_t^2]$ is always smaller than σ_0^2.

Returning back to the probability limit of s^2, we can write:

$$\operatorname*{plim}_{n\to\infty} s^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \sigma_0^2(1 - h_t).$$

Noting that $\sum_{t=1}^{n} h_t = k$,[15] so

$$\operatorname*{plim}_{n\to\infty} s^2 = \lim_{n\to\infty} \frac{n-k}{n} \sigma_0^2 = \sigma_0^2. \tag{79}$$

Therefore, s^2, is indeed a consistent estimator. ■

4.2.7 Asymptotic t-test

Statistical inference in large-sample theory is based on test statistics whose asymptotic distributions are known under the truth of the null hypothesis. Derivation of the distribution of test statistics is easier than in finite-sample theory because we are only concerned about the large-sample approximation to the exact distribution. In this section, we derive test statistics, assuming throughout that a consistent estimator, $\hat{\mathbf{S}}$, of $\mathbf{S} = \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top]$.

Consider testing a hypothesis about the i-th coefficient β_i, $H_0 : \beta_i = \beta_0$, we would then have

$$\sqrt{n}(\hat{\beta}_i - \beta_0) \xrightarrow{d} N(0, \text{AVar}(\hat{\beta}_i),$$
$$\widehat{\text{AVar}}(\hat{\beta}_i) \xrightarrow{p} \text{AVar}(\hat{\beta}_i),$$

where $\hat{\beta}_i$ is is the i-th element of $\hat{\boldsymbol{\beta}}$ and $\text{AVar}(\hat{\beta}_i)$ is the (i,i) element of the $k \times k$ matrix $\text{AVar}(\hat{\boldsymbol{\beta}})$. So

[15] This is because

$$\sum_{t=1}^{n} h_t = \text{Tr}(\mathbf{P}_\mathbf{X}) = \text{Tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)$$
$$= \text{Tr}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}) = \text{Tr}(\mathbf{I}_k) = k.$$

by Slutsky's Theorem, we can write

$$t_{\beta_i} \equiv \frac{\sqrt{n}(\beta_i - \beta_0)}{\sqrt{\widehat{\text{AVar}}(\hat{\beta}_i)}} = \frac{\beta_i - \beta_0}{\text{SE}^*(\hat{\beta}_i)} \overset{a}{\sim} N(0,1), \tag{80}$$

where

$$\text{SE}^*(\hat{\beta}_i) \equiv \sqrt{n^{-1}\widehat{\text{AVar}}(\hat{\beta}_i)} \equiv \sqrt{n^{-1}(\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1})_{ii}}.$$

The denominator in this t-statistic, $\text{SE}^*(\hat{\beta}_i)$, is called the heteroskedasticity-consistent standard standard error, or White's standard error. The reason for this terminology is that the error term can be conditionally heteroskedastic; recall that we have no assumed conditional homoskedasticity (see Assumption (3)) to derive the asymptotic distribution of t_{β_i}. This t-statistic is called the robust t-statistic.

Given this t-statistic, testing the null hypothesis at a significance level of α is as follows:

1. Calculate t_{β_i} using (80).
2. Look up the table of $N(0,1)$ to find the critical value of $z_{\alpha/2}$ which leaves $\alpha/2$ to the upper tail of the standard normal distribution. e.g., If $\alpha = 0.05$, $z_{\alpha/2} = 1.96$.
3. Do not reject the null hypothesis if $|t_{\beta_i}| < z_{\alpha/2}$; else reject.

The differences from the finite-sample t-test are: i) the way the standard error is calculated is different, ii) we use the table of $N(0,1)$ rather than that of $t(n-k)$, and iii) the actual size or exact size of the test (the probability of a Type I error given the sample size) equals the nominal size (i.e., the desired significance level α) only approximately, although the approximation becomes arbitrarily good as the sample size increases. The difference between the exact size and nominal size of a test is called the size distortion. Since t_{β_1} is asymptotically standard normal, the size of the distortion of the t-test converges to zero as $n \to \infty$.

4.2.8 Asymptotic F-test

Now suppose we wish to apply multiple linear restrictions, such that the null hypothesis is

$$H_0 : \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{r},$$

where $\mathbf{R}$ is an $r \times k$ matrix (where r is the dimension of $\mathbf{r}$, the number of restrictions) of full row rank. Then the F-test statistic is

$$W = n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\mathbf{R}\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) = rF \overset{a}{\sim} \chi^2(r). \tag{81}$$

I've used W to denote the test statistic as this formation of the F-test statistic is often referred to as a Wald test statistic. I will expand on what this means later, but for now let's just focus on proving that W is asymptotically distributed as $\chi^2(r)$.

We can first take the expression from (81) and distribute the n term:

$$W = \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top \left[\mathbf{R}\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1} \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}),$$

and notice that under the null, the first and third product terms on the RHS are (ignore the transpose for now):

$$\mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

and we know that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \text{AVar}(\hat{\boldsymbol{\beta}}))$, so by Slutsky's Theorem we can write

$$\mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top).$$

Also, we can say that

$$\left[\mathbf{R}\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1} \xrightarrow{p} \left[\mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1},$$

and remember that this inverse exists since $\mathbf{R}$ is of full rank and $\text{AVar}(\hat{\boldsymbol{\beta}})$ is positive definite. So, to collect what we know:

$$W = \underbrace{\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top}_{\xrightarrow{d} N(\mathbf{0}, \mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top)} \underbrace{\left[\mathbf{R}\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1}}_{\xrightarrow{p} \left[\mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top\right]^{-1}} \underbrace{\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}_{\xrightarrow{d} N(\mathbf{0}, \mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top)}.$$

Looks horrendous, but we can make use of a theorem pertaining to quadratic forms in normal vectors.

Theorem (4.1 in Davidson and MacKinnon (2004)):

1. If the m-vector $\mathbf{x}$ is distributed as $N(\mathbf{0}, \mathbf{\Omega})$, then the quadratic form $\mathbf{x}^\top \mathbf{\Omega} \mathbf{x}$ is distributed as $\chi^2(m)$.

2. If $\mathbf{P}$ is a projection matrix with rank r and $\mathbf{z}$ is an n-vector that is distributed as $N(\mathbf{0}, \mathbf{I})$, then the quadratic form $\mathbf{z}^\top \mathbf{P} \mathbf{z}$ is distributed as $\chi^2(r)$.

It then follows that since the column vector $\mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top)$ and since $\mathbf{R}\text{AVar}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top = \text{Var}\left(\mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right)$, we have

$$W \xrightarrow{d} \chi^2(r). \quad \blacksquare$$

The χ^2 statistic W is a Wald statistic because it is based on unrestricted estimates ($\hat{\boldsymbol{\beta}}$ and $\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})$ here) not constrained by the null hypothesis H_0. Testing H_0 at a significance level of α proceeds as follows.

1. Calculate W by (81).

2. Look up the table of the$\chi^2(r)$ distribution to find the critical value $\chi^2_\alpha(r)$ distribution.

3. If $W < \chi^2_\alpha(r)$, then do not reject H_0; else reject.

The probability of a Type I error approaches α as $n \to \infty$. As will be made clear later, this Wald statistic is closely related to the familiar F-test under conditional homoskedasticity.

4.2.9 Estimating $\mathbb{E}[u_t^2 \mathbf{X}_t^\top \mathbf{X}_t]$ consistently

The theory developed so far presumes that there is available a consistent estimator, $\hat{\mathbf{S}}$, of $\mathbf{S} = \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top] \equiv \mathbb{E}[u_t^2 \mathbf{X}_t^\top \mathbf{X}_t]$ to be used to calculate the estimated asymptotic variance, $\widehat{\text{AVar}}(\hat{\boldsymbol{\beta}})$. This section explains how to obtain $\hat{\mathbf{S}}$ from the sample $(\mathbf{y}, \mathbf{X})$.

If the error were observable, then the sample mean of $u_t^2 \mathbf{X}_t^\top \mathbf{X}_t$ is obviously consistent by ergodic stationarity. But we do not observe the error term, and the substitution of some consistent estimate of it results in

$$\hat{\mathbf{S}} = \frac{1}{n} \sum_{t=1}^{n} \hat{u}_t^2 \mathbf{X}_t^\top \mathbf{X}_t, \tag{82}$$

where $\hat{u}_t = y_t - \mathbf{X}_t^\top \hat{\boldsymbol{\beta}}$. For this estimator to be consistent for $\mathbf{S}$, we need to make a fourth-moment assumption about the regressors: we assume that $\mathbb{E}[(x_{ti}x_{tj})^2]$ exists and is finite for all $i, j = 1, 2, ..., k$. This leads us to the following proposition.

Proposition: Suppose the coefficient estimate $\hat{\boldsymbol{\beta}}$ used for calculating the residual $\hat{u}_t$ for $\hat{\mathbf{S}}$ in (82) is consistent, and suppose $\mathbf{S} = \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^\top]$ exists and is finite. Then, assuming that the the model is linear, that the DGP is ergodic stationary, and that $\mathbb{E}[(x_{ti}x_{tj})^2]$ exists and is finite, then $\hat{\mathbf{S}}$ is consistent for $\mathbf{S}$.

See why the fourth moment assumption is needed for the regressors, consider the case where $k = 1$ (only one regressor). So $\mathbf{X}_t$ is now simply a scalar X_t, $\mathbf{g}_t$ is a scalar $g_t = X_t u_t$, and so

$$\begin{aligned}\hat{u}_t &= y_t - X_t\hat{\beta} \\ &= y_t - X_t\beta - X_t(\hat{\beta} - \beta) \\ &= u_t - X_t(\hat{\beta} - \beta),\end{aligned}$$

and

$$\hat{u}_t^2 = u_t^2 - 2(\hat{\beta} - \beta)X_t u_t + (\hat{\beta} - \beta)^2 X_t^2. \tag{83}$$

By multiplying both sides by X_t^2 and summing over t,

$$\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2 X_t^2 - \frac{1}{n}\sum_{t=1}^{n} u_t^2 X_t^2 = -2(\hat{\beta} - \beta)\frac{1}{n}\sum_{t=1}^{n} u_t X_t^3 + (\hat{\beta} - \beta)^2 \frac{1}{n}\sum_{t=1}^{n} X_t^4. \tag{84}$$

Now we can see why the finite fourth-moment assumption on X_t is required: if the fourth moment $\mathbb{E}[X_t^4]$ is finite, then by ergodic stationarity the sample average of X_t^4 converges in probability to some finite number, so that the last term in (84) vanishes (converges to 0 in probability) if $\hat{\beta}$ is consistent for β.

We can sketch a data matrix representation of $\mathbf{S}$. If $\hat{\boldsymbol{\Omega}}$ is the $n \times n$ diagonal matrix whose t-th diagonal element is $\hat{u}_t^2$, then the $\hat{\mathbf{S}}$ in (82) can be represented in terms of data matrices as

$$\hat{\mathbf{S}} = \frac{1}{n}\mathbf{X}^\top \hat{\boldsymbol{\Omega}} \mathbf{X}, \tag{85}$$

with

$$\hat{\boldsymbol{\Omega}} = \begin{bmatrix} \hat{u}_1^2 & & \\ & \ddots & \\ & & \hat{u}_n^2 \end{bmatrix}.$$

So (78) can be written as

$$\begin{aligned} \widehat{\text{AVar}}(\hat{\boldsymbol{\beta}}) &= \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1} \hat{\mathbf{S}} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1} \\ &= \frac{1}{n} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Omega}} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \\ &= n(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Omega}} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

This form of the variance covariance is known as a "sandwich covariance matrix." In practice we ignore all the factors of n and use the matrix to directly estimate the variance covariance matrix of $\hat{\boldsymbol{\beta}}$.

4.3 Asymptotic theory with conditional homoskedasticity

In this section I use the notation in Davidson and MacKinnon (2004), and work with the model (47):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I}),$$

where the errors are now identically distributed with mean 0 and are now conditionally homoskedastic ($\mathbb{E}[u_t^2|\mathbf{X}] = \sigma_0^2$). We retain the assumption that the regressors are predetermined ($\mathbb{E}[u_t|\mathbf{X}_t] = 0$), and that the DGP allows us to assume

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \mathbf{X}^\top \mathbf{X} = \mathbf{S}_{\mathbf{X}^\top\mathbf{X}},$$

where, as before, $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$ is a finite, deterministic, positive definite matrix.

The key point of this section is to really just highlight how important model assumptions (and notation) are. For instance, Hayashi (2000) and Davidson and MacKinnon (2004) treat the presentation and importance of the sandwich covariance matrix very differently. Hayashi derives the asymptotic properties of the OLS estimator and its distribution by assuming conditional heteroskedasticity and

unconditional homoskedasticity from the outset (see Section 4.2.4). So he naturally ends up with a sandwich form for the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. On the other hand, Davidson and MacKinnon do not allow for conditional heteroskedasticity when deriving the OLS estimator, and only include conditional heteroskedasticity as a special case when discussing heteroskedasticity-consistent covariance matrix estimation (HCCME).

With all that said, let's look at the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ with conditional homoskedasticity. Our conjecture is that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{\sim} N(\mathbf{0}, \text{AVar}(\hat{\boldsymbol{\beta}})),$$

where the asymptotic variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\text{AVar}(\hat{\boldsymbol{\beta}}) = \sigma_0^2 \mathbf{S}_{\mathbf{X}^\top \mathbf{X}}^{-1}$. Begin by writing the sampling error,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u},$$

then distribute the appropriate n's through the equation, and then multiply both the LHS and RHS by $\sqrt{n}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \mathbf{u} \\ \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}^\top \mathbf{u}. \end{aligned}$$

Now, notice that:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \underbrace{\left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1}}_{\xrightarrow{p} \mathbf{S}_{\mathbf{X}^\top \mathbf{X}}^{-1}} \underbrace{\frac{1}{\sqrt{n}} \mathbf{X}^\top \mathbf{u}}_{\xrightarrow{d} N(\mathbf{0}, \sigma_0^2 \mathbf{S}_{\mathbf{X}^\top \mathbf{X}})}.$$

Before continuing, one thing to be careful of is the dimension of the variance covariance matrix. If we note the second term on the RHS as $\mathbf{v}$:

$$\mathbf{v} \equiv \frac{1}{\sqrt{n}} \mathbf{X}^\top \mathbf{u} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \mathbf{X}_t^\top, \tag{86}$$

and remember we assume that the regressors are predetermined so $\mathbb{E}[u_t | \mathbf{X}_t] = 0 \implies \mathbb{E}[u_t \mathbf{X}_t^\top] = \mathbf{0}$.

Due to this assumption, we can apply a CLT to $\mathbf{v}$, which tells us that

$$\mathbf{v} \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \text{Var}(u_t \mathbf{X}_t^\top)\right) = N\left(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[u_t^2 \mathbf{X}_t^\top \mathbf{X}_t\right]\right),$$

and note that because $\mathbf{X}_t$ is a $1 \times k$ row vector, the covariance matrix here is $k \times k$, as it must be for conformity (and since we want the variance covariance matrix associated with the parameter vector, which is k-dimensional). It follows that, asymptotically,

$$\text{Var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) = \sigma_0^2 \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}. \tag{87}$$

Moreover, since $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is, asymptotically, just a deterministic linear combination of the components of the multivariate normal random vector, $n^{-1/2}\mathbf{X}^\top\mathbf{u}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{\sim} N\left(\mathbf{0}, \sigma_0^2 \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}^{-1}\right). \quad \blacksquare \tag{88}$$

As a bonus, we can write the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ itself as

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left(\boldsymbol{\beta}_0, \sigma_0^2 (\mathbf{X}^\top\mathbf{X})^{-1}\right). \tag{89}$$

There is another thing to be aware of here: Davidson and MacKinnon often use the term $\text{Var}(\hat{\boldsymbol{\beta}})$ to denote the variance of the distribution of the estimator itself, as in (89). For the sake of consistency in these notes, I am going to use the notation we've used thus far: when I use $\text{Var}(\hat{\boldsymbol{\beta}})$ or $\text{AVar}(\hat{\boldsymbol{\beta}})$, I refer to the variance of the asymptotic distribution of the vector as in (88).

One benefit of writing (89) is that we can see why $\hat{\boldsymbol{\beta}}$ is referred to as being $\sqrt{n}$-consistent. Remember that when $n^{-1}\mathbf{X}^\top\mathbf{X} \xrightarrow{p} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$ as $n \to \infty$, the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$, without the factor of n^{-1}, simply tends to a zero matrix. The rate of convergence of $\hat{\boldsymbol{\beta}}$ to its probability limit, $\boldsymbol{\beta}_0$, is $\sqrt{n}$, as multiplying (88) by $n^{-1/2}$ will result in an expression with zero mean and finite covariance matrix.

Next, we will look at the t- and F-tests, this time with conditional homoskedasticity. Additionally, this will give us a chance to use orthogonal projections to set up the test statistics, avoiding tedious

algebra.

4.3.1 Asymptotic t-test (again)

First, rewrite the model so that we isolate the coefficient of interest as a scalar:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \beta_2\mathbf{x}_2 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2\mathbf{I}),$$

where $\mathbf{X}_1$ is $n \times (k-1)$, $\boldsymbol{\beta}_1$ is a $(k-1)$-vector, and $\mathbf{x}_2$ is an n-vector, so we have $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{x}_2\end{bmatrix}$. Now, we can state our null hypothesis,

$$H_0 : \beta_2 = 0.$$

By the FWL Theorem, the OLS estimate of β_2, $\hat{\beta}_2$, can be attained from the following regression:

$$\mathbf{M}_1\mathbf{y} = \beta_2\mathbf{M}_1\mathbf{x}_2 + \mathbf{M}_1\mathbf{u},$$

where $\mathbf{M}_1$ is the elimination matrix $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top$ which projects on to $\mathcal{S}^\perp(\mathbf{X}_1)$. Using some matrix algebra, we can write the expression for our estimates as

$$\begin{aligned}\hat{\beta}_2 &= \frac{\mathbf{x}_2^\top\mathbf{M}_1\mathbf{y}}{\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2},\\ \mathbb{E}\left[(\hat{\beta}_2 - \beta_2)^2\right] &= \sigma_0^2(\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2)^{-1},\end{aligned}$$

where we can write them in this form as they're scalar products. In order to test the null hypothesis, we have to minus β_2 from $\hat{\beta}_2$ and divide by the square root of the variance. So this will give us (noting that our null is that $\beta_2 = 0$):

$$\begin{aligned}\frac{\hat{\beta}_2 - \beta_2}{\sigma_0(\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2)^{-1/2}} &= \frac{\mathbf{x}_2^\top\mathbf{M}_1\mathbf{y}}{\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2} \div \sigma_0(\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2)^{-1/2},\\ &= \frac{\mathbf{x}_2^\top\mathbf{M}_1\mathbf{y}}{\sigma_0(\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2)^{1/2}}.\end{aligned}$$

Of course, in practice, we don't know what σ_0 is, so we replace it by its estimate, s. But the further trick, and why projection matrices are so convenient, is to notice that we can write estimated variance as

$$s^2 = \frac{\mathbf{u}^\top \mathbf{u}}{n-k} = \frac{\mathbf{y}^\top \mathbf{M}_\mathbf{X} \mathbf{y}}{n-k},$$

leading to

$$\begin{aligned} t_{\beta_2} &= \frac{\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}}{s(\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{1/2}} \\ &= \left(\frac{\mathbf{y}^\top \mathbf{M}_\mathbf{X} \mathbf{y}}{n-k} \right)^{-1/2} \frac{\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}}{(\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{1/2}}, \end{aligned}$$

where we know that s^2 is a consistent estimate for σ_0^2 (see Equation (79)). Now, in order to prove that t_{β_2} has an asymptotic distribution, we need to write it as a function of quantities to which we can apply either a LLN or CLT. Multiplying the numerator and denominator by $n^{-1/2}$ gives us

$$t_{\beta_2} = \left(\frac{\mathbf{y}^\top \mathbf{M}_\mathbf{X} \mathbf{y}}{n-k} \right)^{-1/2} \frac{n^{-1/2} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}}{(n^{-1} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{1/2}}. \tag{90}$$

Now, a few things stick out from (90): first, as stated, s^2 is consistent as it is $n/(n-k)$ multiplied by the average of the $\hat{u}_t^2$, and $\operatorname{plim} n/(n-k) = 1$, and the average of the $\hat{u}_t^2$ tends to σ_0^2 by a LLN. So the first factor on the LHS of (90) tends to $1/\sigma_0$ as $n \to \infty$. When the data is generated under the null ($\beta_2 = 0$), we have that $\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{u}$, and so (90) is asymptotically equivalent to

$$\frac{n^{-1/2} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{u}}{\sigma_0 (n^{-1} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{1/2}}. \tag{91}$$

If we assumed that the regressors were exogenous, then proving that $t_{\beta_2} \overset{a}{\sim} N(0,1)$ would be a piece of cake. Let's just assume this briefly, just for illustrative purposes. With exogenous regressors, we can now apply conditional expectations, which means that the only part of (91) that is treated as random is $\mathbf{u}$. The numerator is $n^{-1/2}$ times a weighted sum of the u_t, each of which has mean 0, and the

conditional variance of this weighted sum is:

$$\mathbb{E}\left[\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{u}\mathbf{u}^\top \mathbf{M}_1 \mathbf{x}_2 | \mathbf{X}\right] = \sigma_0^2 \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}.$$

Thus, (91) evidently has mean 0 and variance 1, conditional on $\mathbf{X}$. But since 0 and 1 do not depend on $\mathbf{X}$, these are also the unconditional mean and variance of (91). Provided that we can apply a CLT to the numerator of (91), the numerator of t_{β_2} must be asymptotically normally distributed, and we can conclude that, under the null hypothesis and with exogenous regressors,

$$t_{\beta_2} \overset{a}{\sim} N(0,1).$$

If we tighten our assumption, and go back to predetermined regressors, then much like we did before in (86), start by applying a CLT to the k-vector

$$\mathbf{v} \equiv \frac{1}{\sqrt{n}} \mathbf{X}^\top \mathbf{u} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \mathbf{X}_t^\top.$$

Since we assume $\mathbb{E}[u_t | \mathbf{X}_t] = 0$, this implies that $\mathbb{E}[\mathbf{X}_t^\top u_t] = \mathbf{0}$, as required for the CLT, which then tells us that

$$\mathbf{v} \xrightarrow{d} N\left(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{Var}(u_t \mathbf{X}_t^\top)\right) = N\left(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[u_t^2 \mathbf{X}_t^\top \mathbf{X}_t\right]\right),$$

where we once again carefully note that because $\mathbf{X}_t$ is a $1 \times k$ row vector, the variance covariance matrix here is $k \times k$. Due to our assumption of conditional homoskedasticity, $\mathbb{E}[u_t^2 | \mathbf{X}_t] = \sigma_0^2$, we can

simplify the limiting covariance matrix:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[u_t^2\mathbf{X}_t^\top\mathbf{X}_t\right] &= \lim_{n\to\infty}\sigma_0^2\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}[\mathbf{X}_t^\top\mathbf{X}_t]\\ &= \sigma_0^2\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t^\top\mathbf{X}_t\\ &= \sigma_0^2\operatorname*{plim}_{n\to\infty}\mathbf{X}^\top\mathbf{X}\\ &= \sigma_0^2\mathbf{S}_{\mathbf{X}^\top\mathbf{X}},\end{aligned}$$

where we applied an LLN in reverse to go from the first line to the second. Notice that we applied this result in (87).

Looking at (91), we can write the numerator as

$$n^{-1/2}\mathbf{x}_2^\top\mathbf{u} - n^{-1/2}\mathbf{x}_2^\top\mathbf{P}_1\mathbf{u}. \tag{92}$$

Since $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{x}_2\end{bmatrix}$, such that $\mathbf{x}_2$ is the k-th column of $\mathbf{X}$, the first term in (92) is simply the k-th element in $\mathbf{v}$, which we denote as v_2. The second term is

$$\begin{aligned}n^{-1/2}\mathbf{x}_2^\top\mathbf{P}_1\mathbf{u} &= n^{-1/2}\mathbf{x}_2^\top\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{u}\\ &= \underbrace{n^{-1}\mathbf{x}_2^\top\mathbf{X}_1}_{\xrightarrow{p}\mathbf{S}_{21}}\underbrace{(n^{-1}\mathbf{X}_1^\top\mathbf{X}_1)^{-1}}_{\xrightarrow{p}\mathbf{S}_{11}^{-1}}n^{-1/2}\mathbf{X}_1^\top\mathbf{u},\end{aligned}$$

where $\mathbf{S}_{21}$ and $\mathbf{S}_{11}$ are submatrices of $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$. Thus, the term of interest for us the third term, as it's the only factor that remains random as $n\to\infty$. It is just a subvector $\mathbf{v}$ consisting of the first $k-1$ components, which we denote as $\mathbf{v}_1$. Asymptotically, in partitioned matrix notation, (92) becomes

$$v_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{v}_1 = \begin{bmatrix}-\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1\end{bmatrix}\underbrace{\begin{bmatrix}\mathbf{v}_1\\ v_2\end{bmatrix}}_{\mathbf{v}}.$$

Since $\mathbf{v}$ is asymptotically multivariate normal, by Slutsky's Theorem, this scalar expression is asymp-

totically normal, with mean zero and variance

$$\sigma_0^2 \begin{bmatrix} -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1 \end{bmatrix} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}} \begin{bmatrix} -\mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ 1 \end{bmatrix},$$

where, since $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$ is symmetric, $\mathbf{S}_{12}$ is just the transpose of $\mathbf{S}_{21}$. If we now write $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$ as a partitioned matrix, the variance of (92) is seen to be

$$\sigma_0^2 \begin{bmatrix} -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} -\mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ 1 \end{bmatrix} = \sigma_0^2(\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}). \tag{93}$$

The denominator of (91) is easier to deal with. Ignoring the σ_0^2, the square of the second term in the denominator is

$$\begin{aligned} n^{-1}\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2 &= n^{-1}\mathbf{x}_2^\top\mathbf{x}_2 - n^{-1}\mathbf{x}_2^\top\mathbf{P}_1\mathbf{x}_2 \\ &= n^{-1}\mathbf{x}_2^\top\mathbf{x}_2 - n^{-1}\mathbf{x}_2^\top\mathbf{X}_1(n^{-1}\mathbf{X}_1^\top\mathbf{X}_1)^{-1}n^{-1}\mathbf{X}_1^\top\mathbf{x}_2. \end{aligned}$$

In the limit, all the pieces of this expression become submatrices of $\mathbf{S}_{\mathbf{X}^\top\mathbf{X}}$, and we find

$$n^{-1}\mathbf{x}_2^\top\mathbf{M}_1\mathbf{x}_2 = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}.$$

When this is multiplied by σ_0^2, it's nothing but (93), the variance of the numerator of expression (91). Thus, asymptotically, we have shown that t_{β_2} is the ratio of a normal random variable with mean zero the standard deviation of that random variable. Consequently, we have established that, under the null hypothesis, with regressors that are not exogenous but merely predetermined,

$$t_{\beta_2} \overset{a}{\sim} N(0,1),$$

which is what we had for the exogenous regressors case.

4.3.2 Asymptotic F-test (again)

Now, let's look at the F-test, starting by stating our model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{X}_1$ is $n \times k_1$, $\mathbf{X}_2$ is $n \times k_2$, $k = k_1 + k_2$,and, again, we have $\mathbb{E}[u_t|\mathbf{X}_t] = 0$ and $\mathbb{E}[u_t^2|\mathbf{X}_t] = \sigma_0^2$, and our hypothesis is

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

Under the null, $F_{\boldsymbol{\beta}_2}$ is equal to expression (62):

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{\epsilon}^\top \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2^\top \mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^\top \mathbf{M}_1\boldsymbol{\epsilon}/r}{\boldsymbol{\epsilon}^\top \mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}/(n-k)},$$

where $\boldsymbol{\epsilon} \equiv \mathbf{u}/\sigma_0$ and $r = k_2$, the dimension of $\boldsymbol{\beta}_2$. As usual with asymptotic theory, we dispense some select quantities of n so that we can apply some LLNs and CLTs

$$F_{\beta_2} = \frac{\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^\top \mathbf{M}_1\mathbf{X}_2 \left(\frac{1}{n}\mathbf{X}_2^\top \mathbf{M}_1\mathbf{X}_2\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{X}_2^\top \mathbf{M}_1\boldsymbol{\epsilon}/r}{\boldsymbol{\epsilon}^\top \mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}/(n-k)}. \tag{94}$$

It is not hard to use the results we obtained for the t statistic to show that, as $n \to \infty$,

$$rF_{\boldsymbol{\beta}_2} \overset{a}{\sim} \chi^2(r), \tag{95}$$

under the null hypothesis. Since $1/r$ times a random variable t hat follows the $\chi^2(r)$ distribution is distributed as $F(r, \infty)$, we can also conclude that

$$F_{\boldsymbol{\beta}_2} \overset{a}{\sim} F(r, n-k).$$

Proof: Begin with the denominator of (94), $\boldsymbol{\epsilon}^\top \mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}$, where each element of $\boldsymbol{\epsilon}$ has mean 0 and variance 1 and is independent of the others. Under the normality assumption $\boldsymbol{\epsilon}^\top \mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}$ is distributed

as $\chi^2(n-k)$, and so it has mean $n-k$ (convergence in probability due to weak LLN).[16] Applying an LLN, we see that the denominator must tend to 1 asymptotically.

The numerator of the F statistic, multiplied by r, is

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^\top \mathbf{M}_1\mathbf{X}_2\left(\frac{1}{n}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2\right)^{-1}\frac{1}{\sqrt{n}}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon}. \tag{96}$$

Let $\mathbf{q} \equiv n^{-1/2}\mathbf{X}^\top\boldsymbol{\epsilon}$. Then a CLT shows that, asymptotically,

$$\mathbf{q} \overset{a}{\sim} N(\mathbf{0}, \mathbf{S}_{\mathbf{X}^\top\mathbf{X}}),$$

similar to the discussion of the t-statistic. If we partition $\mathbf{q}$, conformably with the partition of $\mathbf{X}$, into two subvectors $\mathbf{q}_1$ and $\mathbf{q}_2$, we have

$$\begin{aligned} n^{-1/2}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon} &= n^{-1/2}\mathbf{X}_2^\top\boldsymbol{\epsilon} - n^{-1/2}\mathbf{X}_2\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\boldsymbol{\epsilon} \\ &= n^{-1/2}\mathbf{X}_2^\top\boldsymbol{\epsilon} - n^{-1}\mathbf{X}_2\mathbf{X}_1(n^{-1}\mathbf{X}_1^\top\mathbf{X}_1)^{-1}n^{-1/2}\mathbf{X}_1^\top\boldsymbol{\epsilon} \\ \implies \operatorname*{plim}_{n\to\infty} n^{-1/2}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon} &= \mathbf{q}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{q}_1, \end{aligned} \tag{97}$$

and we can write this as

$$\begin{bmatrix} -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1 \end{bmatrix}\begin{bmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \end{bmatrix}.$$

Since $\mathbf{q}$ is asymptotically multivariate normal, this scalar expression is asymptotic normal, with mean

[16] Recall the following theorem:

Theorem :

1. If the m-vector $\mathbf{x}$ is distributed as $N(\mathbf{0}, \boldsymbol{\Omega})$, then the quadratic form $\mathbf{x}^\top\boldsymbol{\Omega}\mathbf{x}$ is distributed as $\chi^2(m)$.
2. If $\mathbf{P}$ is a projection matrix with rank r and $\mathbf{z}$ is an n-vector that is distributed as $N(\mathbf{0}, \mathbf{I})$, then the quadratic form $\mathbf{z}^\top\mathbf{P}\mathbf{z}$ is distributed as $\chi^2(r)$.

Also recall that $\mathbb{E}[\boldsymbol{\epsilon}^\top\mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}] = \mathrm{Tr}(\mathbf{M}_\mathbf{X}) = n-k$.

zero and variance

$$\begin{bmatrix} -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1 \end{bmatrix} \mathbf{S}_{\mathbf{X}^\top\mathbf{X}} \begin{bmatrix} -\mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ 1 \end{bmatrix} = \begin{bmatrix} -\mathbf{S}_{21}\mathbf{S}_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} -\mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ 1 \end{bmatrix}$$
$$= \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}. \tag{98}$$

The middle term in the numerator (94) is

$$\begin{aligned} \frac{1}{n}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2 &= n^{-1}\mathbf{X}_2^\top\mathbf{X}_2 - n^{-1}\mathbf{X}_2\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2 \\ &= n^{-1}\mathbf{X}_2^\top\mathbf{X}_2 - n^{-1}\mathbf{X}_2\mathbf{X}_1(n^{-1}\mathbf{X}_1^\top\mathbf{X}_1)^{-1}n^{-1}\mathbf{X}_1^\top\mathbf{X}_2 \\ \implies \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2 &= \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}. \end{aligned} \tag{99}$$

Thus, asymptotically, (96) is a quadratic form in the normal r-vector $\mathbf{q}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{q}_1$ and the inverse of the covariance matrix of this vector.

Collecting our information, we have

$$rF_{\beta_2} = \frac{\underbrace{\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}^\top\mathbf{M}_1\mathbf{X}_2}_{\xrightarrow{d} N(\mathbf{0},\mathbf{S}_{22}-\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12})} \underbrace{\left(\frac{1}{n}\mathbf{X}_2^\top\mathbf{M}_1\mathbf{X}_2\right)^{-1}}_{\xrightarrow{p}\left(\mathbf{S}_{22}-\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\right)^{-1}} \underbrace{\frac{1}{\sqrt{n}}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon}}_{\xrightarrow{d} N(\mathbf{0},\mathbf{S}_{22}-\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12})}}{\underbrace{\boldsymbol{\epsilon}^\top\mathbf{M}_\mathbf{X}\boldsymbol{\epsilon}/(n-k)}_{\xrightarrow{p}1}}.$$

That's rough. But, the numerator is a quadratic with r-vector $n^{-1/2}\mathbf{X}_2^\top\mathbf{M}_1\boldsymbol{\epsilon}$ which is asymptotically distributed as $N(\mathbf{0}, \mathbf{S}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12})$ and the asymptotic variance of the r-vector. Therefore, as the numerator is 1 in the limit, $rF_{\boldsymbol{\beta}_2}$ is asymptotically distributed as $\chi^2(r)$. ■

References

Davidson, R. and MacKinnon, J. G. (2004), *Econometric Theory and Methods* (Oxford University Press).

Hayashi, F. (2000), *Econometrics* (Princeton University Press).

Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd Edition, John Wiley and Sons).

Turkington, D. A. (2007), *Mathematical Tools for Economics* (Blackwell Publishing).

Turkington, D. A. (2013), *Generalized Vectorization, Cross-Products, and Matrix Calculus* (Cambridge University Press).