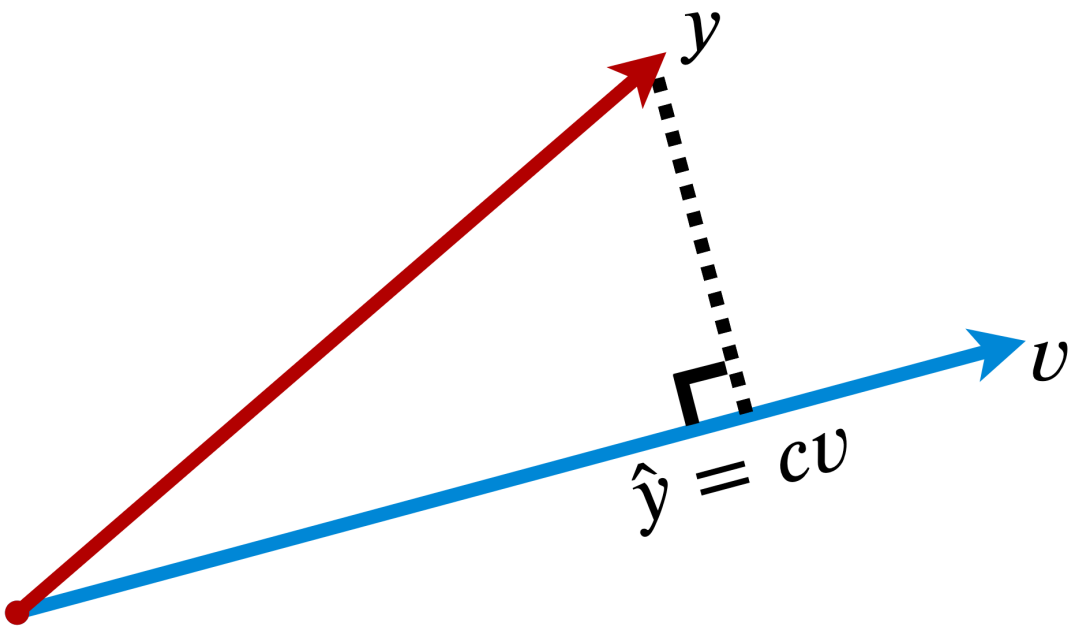
10 Fundamental Theorems for Econometrics

Thomas S. Robinson (https://ts-robinson.com)

2020-08-13 – v0.1

Preface



A list of the 10 most important econometric theorems was circulated on Twitter, based on what Jeffrey Wooldridge claims are the bases for most econometrics. As a political scientist with applied statistics training, this list caught my attention because it contains many of the theorems I see used in (methods) papers, but which I typically glaze over for lack of understanding. The complete list (slightly paraphrased) is:

1. Law of Iterated Expectations, Law of Total Variance
2. Linearity of Expectations, Variance of a Sum
3. Jensen's Inequality, Chebyshev's Inequality
4. Linear Projection and its Properties
5. Weak Law of Large Numbers, Central Limit Theorem
6. Slutsky's Theorem, Continuous Convergence Theorem, Asymptotic Equivalence Lemma
7. Big Op, Little op, and the algebra of them
8. Delta Method
9. Frisch-Waugh Partialling Out
10. For PD matrices A and B, A-B is PSD if and only if $B^{-1} - A^{-1}$ is PSD.

As an exercise in improving my own knowledge of these fundamentals, I decided to work through each theorem – using various lecture notes found online, and excellent textbooks like Aronow & Miller's (2019) Foundations of Agnostic Statistics, Angrist and Pischke's (2008) Mostly Harmless Econometrics, and Wasserman's (2004)

All of Statistics.

I found for a list of important theorems there were few consistent sources that contained explanations and proofs of each item. Often, textbooks had excellent descriptive intuitions but would hold back on offering full, annotated proofs. Or full proofs were offered without explaining the wider significance of the theorems. Some of the concepts, moreover, had different definitions dependent on the field or source of the proof (like Slutsky's Theorems)!

This resource is an attempt to collate my writing on these theorems – the intuitions, proofs, and examples – into a single document. I have taken some liberties in doing so – for instance combining Wooldridge's first two points into a single chapter on 'Expectation Theorems', and often omit continuous proofs where discrete proofs are similar and easier to follow. That said, I have tried to be reasonably exhaustive in my proofs so that they are accessible to those (like me) without a formal statistics background.

The inspiration for this project was Jeffrey Wooldridge's list, an academic whose work I admire greatly. This document, however, is in no way endorsed by or associated with him. Most of the applied examples (and invisible corrections to my maths) stem from discussions with Andy Eggers and Musashi Harukawa. There will inevitably still be some errors, omissions, and confusing passages. I would be more than grateful to receive any feedback at thomas.robinson@durham.ac.uk or via the GitHub repo for this project.

Prerequisites

I worked through these proofs learning the bits of maths I needed as I went along. For those who want to consult Google a little less than I had to, the following should ease you into the more formal aspects of this document:

- A simple working knowledge of probability theory
- The basics of expectation notation, but you don't need to know any expectation rules (I cover the important ones in Chapter 1).
- A basic understanding of linear algebra (i.e. how you multiply matrices, what transposition is, and what the identity matrix looks like). More complicated aspects like eigenvalues and Gaussian elimination make fleeting appearances, particularly in Chapter 9, but these are not crucial.
- Where relevant, I provide coded examples in R. I've kept my use of packages to a minimum so the code should be reasonably easy to read/port to other programming languages.

Version notes

v0.1

This is the first complete draft, and some sections are likely to be changed in future versions. For instance, in Chapter 9 I would like to provide a more comprehensive overview of quadratic form in linear algebra, how we derive gradients, and hence the shape of PD matrices. Again, any suggestions on ways to improve/add to this resource are very much welcome!

Chapter 1

Expectation Theorems

This chapter sets out some of the basic theorems that can be derived from the definition of expectations, as highlighted by Wooldridge. I have combined his first two points into a single overview of expectation maths. The theorems themselves are not as immediately relevant to applied research as some of the later theorems on Wooldridge's list. However, they often form the fundamental basis upon which future proofs are conducted.

1.1 Law of Iterated Expectations

The Law of Iterated Expectations (LIE) states that:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] \tag{1.1}$$

In plain English, the expected value of X is equal to the expectation over the conditional expectation of X given Y. More simply, the mean of X is equal to a weighted mean of conditional means.

Aronow & Miller (2019) note that LIE is 'one of the most important theorems', because being able to express unconditional expectation functions in terms of conditional expectations allow you to hold some parameters fixed, making calculations more tractable.

1.1.1 Proof of LIE

First, we can express the expectation over conditional expectations as a weighted sum over all possible values of Y, and similarly express the conditional expectations using summation too:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y=y]P(Y=y) \tag{1.2}$$

$$= \sum_y \sum_x xP(X=x|Y=y)P(Y=y) \tag{1.3}$$

$$= \sum_y \sum_x xP(Y=y|X=x)P(X=x)., \tag{1.4}$$

Note that the final line follows due to Bayes' Rule.[1] And so:

$$... = \sum_y \sum_x xP(X = x)P(Y = y|X = x) \tag{1.5}$$
$$= \sum_x xP(X = x) \sum_y P(Y = y|X = x) \tag{1.6}$$
$$= \sum_x xP(X = x) \tag{1.7}$$
$$= \mathbb{E}[X] \quad \square \tag{1.8}$$

The last steps of the proof are reasonably simple. Equation 1.5 is a trivial rearrangement of terms. The second line follows since y does not appear in $xP(X = x)$ and so we can move the summation over Y to within the summation over X. The final line follows from the fact that the sum of the conditional probabilities $P(Y = y|X = x) = 1$ (by simple probability theory).

1.2 Law of Total Variance

The Law of Total Variance (LTV) states the following:

$$var[Y] = \mathbb{E}[var[Y|X]] + var(\mathbb{E}[Y|X]) \tag{1.9}$$

1.2.1 Proof of LTV

LTV can be proved almost immediately using LIE and the definition of variance:

$$var(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \tag{1.10}$$
$$= \mathbb{E}[\mathbb{E}[Y^2|X]] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \tag{1.11}$$
$$= \mathbb{E}[var[Y|X] + \mathbb{E}[Y]^2]] - \mathbb{E}[\mathbb{E}[Y|X]]^2 \tag{1.12}$$
$$= \mathbb{E}[var[Y|X]] + (\mathbb{E}[\mathbb{E}[Y]^2] - \mathbb{E}[\mathbb{E}[Y|X]]^2) \tag{1.13}$$
$$= \mathbb{E}[var[Y|X]] + var(\mathbb{E}[Y|X]) \quad \square \tag{1.14}$$

The second line applies LIE to both Y^2 and Y separately. Then we apply the definition of variance to $\mathbb{E}[Y^2|X]$, and subsequently decompose this term (since $\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$.

1.3 Linearity of Expectations

The Linearity of Expectations (LOE) simply states that:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \tag{1.15}$$

where a and b are real numbers, and X and Y are random variables.

[1] Bayes' Rule states $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. Therefore:

$$P(X = x|Y = y) \times P(Y = y) = \frac{P(Y = y|X = x)P(X = x)P(Y = y)}{P(Y = y)} = P(Y = y|X = x)P(X = x).$$

1.3.1 Proof of LOE

$$\mathbb{E}[aX + bY] = \sum_x \sum_y (ax + by)P(X = x, Y = y) \tag{1.16}$$
$$= \sum_x \sum_y axP(X = x, Y = y) + \sum_x \sum_y byP(X = x, Y = y) \tag{1.17}$$
$$= a\sum_x x \sum_y P(X = x, Y = y) + b\sum_y y \sum_x P(X = x, Y = y) \tag{1.18}$$
$$\tag{1.19}$$

The first line simply expands the expectation into summation form i.e. the expectation is the sum of $aX+bY$ for each (discrete) value of X and Y weighted by their joint probability. We then expand out these terms. Since summations are commutative, we can rearrange the order of the summations for each of the two parts in the final line, and shift the real numbers and random variables outside the various operators.

Now note that $\sum_i P(I = i, J = j) \equiv P(J = j)$ by probability theory. Therefore:

$$... = a\sum_x xP(X = x) + b\sum_y yP(Y = y) \tag{1.20}$$

The two terms within summations are just the weighted averages of X and Y respectively, i.e. the expectations of X and Y, so:

$$... = a\mathbb{E}[X] + b\mathbb{E}[Y] \quad \square \tag{1.21}$$
$$\tag{1.22}$$

1.4 Variance of a Sum

There are two versions of the Variance of a Sum (VOS) law:

- $var(X + Y) = var(X) + var(Y)$, when X and Y are independent
- $var(X + Y) = var(X) + var(Y) + 2Cov(X, Y)$, when X and Y are correlated

1.4.1 Proof of VoS: X, Y are independent

$$var(X + Y) = \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 \tag{1.23}$$
$$= \mathbb{E}[(X^2 + 2XY + Y^2)] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \tag{1.24}$$

The first line of the proof is simply the definition of variance. In the second line, we expand the equation in the first term and using LOE decompose the second term. We can expand this equation further, continuing to use LOE and noting that :

$$... = \mathbb{E}[X^2] + \mathbb{E}[2XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2) \tag{1.25}$$
$$= \mathbb{E}[X^2] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - \mathbb{E}[Y]^2 \tag{1.26}$$
$$= Var[X] + Var[Y] \quad \square \tag{1.27}$$

since $\mathbb{E}[A]\mathbb{E}[B] = \mathbb{E}[AB]$ when A and B are independent.

1.4.2 Proof of VoS: X, Y are dependent

As before, we can expand out the variance of a sum into its expected values:

$$var(X+Y) = \mathbb{E}[X^2] + \mathbb{E}[2XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2). \tag{1.28}$$

Since X and Y are assumed to be dependent, the non-squared terms do not necessarily cancel each other out anymore. Instead, we can rearrange as follows:

$$\begin{aligned} var(X+Y) &= var(X) + var(Y) + \mathbb{E}[2XY] - 2\mathbb{E}[X]\mathbb{E}[Y] && (1.29)\\ &= var(X) + var(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]), && (1.30) \end{aligned}$$

and note that $\mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B] = Cov(A, B)$:

$$... = var(X) + var(Y) + 2Cov(A, B) \quad \square \tag{1.31}$$

Two further points are worth noting. First, the independent version of the proof is just a special case of the dependent version of the proof. When X and Y are independent, the covariance between the two random variables is zero, and therefore the the variance of the sum is just equal to the sum of the variances.

Second, nothing in the above proofs rely on there being just two random variables. In fact, $var(\sum_i^n X_i) = \sum_i^n var(X_i)$ when all variables are independent from each other, and equal to $\sum_i^n var(X_i) + 2\sum_{1\leq i<j\leq n}^n Cov(X_i, X_j)$. This can be proved by induction using the above proofs, but intuitively: we can replace, for example, Y with $Y = (Y_1 + Y_2)$ and iteratively apply the above proof first to $X + Y$ and then subsequently expand $var(Y)$ as $var(Y_1 + Y_2)$.

Chapter 2

Inequalities involving expectations

This chapter discusses and proves two inequalities that Wooldridge highlights - Jensen's and Chebyshev's. Both involve expectations (and the theorems derived in the previous chapter).

2.1 Jensen's Inequality

Jensen's Inequality is a statement about the relative size of the expectation of a function compared with the function over that expectation (with respect to some random variable). To understand the mechanics, I first define convex functions and then walkthrough the logic behind the inequality itself.

2.1.1 Convex functions

A function f is convex (in two dimensions) if all points on a straight line connecting any two points on the graph of f is above or on that graph. More formally, f is convex if for $\forall x_1, x_2 \in \mathbb{R}$, and $\forall t \in [0, 1]$:

$$f(tx_1, (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

Here, t is a weighting parameter that allows us to range over the full interval between points x_1 and x_2.

Note also that concave functions are defined as the opposite of convex functions i.e. a function h is concave if and only if $-h$ is convex.

2.1.2 The Inequality

Jensen's Inequality (JI) states that, for a convex function g and random variable X:

$$\mathbb{E}[g(X)] \geq g(E[X])$$

This inequality is exceptionally general – it holds for any convex function. Moreover, given that concave functions are defined as negative convex functions, it is easy to see that JI also implies that if h is a concave function, $h(\mathbb{E}[X]) \geq \mathbb{E}[h(X)]$.[1]

Interestingly, note the similarity between this inequality and the definition of variance in terms of expectations:

[1] Since $-h(x)$ is convex, $\mathbb{E}[-h(X)] \geq -h(\mathbb{E}[X])$ by JI. Hence, $h(\mathbb{E}[X]) - \mathbb{E}[h(X)] \geq 0$ and so $h(\mathbb{E}[X]) \geq \mathbb{E}[h(X)]$.

$$var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2,$$

and since $var(X)$ is always positive:

$$\mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$$
$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2).$$

We can therefore define $g(X) = X^2$ (a convex function), and see that variance itself is an instance of Jensen's Inequality.

2.1.3 Proof

Assume $g(X)$ is a convex function, and $L(X) = a + bX$ is a linear function tangential to $g(X)$ at point $\mathbb{E}[X]$.

Hence, since g is convex and L is tangential to g, we know by definition that:

$$g(x) \geq L(x), \forall x \in X. \tag{2.1}$$

So, therefore:

$$\begin{aligned}
\mathbb{E}[g(X)] &\geq \mathbb{E}[L(X)] && (2.2)\\
&\geq \mathbb{E}[A + bX] && (2.3)\\
&\geq a + b\mathbb{E}[X] && (2.4)\\
&\geq L(\mathbb{E}[X]) && (2.5)\\
&\geq g(\mathbb{E}[X]) \quad \square && (2.6)
\end{aligned}$$

The majority of this proof is straightforward. If one function is always greater than or equal to another function, then the unconditional expectation of the first function must be at least as big as that of the second. The interior lines of the proof follow from the definition of L, the linearity of expectations, and another application of the definition of L respectively.

The final line then follows because, by the definition of the straight line L, we know that $L[\mathbb{E}[X]]$ is tangential with g at $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] = g(\mathbb{E}[X])$.[2]

2.1.4 Application

In Chapter 2 of Agnostic Statistics (2019), the authors note (almost in passing) that the standard error of the mean is not unbiased, i.e. that $\mathbb{E}[\hat{\sigma}] \neq \sigma$, even though it is consistent i.e. that $\hat{\sigma} \xrightarrow{p} \sigma$. The bias of the mean's standard error is somewhat interesting (if not surprising), given how frequently we deploy the standard error (and, in a more general sense, highlights how important asymptotics are not just for the estimation of parameters, but also those parameters' uncertainty). The proof of why $\hat{\sigma}$ is biased also, conveniently for this chapter, uses Jensen's Inequality.

The standard error of the mean is denoted as

$$\sigma = \sqrt{V(\bar{X})}$$

[2]Based on lecture notes by Larry Wasserman.

,
where $V(\bar{X}) = \frac{V(X)}{n}$.

Our best estimate of this quantity $\hat{\sigma} = \sqrt{\hat{V}(\bar{X})}$ is simply the square root of the sample variance estimator. We know that the variance estimator itself is unbiased and a consistent estimator of the sampling variance (see Agnostic Statistics Theorem 2.1.9).

The bias in the estimate of the sample mean's standard error originates from the square root function Note that the square root is a strictly concave function. This means we can make two claims about the estimator. First, as with any concave function we can use the inverse version of Jensen's Inequality, i.e. that $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$. Second, since the square root is a strictly concave function, we can use the weaker "less than or equal to" operator with the strict "less than" inequality. Hence, the proof is reasonably easy:

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}\right] = \mathbb{E}\left[\sqrt{\hat{V}(\bar{X})}\right] &< \sqrt{\mathbb{E}[\hat{V}(\bar{X})]} \quad \text{(by Jensen's Inequality)} \\
&< \sqrt{V(\bar{X})} \quad \text{(since the sampling variance is unbiased)} \\
&< \sigma. \quad \square
\end{aligned}$$

The first line follows by first defining the conditional expectation of the sample mean's standard error, and then applying the noted variant of Jensen's inequality. Then, since we know that the standard error estimator of the variance is unbiased, we can replace the expectation with the true sampling variance, and note finally that the square root of the true sampling variance is, by definition, the true standard error of the sample mean. Hence, we see that our estimator of the sampling mean's standard error is strictly less than the true value and therefore is biased.

2.2 Chebyshev's Inequality

The other inequality Wooldridge highlights is the Chebyshev Inequality. This inequality states that for a set of probability distributions, no more than a specific proportion of that distribution is more than a set distance from the mean.

More formally, if $\mu = \mathbb{E}[X]$ and $\sigma^2 = var(X)$, then:

$$P(|Z| \geq k) \leq \frac{1}{k^2}, \tag{2.7}$$

where $Z = (X - \mu)/\sigma$ (Wasserman, 2004, p.64) and k indicates the number of standard deviations.

2.2.1 Proof

First, let us define the variance (σ^2) as:

$$\sigma^2 = \mathbb{E}[(X - \mu)^2]. \tag{2.8}$$

By expectation theory, we know that we can express any unconditional expectation as the weighted sum of its conditional components i.e. $\mathbb{E}[A] = \sum_i \mathbb{E}[A|c_i]P(c_i)$, where $\sum_i P(c_i) = 1$. Hence:

$$... = \mathbb{E}[(X - \mu)^2 | k\sigma \leq |X - \mu|]P(k\sigma \leq |X - \mu|) + \mathbb{E}[(X - \mu)^2 | k\sigma > |X - \mu|]P(k\sigma > |X - \mu|) \tag{2.9}$$

Since any probability is bounded between 0 and 1, and variance must be greater than or equal to zero, the second term must be non-negative. If we remove this term, therefore, the right-hand side is necessarily either the same size or smaller. Therefore we can alter the equality to the following inequality:

$$\sigma^2 \geq \mathbb{E}[(X-\mu)^2 | k\sigma \leq X-\mu] P(k\sigma \leq |X-\mu|) \tag{2.10}$$

This then simplifies:

$$\begin{aligned} \sigma^2 &\geq (k\sigma)^2 P(k\sigma \leq |X-\mu|) \\ &\geq k^2\sigma^2 P(k\sigma \leq |X-\mu|) \\ \frac{1}{k^2} &\geq P(|Z| \geq k) \quad \square \end{aligned}$$

Conditional on $k\sigma \leq |X-\mu|$, $(k\sigma)^2 \leq (X-\mu)^2$, and therefore $\mathbb{E}[(k\sigma)^2] \leq \mathbb{E}[(X-\mu)^2]$. Then, the last step simply rearranges the terms within the probability function.[3]

2.2.2 Applications

Wasserman notes that this inequality is useful when we want to know the probable bounds of an unknown quantity, and where direct computation would be difficult. It can also be used to prove the Weak Law of Large Numbers (point 5 in Wooldridge's list!) I delay discussion of this application until Section X.

It is worth noting, however, that the inequality is really powerful – it guarantees that a certain amount of a probability distribution is within a certain region – irrespective of the shape of that distribution (so long as we can estimate the mean and variance)!

For some well-defined distributions, this theorem is weaker than what we know by dint of their form. For example, we know that for a normal distribution, approximately 95 percent of values lie within 2 standard deviations of the mean. Chebyshev's Inequality only guarantees that 75 percent of values lie within two standard deviations of the mean (since $P(|Z| \geq k) \leq \frac{1}{2^2}$). Crucially, however, even if we didn't know whether a given distribution was normal, so long as it is a well-behaved probability distribution (i.e. the unrestricted integral sums to 1) we can guarantee that 75 percent will lie within two standard deviations of the mean.

[3] $k\sigma \leq |X-\mu| \equiv k \leq |X-\mu|/\sigma \equiv |Z| \geq k$, since σ is strictly non-negative.

Chapter 3

Linear Projection

This chapter provides a basic introduction to projection using both linear algebra and geometric demonstrations. I discuss the derivation of the orthogonal projection, its general properties as an "operator", and explore its relationship with ordinary least squares (OLS) regression. I defer a discussion of linear projections' applications until the penultimate chapter on the Frisch-Waugh Theorem, where projection matrices feature heavily in the proof.

3.1 Projection

Formally, a projection P is a linear function on a vector space, such that when it is applied to itself you get the same result i.e. $P^2 = P$.[1]

This definition is slightly intractable, but the intuition is reasonably simple. Consider a vector v in two-dimensions. v is a finite straight line pointing in a given direction. Suppose there is some point x not on this straight line but in the same two-dimensional space. The projection of x, i.e. Px, is a function that returns the point "closest" to x along the vector line v. Call this point $\bar{x}$. In most contexts, closest refers to Euclidean distance, i.e. $\sqrt{\sum_i (x_i - \bar{x}_i)^2}$, where i ranges over the dimensions of the vector space (in this case two dimensions).[fn^euclid] Figure @ref(fig:lp_basic) depicts this logic visually. The green dashed line shows the orthogonal projection, and red dashed lines indicate other potential (non-orthgonal) projections that are further away in Euclidean space from x than $\bar{x}$.

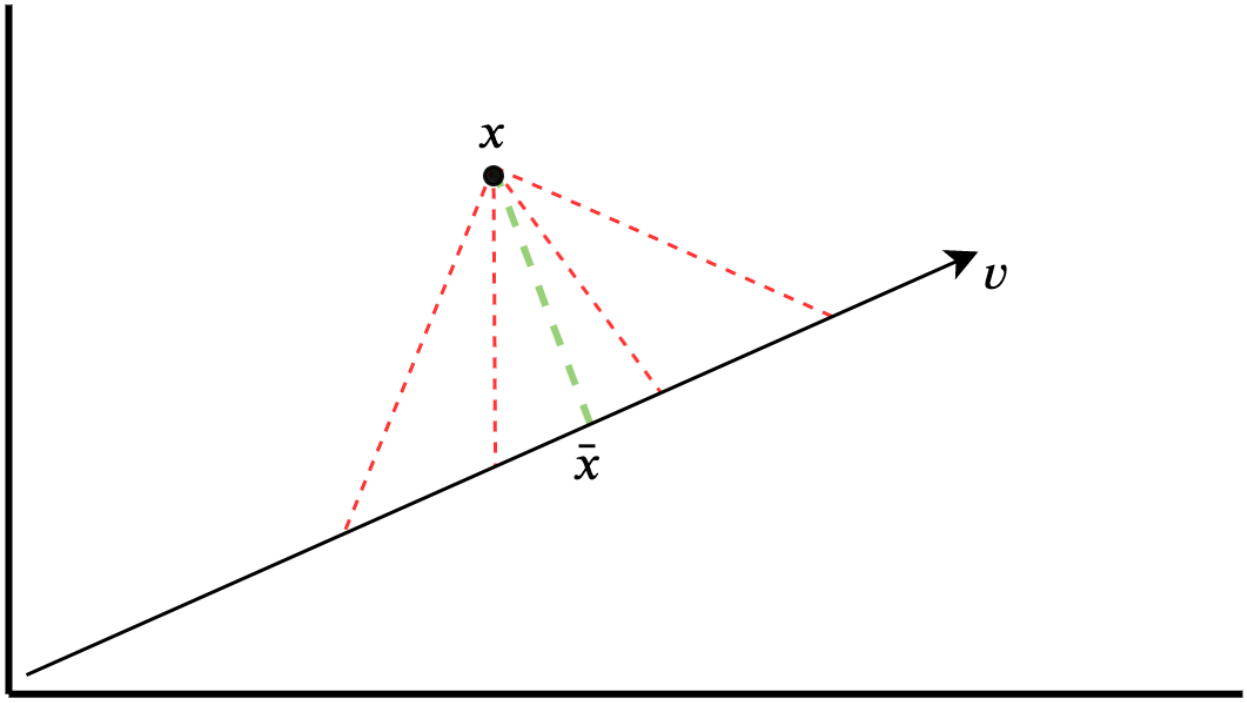
In short, projection is a way of simplifying some n-dimensional space – compressing information onto a (hyper-) plane. This is useful especially in social science settings where the complexity of the phenomena we study mean exact prediction is impossible. Instead, we often want to construct models that compress busy and variable data into simpler, parsimonious explanations. Projection is the statistical method of achieving this – it takes the full space and simplifies it with respect to a certain number of dimensions.

While the above is (reasonably) intuitive it is worth spelling out the maths behind projection, not least because it helps demonstrate the connection between linear projection and linear regression.

To begin, we can take some point in n-dimensional space, x, and the vector line v along which we want to project x. The goal is the following:

[1] Since P is (in the finite case) a square matrix, a projection matrix is an idempotent matrix – I discuss this property in more detail later on in this note.

Figure 3.1: Orthogonal projection of a point onto a vector line.



$$\begin{aligned} arg\,min_c \sqrt{\sum_i (\bar{x}_i - x)^2} &= arg\,min_c \sum_i (\bar{x}_i - x)^2 \\ &= arg\,min_c \sum_i (cv_i - x)^2 \end{aligned}$$

This rearrangement follows since the square root is a monotonic transformation, such that the optimal choice of c is the same across both $arg\ min$'s. Since any potential $\bar{x}$ along the line drawn by v is some scalar multiplication of that line (cv), we can express the function to be minimised with respect to c, and then differentiate:

$$\begin{aligned} \frac{d}{dc} \sum_i (cv_i - x)^2 &= \sum_i 2v_i(cv_i - x) \\ &= 2(\sum_i cv_i^2 - \sum_i v_i x) \\ &= 2(cv'v - v'x) \Rightarrow 0 \end{aligned}$$

Here we differentiate the equation and rearrange terms. The final step simply converts the summation notation into matrix multiplication. Solving:

$$\begin{aligned} 2(cv'v - v'x) &= 0 \\ cv'v - v'x &= 0 \\ cv'v &= v'x \\ c &= (v'v)^{-1}v'x. \end{aligned}$$

From here, note that $\bar{x}$, the projection of x onto the vector line, is $vc = v(v'v)^{-1}v'x$. Hence, we can define the projection matrix of x onto v as:

$$P_v = v(v'v)^{-1}v'.$$

In plain English, for any point in some space, the orthogonal projection of that point onto some subspace, is the point on a vector line that minimises the Euclidian distance between itself and the original point. A visual demonstration of this point is shown and discussed in Figure ?? below.

Note also that this projection matrix has a clear analogue to the linear algebraic expression of linear regression. The vector of coefficients in a linear regression $\widehat{\beta}$ can be expressed as $(X'X)^{-1}X'y$. And we know that multiplying this vector by the matrix of predictors X results in the vector of predicted values $\widehat{y}$. Now we have $\widehat{y} = X(X'X)^{-1}X'Y \equiv P_X y$. Clearly, therefore, linear projection and linear regression are closely related – and I return to this point below.

3.2 Properties of the projection matrix

The projection matrix P has several interesting properties. First, and most simply, the projection matrix is square. Since v is of some arbitrary dimensions $n \times k$, its transpose is of dimensions $k \times n$. By linear algebra, the shape of the full matrix is therefore $n \times n$, i.e. square.

Projection matrices are also symmetric, i.e. $P = P'$. To prove symmetry, note that transposing both sides of the projection matrix definition:

$$\begin{aligned} P' &= (v(v'v)^{-1}v')' && (3.1)\\ &= v(v'v)^{-1}v' && (3.2)\\ &= P, && (3.3) \end{aligned}$$

since $(AB)' = B'A'$ and $(A^{-1})' = (A')^{-1}$.

Projection matrices are idempotent:

$$\begin{aligned} PP &= v(v'v)^{-1}v'v(v'v)^{-1}v' && (3.4)\\ &= v(v'v)^{-1}v' && (3.5)\\ &= P, && (3.6) \end{aligned}$$

since $(A)^{-1}A = I$ and $BI = B$.

Since, projection matrices are idempotent, this entails that projecting a point already on the vector line will just return that same point. This is fairly intuitive: the closest point on the vector line to a point already on the vector line is just that same point.

Finally, we can see that the projection of any point is orthogonal to the respected projected point on vector line. Two vectors are orthogonal if $ab = 0$. Starting with the expression in Equation 3.1 (i.e. minimising the Euclidean distance with respect to c):

$$\begin{aligned} 2(cv'v - v'x) &= 0\\ v'cv - v'x &= 0\\ v'(cv - x) &= 0\\ v'(\bar{x} - x) &= 0, \end{aligned}$$

hence the line connecting the original point x is orthogonal to the vector line.

The projection matrix is very useful in other fundamental theorems in econometrics, like Frisch Waugh Lovell Theorem discussed in Chapter 8.

3.3 Linear regression

Given a vector of interest, how do we capture as much information from it as possible using set of predictors? Projection matrices essentially simplify the dimensionality of some space, by casting points onto a lower-dimensional plane. Think of it like capturing the shadow of an object on the ground. There is far more detail in the actual object itself but we roughly know its position, shape, and scale from the shadow that's cast on the 2d plane of the ground.

Note also this is actually quite similar to how we think about regression. Loosely, when we regress Y on X, we are trying to characterise how the components (or predictors) within X characterise or relate to Y. Of course, regression is also imperfect (after all, the optimisation goal is to minimise the errors of our predictions). So, regression also seems to capture some lower dimensional approximation of an outcome.

In fact, linear projection and linear regression are very closely related. In this final section, I outline how these two statistical concepts relate to each other, both algebraically and geometrically,

Suppose we have a vector of outcomes y, and some n-dimensional matrix X of predictors. We write the linear regression model as:

$$y = X\beta + \epsilon, \tag{3.7}$$

where β is a vector of coefficients, and ϵ is the difference between the prediction and the observed value in y. The goal of linear regression is to minimise the sum of the squared residuals:

$$arg\,min\ \epsilon^2 = arg\,min(y - X\beta)'(y - X\beta)$$

Differentiating with respect to and solving:

$$\begin{aligned}
\frac{d}{d\beta}(y - X\beta)'(y - X\beta) &= -2X(y - X\beta) \\
&= 2X'X\beta - 2X'y \Rightarrow 0 \\
X'X\hat{\beta} &= X'y \\
(X'X)^{-1}X'X\hat{\beta} &= (X'X)^{-1}X'y \\
\hat{\beta} &= (X'X)^{-1}X'y.
\end{aligned}$$

To get our prediction of y, i.e. $\hat{y}$, we simply multiply our beta coefficient by the matrix X:

$$\hat{y} = X(X'X)^{-1}X'y.$$

Note how the OLS derivation of $\hat{y}$ is very similar to $P = X(X'X)^{-1}X$, the orthogonal prediction matrix. The two differ only in that that $\hat{y}$ includes the original outcome vector y in its expression. But, note that $Py = X(X'X)^{-1}X'y = \hat{y}$! Hence the predicted values from a linear regression simply are an orthogonal projection of y onto the space defined by X.

3.3.1 Geometric interpretation

It should be clear now that linear projection and linear regression are connected – but it is probably less clear why this holds. To understand what's going on, let's depict the problem geometrically.[2]

To appreciate what's going on, we first need to invert how we typically think about observations, variables and datapoints. Consider a bivariate regression problem with three observations. Our data will include three variables: a constant (c, a vector of 1's), a predictor (X), and an outcome variable (Y). As a matrix, this might look something like the following:

Y	X	c
2	3	1
3	1	1
2	1	1

Typically we would represent the relationship geometrically by treating the variables as dimensions, such that every datapoint is an observation (and we would typically ignore the constant column since all its values are the same).

An alternative way to represent this data is to treat each observation (i.e. row) as a dimension and then represent each variable as a vector. What does that actually mean? Well consider the column $Y = (2, 3, 2)$. This vector essentially gives us the coordinates for a point in three-dimensional space: $d_1 = 2, d_2 = 3, d_3 = 2$. Drawing a straight line from the origin (0,0,0) to this point gives us a vector line for the outcome. While visually this might seem strange, from the perspective of our data it's not unusual to refer to each variable as a column vector, and that's precisely because it is a quantity with a magnitude and direction (as determined by its position in n dimensions).

Our predictors are the vectors X and c (note the vector c is now slightly more interesting because it is a diagonal line through the three-dimensional space). We can extend either vector line by multiplying it by a constant e.g. $2X = (6, 2, 2)$. With a single vector, we can only move forwards or backwards along a line. But if we combine two vectors together, we can actually reach lots of points in space. Imagine placing the vector X at the end of the c. The total path now reaches a new point that is not intersected by either X or c. In fact, if we multiply X and c by some scalars (numbers), we can snake our way across a whole array of different points in three-dimensional space. Figure 3.2 demonstrates some of these combinations in the two dimensional space created by X and c.

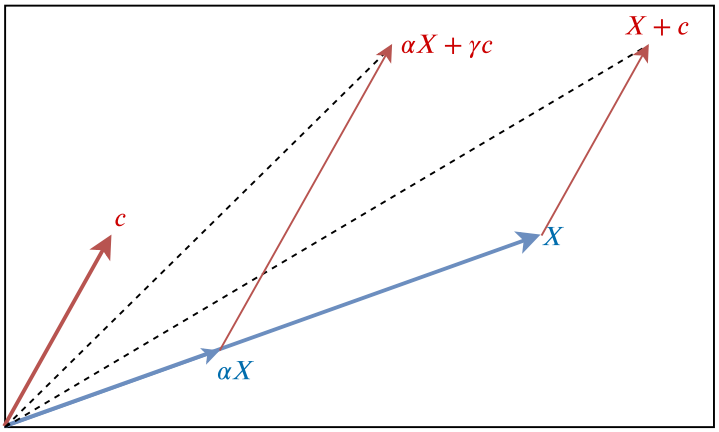
The comprehensive set of all possible points covered by linear combinations of X and c is called the span or column space. In fact, with the specific set up of this example (3 observations, two predictors), the span of our predictors is a flat plane. Imagine taking a flat bit of paper and aligning one corner with the origin, and then angling surface so that the end points of the vectors X and c are both resting on the card's surface. Keeping that alignment, any point on the surface of the card is reachable by some combination of X and c. Algebraically we can refer to this surface as $col(X, c)$, and it generalises beyond two predictors (although this is much harder to visualise).

Crucially, in our reduced example of three-dimensional space, there are points in space not reachable by combining these two vectors (any point above or below the piece of card). We know, for instance that the vector line y lies off this plane. The goal therefore is to find a vector that is on the column space of (X, c) that gets closest to our off-plane vector y as possible. Figure 3.3 depicts this set up visually – each dimension is an observation, each column in the matrix is represented a vector, and the column space of (X, c) is the shaded grey plane. The vector y lies off this plane.

From our discussion in Section 3.1, we know that the "best" vector is the orthogonal projection from the column space to the vector y. This is the shortest possible distance between the flat plane and the observed

[2]This final section borrows heavily from Ben Lambert's explanation of projection and a demonstration using R by Andy Eggers.

Figure 3.2: Potential combinations of two vectors.



outcome, and is just $\hat{y}$. Moreover, since $\hat{y}$ lies on the column space, we know we only need to combine some scaled amount of X and c to define the vector $\hat{y}$, i.e., $\beta_1 X + \beta_0 c$. Figure 3.4 shows this geometrically. And in fact, the scalar coefficients β_1, β_0 in this case are just the regression coefficients derived from OLS. Why? Because we know that the orthogonal projection of y onto the column space minimises the error between our prediction $\hat{y}$ and the observed outcome vector y. This is the same as the minimisation problem that OLS solves, as outlined at the beginning of this section!

Consider any other vector on the column space, and the distance between itself and and y. Each non-orthogonal vector would be longer, and hence have a larger predictive error, than $\hat{y}$. For example, Figure 3.5 plots two alternative vectors on $col(X, c)$ alongside $\hat{y}$. Clearly, $\hat{\epsilon} < \epsilon' < \epsilon''$, and this is true of any other vector on the column space too.

Hence, linear projection and linear regression can be seen (both algebraically and geometrically) to be solving the same problem – minimising the (squared) distance between an observed vector y and prediction vector $\hat{y}$. This demonstration generalises to many dimensions (observations), though of course it becomes much harder to intuit the geometry of highly-dimensional data. And similarly, with more observations we could also extend the number of predictors too such that X is not a single column vector but a matrix of predictor variables (i.e. multivariate regression). Again, visualising what the column space of this matrix would look like geometrically becomes harder.

To summarise, this section has demonstrated two features. First, that linear regression simply is an orthogonal projection. We saw this algebraically by noting that the derivation of OLS coefficients, and subsequently the predicted values from a linear regression, is identical to Py (where P is a projection matrix). Second, and geometrically, we intuited why this is the case: namely that projecting onto a lower-dimensional column space involves finding the linear combination of predictors that minimises the Euclidean distance to y, i.e. $\hat{y}$. The scalars we use to do so are simply the regression coefficients we would generate using OLS regression.

Figure 3.3: Schematic of orthogonal projection as a geometric problem

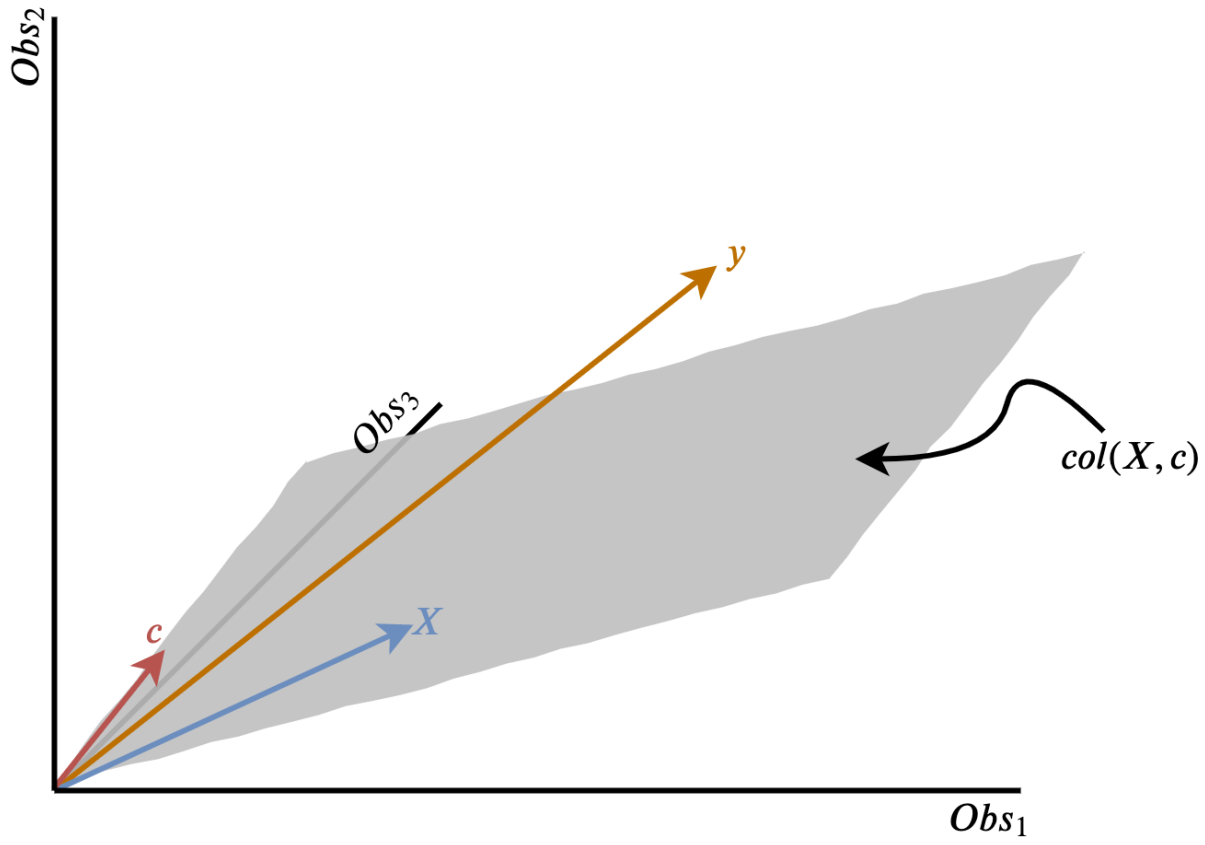


Figure 3.4: Relation of orthogonal projection to linear regression.

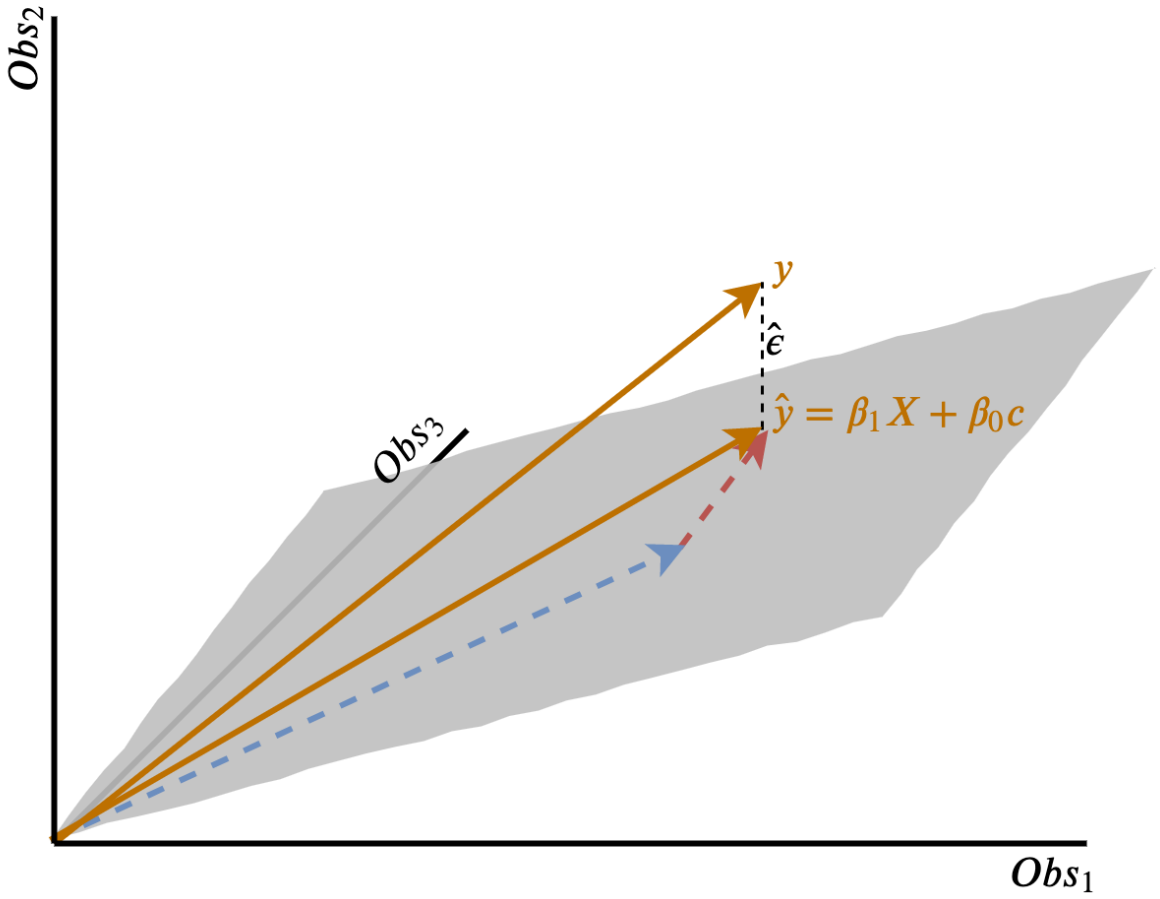
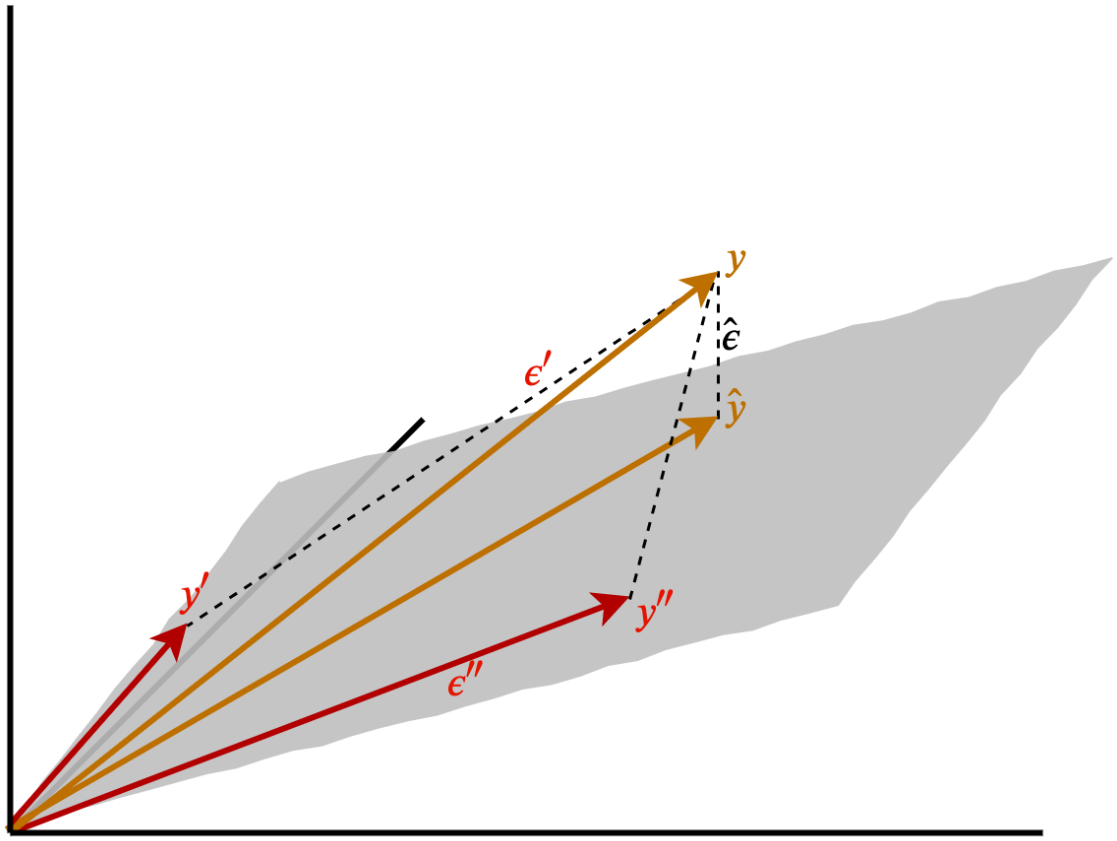


Figure 3.5: Alternative vectors on the column space are further away from y.



Chapter 4

Weak Law of Large Numbers and Central Limit Theorem

This chapter focuses on two fundamental theorems that form the basis of our inferences from samples to populations. The Weak Law of Large Numbers (WLLN) provides the basis for generalisation from a sample mean to the population mean. The Central Limit Theorem (CLT) provides the basis for quantifying our uncertainty over this parameter. In both cases, I discuss the theorem itself and provide an annotated proof. Finally, I discuss how the two theorems complement each other.

4.1 Weak Law of Large Numbers

4.1.1 Theorem in Plain English

Suppose we have a random variable X. From X, we can generate a sequence of random variables $X_1, X_2, ..., X_n$ that are independent and identically distributed (i.i.d.) draws of X. Assuming n is finite, we can perform calculations on this sequence of random numbers. For example, we can calculate the mean of the sequence $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. This value is the sample mean – from a much wider population, we have drawn a finite sequence of observations, and calculated the average across them. How do we know that this sample parameter is meaningful with respect to the population, and therefore that we can make inferences from it?

WLLN states that the mean of a sequence of i.i.d. random variables converges in probability to the expected value of the random variable as the length of that sequence tends to infinity. By 'converging in probability', we mean that the probability that the difference between the mean of the sample and the expected value of the random variable tends to zero.

In short, WLLN guarantees that with a large enough sample size the sample mean should approximately match the true population parameter. Clearly, this is powerful theorem for any statistical exercise: given we are (always) constrained by a finite sample, WLLN ensures that we can infer from the data something meaningful about the population. For example, from a large enough sample of voters we can estimate the average support for a candidate or party.

More formally, we can state WLLN as follows:

$$\bar{X}_n \xrightarrow{p} \mathbb{E}[X], \tag{4.1}$$

where $\xrightarrow{p}$ denotes 'converging in probability'.

4.1.2 Proof

To prove WLLN, we use Chebyshev's Inequality (CI). More specifically we first have to prove Chebyshev's Inequality of the Sample Mean (CISM), and then use CISM to prove WLLN. The following steps are based on the proof provided in Aronow and Miller (2019).

Proof of Chebyshev's Inequality of the Sample Mean. Chebyshev's Inequality for the Sample Mean (CISM) states that:

$$P(|\bar{X}_n - \mathbb{E}[X]| \geq k) \leq \frac{var(X)}{k^2 n}, \tag{4.2}$$

where $\bar{X}_n$ is the sample mean of a sequensce of n independent draws from a random variable X. Recall CI states that $P(|(X-\mu)/\sigma| \geq k) \leq \frac{1}{k^2}$. To help prove CISM, we can rearrange the left hand side of the inequality by multiplying both sides of the inequality within the probability function by σ, such that:

$$P(|(X-\mu)| \geq k\sigma) \leq \frac{1}{k^2}. \tag{4.3}$$

Then, finally, let us define $k' = \frac{k}{\sigma}$. Hence:

$$\begin{aligned} P(|(\bar{X} - \mathbb{E}[X])| \geq k) &= P(|(\bar{X} - \mathbb{E}[X])| \geq k'\sigma) && (4.4)\\ &\leq \frac{1}{k'^2} && (4.5)\\ &\leq \frac{\sigma^2}{k^2} && (4.6)\\ &\leq \frac{var(\bar{X})}{k^2} && (4.7)\\ &\leq \frac{var(X)}{k^2 n} \quad \square && (4.8) \end{aligned}$$

This proof is reasonably straightfoward. Using our definition of k' allows us to us rearrange the probability within CISM to match the form of the Chebyshev Inequality stated above, which then allows us to infer the bounds of the probability. We then replace k' with $\frac{k}{\sigma}$, expand and simplify. The move made between the penultimate and final line relies on the fact that variance of the sample mean is equal to the variance in the random variable divided by the sample size (n).[1]

Applying CISM to WLLN proof. Given that all probabilities are non-negative and CISM, we can now write:

$$0 \leq P(|\bar{X}_n - \mathbb{E}[X]| \geq k) \leq \frac{var(X)}{k^2 n}. \tag{4.9}$$

Note that for the first and third term of this multiple inequality, as n approaches infinity both terms approach 0. In the case of the constant zero, this is trivial. In the final term, note that $var(X)$ denotes the inherent variance of the random variable, and therefore is constant as n increases. Therefore, as the denominator increases, the term converges to zero.

Since the middle term is sandwiched in between these two limits, by definition we know that this term must also converge to zero.[2] Therefore:

[1] See Aronow and Miller 2019, p.98.

[2] To see why this is the case, given the limits of the first and third terms, Equation ?? is of the form $0 \leq A \leq 0$ as $n \to \infty$. The only value of A that satisfies this inequality is 0.

$$\lim_{n\to\infty} P(|\bar{X}_n - \mathbb{E}[X]| \geq k) = 0 \quad \square \tag{4.10}$$

Hence, WLLN is proved: for any value of k, the probability that the difference between the sample mean and the expected value is greater or equal to k converges on zero. Since k's value is arbitrary, it can be set to something infinitesimally small, such that the sample mean and expected value converge in value.

4.2 Central Limit Theorem

WLLN applies to the value of the statistic itself (the mean value). Given a single, n-length sequence drawn from a random variable, we know that the mean of this sequence will converge on the expected value of the random variable. But often, we want to think about what happens when we (hypothetically) calculate the mean across multiple sequences i.e. expectations under repeat sampling.

The Central Limit Theorem (CLT) is closely related to the WLLN. Like WLLN, it relies on asymptotic properties of random variables as the sample size increases. CLT, however, lets us make informative claims about the distribution of the sample mean around the true population parameter.

4.2.1 Theorem in Plain English

CLT states that as the sample size increases, the distribution of sample means converges to a normal distribution. That is, so long as the underlying distribution has a finite variance (bye bye Cauchy!), then irrespective of the underlying distribution of X the distribution of sample means will be a normal distribution!

In fact, there are multiple types of CLT that apply in a variety of different contexts – cases including Bernoulli random variables (de Moivre - Laplace), where random variables are independent but do not need to be identically distributed (Lyapunov), and where random variables are vectors in $\mathbb{R}^k$ space (multivariate CLT).

In what follows, I will discuss a weaker, more basic case of CLT where we assume random variables are scalar, independent, and identically distributed (i.e. drawn from the same unknown distribution function). In particular, this section proves that the standardized difference between the sample mean and population mean for i.i.d. random variables converges in distribution to the standard normal distribution $N(0, 1)$. This variant of the CLT is called the Lindeberg-Levy CLT, and can be stated as:

$$\frac{\bar{X}_n - \mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1), \tag{4.11}$$

where $\xrightarrow{d}$ denotes 'converging in distribution'.

In general, the CLT is useful because proving that the sample mean is normally distributed allows us to quantify the uncertainty around our parameter estimate. Normal distributions have convenient properties that allow us to calculate the area under any portion of the curve, given just the same mean and standard deviation. We already know by WLLN that the sample mean will (with a sufficiently large sample) approximate the population mean, so we know that the distribution is also centred around the true population mean. By CLT, the dispersion around that point is therefore normal, and to quantify the probable bounds of the point estimate (under the assumption of repeat sampling) requires only an estimate of the variance.

4.2.2 Primer: Characteristic Functions

CLT is harder (and lengthier) to prove than other proofs we've encountered so far – it relies on showing that the sample mean converges in distribution to a known mathematical form that uniquely and fully describes

the normal distribution. To do so, we use the idea of a characteristic functions, which simply denotes a function that completely defines a probability function.

For example, and we will use this later on, we know that the characteristic function of the normal distribution is $e^{it\mu-\frac{\sigma^2t^2}{2}}$. A standard normal distriibution (where $\mu=0,\sigma^2=1$) therefore simplifies to $e^{-\frac{t^2}{2}}$.

More generally, we know that for any scalar random variable X, the characteristic function of X is defined as:

$$\phi_X(t)=\mathbb{E}[e^{itX}], \tag{4.12}$$

where $t\in\mathbb{R}$ and i is the imaginary unit. Proving why this is the case is beyond the purview of this section, so unfortunately I will just take it at face value.

We can expand e^{itX} as an infinite sum, using a Taylor Series, since $e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+....$ Hence:

$$\phi_X(t)=\mathbb{E}[1+itX+\frac{(itX)^2}{2!}+\frac{(itX)^3}{3!}+...], \tag{4.13}$$

Note that $i^2=-1$, and since the latter terms tend to zero faster than the second order term we can summarise them as $o(t^2)$ (they are no larger than of order t^2). Therefore we can rewrite this expression as:

$$\phi_X(t)=\mathbb{E}[1+itX-\frac{t^2}{2}X^2+o(t^2)]. \tag{4.14}$$

In the case of continuous random variables, the expected value can be expressed as the integral across all space of the expression multiplied by the probability density, such that:

$$\phi_X(t)=\int_{-\infty}^{\infty}[1+itX-\frac{t^2}{2}X^2+o(t^2)]f_X dX, \tag{4.15}$$

and this can be simplified to:

$$\phi_X(t)=1+it\mathbb{E}[X]-\frac{t^2}{2}\mathbb{E}[X^2]+o(t^2)], \tag{4.16}$$

since $1\times f_X=f_X$, the total area under a probability density necessarily sums to 1; $\int Xf_X dX$ is the definition of the expected value of X, and so by similar logic $\int X^2 f_X dX=\mathbb{E}[X^2]$.

In Ben Lambert's video introducing the CLT proof, he notes that if we assume X has mean 0 and variance 1, the characteristic function of that distribution has some nice properties, namely that it simplifies to:

$$\phi_X(t)=1-\frac{t^2}{2}+o(t^2)], \tag{4.17}$$

since $\mathbb{E}[X]=0$ cancelling the second term, and $\mathbb{E}[X^2]\equiv\mathbb{E}[(X-0)^2]=\mathbb{E}[(X-\mu)^2]=var(X)=1$.

One final piece of characteristic function math that will help finalise the CLT proof is to note that if we define some random variable $Q_n=\sum_{i=1}^{n}R_i$, where all R_i are i.i.d., then the characteristic function of Q_n can be expressed as $\phi_{Q_n}(t)=[\phi_R(t)]^n$. Again, I will not prove this property here.

4.2.3 Proof of CLT

This proof is based in part on Ben Lambert's excellent YouTube series, as well as Lemons et al. (2002).

Given the above discussion of a characteristic function, let us assume a sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, ..., X_n$, each with mean μ and finite[3] variance σ^2. The sum of these random variables has mean $n\mu$ (since each random variable has the same mean) and the variance equivalent to $n\sigma^2$ (because the random variables are i.i.d. we know that $var(A, B) = var(A)var(B)$).

Now let's consider the standardized difference between the actual sum of the random variables and the mean. Standardization simply means dividing a parameter estimate by its standard deviation. In particular, we can consider the following standardized random variable:

$$Z_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}, \tag{4.18}$$

where Z_n, in words, is the standardised difference between the sum of i.i.d. random variables and the expected value of the sequence. Note that we use the known variance in the denominator.

We can simplify this further:

$$Z_n = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i, \tag{4.19}$$

where we define a new random variable $Y_i = \frac{X_i - \mu}{\sigma}$.

Y_i has some convenient properties. First, since each random variable X_i in our sample has mean μ, we know that $\mathbb{E}[Y_i] = 0$ since $\mathbb{E}[X_i] = \mu$ and therefore $\mu - \mu = 0$. Note that this holds irrespective of the distribution and value of $\mathbb{E}[X_i]$.

The variance of Y_i is also recoverable. First note three basic features of variance: if a is a constant, and X and Y are random variables, $var(a) = 0$; $var(aX) = a^2 var(X)$; and from the variance of a sum $var(X - Y) = var(A) - var(B)$. Therefore:

$$var(\frac{1}{\sigma}(X_i - \mu) = \frac{1}{\sigma^2} var(X_i - \mu) \tag{4.20}$$

$$var(X_i - \mu) = var(X_i) - var(\mu) \tag{4.21}$$

$$= var(X_i). \tag{4.22}$$

Hence:

$$var(Y_i) = \frac{var(X_i)}{\sigma^2} = 1, \tag{4.23}$$

since $var(X_i) = \sigma^2$.

At this stage, the proof is tantalisingly close. While we have not yet fully characterised the distribution of Z_n or even Y_i, the fact that Y_i has unit variance and a mean of zero means suggests we are on the right track to proving that this does asymptotically tend in distribution to the standard normal. In fact, recall from the primer on characteristic functions, that Lambert notes for any random variable with unit variance and mean of 0, $\phi_X(t) = 1 - \frac{t^2}{2} + o(t^2)$. Hence, we can now say that:

[3] Hence why degenerate distributions like the Cauchy are not covered by CLT.

$$\phi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2). \tag{4.24}$$

Now let us return to $Z_n = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Y_i$ and using the final bit of characteristic function math in the primer, we can express the characteristic function of Z_n as:

$$\phi_{Z_n}(t) = [\phi_Y(\frac{t}{\sqrt{n}})]^n, \tag{4.25}$$

since Y_i is divided by the square root of the sample size. Given our previously stated expression of the characteristic function of Y_i:

$$\phi_{Z_n}(t) = [1 - \frac{t^2}{2n} + o(t^2)]]^n. \tag{4.26}$$

We can now consider what happens as $n \to \infty$. By definition, we know that $o(t^2)$ converges to zero faster than the other terms, so we can safely ignore it. As a result, and noting that $e^x = \lim(1 + \frac{x}{n})^n$:

$$\lim_{n\to\infty} \phi_{Z_n}(t) = e^{-\frac{t^2}{2}}. \tag{4.27}$$

This expression shows that as n tends to infinity, the characteristic function of Z_n is the standard normal distribution (as noted in the characteristic function primer). Therefore:

$$\lim_{n\to\infty} Z_n = N(0,1) \tag{4.28}$$

$$\lim_{n\to\infty} \frac{\bar{X}_n - \mu}{\sigma\sqrt{n}} = N(0,1). \quad \square \tag{4.29}$$

The last line here simply follows from the definition of Z_n.

4.2.4 Generalising CLT

From here, it is possible to intuit the more general CLT that the distribution of sampling means is normally distributed around the true mean μ with variance $\frac{\sigma^2}{n}$. Note this is only a pseudo-proof, because as Lambert notes, multiplying through by n is complicated by the limit operator with respect to n. However, it is useful to see how these two CLT are closely related.

First, we can rearrange the limit expression using known features of the normal distribution:

$$\lim_{n\to\infty} Z_n \xrightarrow{d} N(0,1) \tag{4.30}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i) - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0,1) \tag{4.31}$$

$$\lim_{n\to\infty} \sum_{i=1}^{n}(X_i) - n\mu \xrightarrow{d} N(0, n\sigma^2) \tag{4.32}$$

$$\lim_{n\to\infty} \sum_{i=1}^{n}(X_i) \xrightarrow{d} N(n\mu, n\sigma^2), \tag{4.33}$$

since $aN(b,c) = N(ab, a^2c)$, and $N(d,e) + f = N(d+f, e)$.

At this penultimate step, we know that the sum of i.i.d. random variables is a normal distribution. To see that the sample mean is also normally distributed, we simply divide through by n:

$$\lim_{n\to\infty} \bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i) \xrightarrow{d} N(\mu, \frac{\sigma^2}{n}). \tag{4.34}$$

4.2.5 Limitation of CLT (and the importance of WLLN)

Before ending, it is worth noting that CLT is a claim with respect to repeat sampling from a population (holding n constant each time). It is not, therefore, a claim that holds with respect to any particular sample draw. We may actually estimate a mean value that, while probable, lies away from the true population parameter (by definition, since the sample means are normally distributed, there is some dispersion). Constructing uncertainty estimates using CLT on this estimate alone does not guarantee that we are in fact capturing either the true variance or the true parameter.

That being said, with sufficiently high-N, we know that WLLN guarantees (assuming i.i.d. observations) that our estimate converges on the population mean. WLLN's asymptotics rely only on sufficiently large sample sizes for a single sample. Hence, both WLLN and CLT are crucial for valid inference from sampled data. WLLN leads us to expect that our parameter estimate will in fact be centred approximately near the true parameter. Here, CLT can only say that across multiple samples from the population the distribution of sample means is centred on the true parameter. With WLLN in action, however, CLT allows us to make inferential claims about the uncertainty of this converged parameter.

Chapter 5

Slutsky's Theorem

5.1 Theorem in plain English

Slutsky's Theorem allows us to make claims about the convergence of random variables. It states that a random variable converging to some distribution X, when multiplied by a variable converging in probability on some constant a, converges in distribution to $a \times X$. Similarly, if you add the two random variables, they converge in distribution to a plus X. More formally, the theorem states that if $X_n \xrightarrow{d} X$ and $A_n \xrightarrow{p}$, where a is a constant, then:

1. $X_n + A_n \xrightarrow{d} X + a$
2. $A_n X_n \xrightarrow{d} aX$

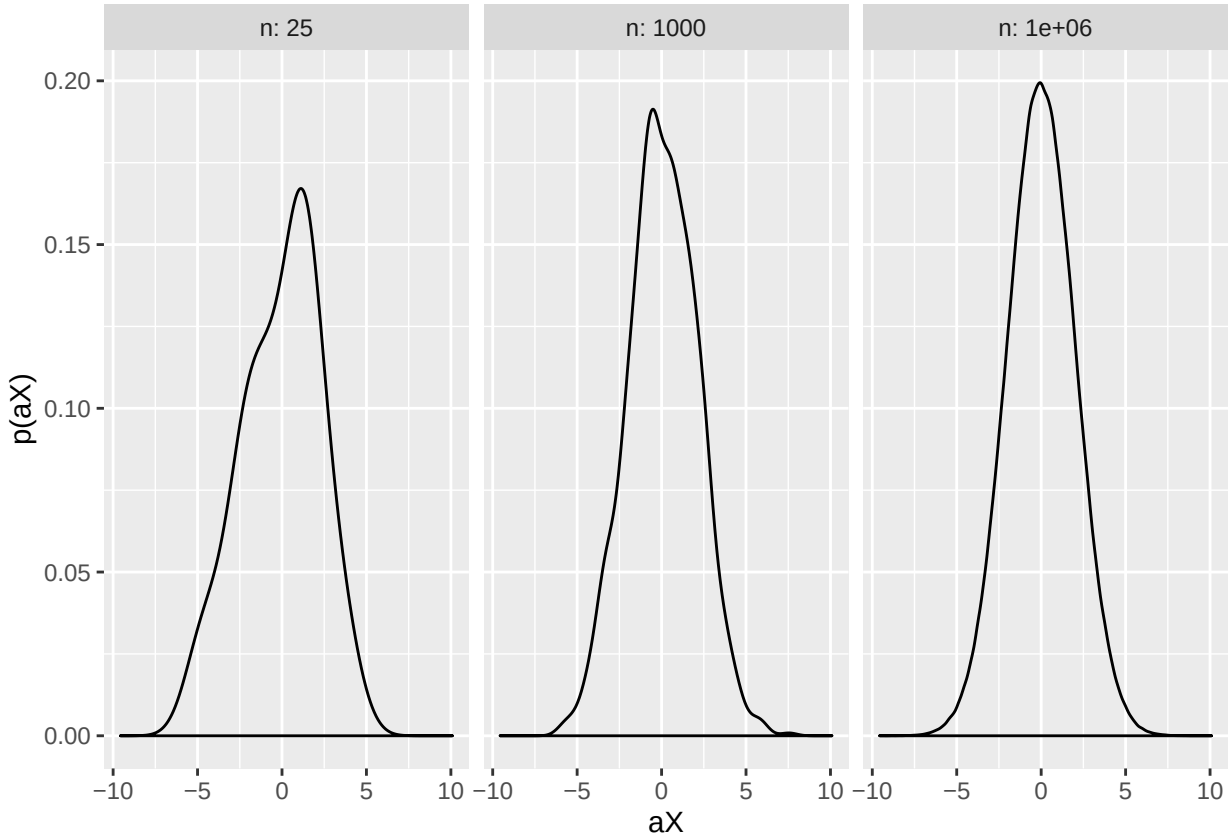
Note that if A_n or B_n do not converge in probability to constants, and instead converge towards some distribution, then Slutsky's Theorem does not hold. More trivially, if all variables converge in probability to constants, then $A_n X_n + B_n \xrightarrow{p} aX + B$.

5.2 Coded demonstration

This theorem is reasonably intuitive. Suppose that the random variable X_n converges in distribution to a standard normal distribution $N(0, 1)$. For part 1) of the Theorem, note that when we multiply a standard normal by a constant we "stretch" the distribution (assuming $|a| > 1$, else we "compress" it). Recall from the discussion of the standard normal in Chapter 5 that $aN(0, 1) = N(0, a^2)$. As n approaches infinity, therefore, by definition $A_n \xrightarrow{p} a$, and so the degree to which the standard normal is stretched will converge to that constant too. To demonstrate this feature visually, consider the following simulation:

```
library(ggplot2)
set.seed(89)
N <- c(25,1000,1000000)

results <- data.frame(n = as.factor(levels(N)),
                      X_n = as.numeric(),
                      A_n = as.numeric(),
                      ax = as.numeric())

for (n in N) {
  X_n <- rnorm(n)
```

```
  A_n <- 2 + exp(-n)
  aX <- A_n * X_n

  results <- rbind(results, cbind(n, X_n, A_n, aX))

}

ggplot(results, aes(x = aX)) +
  facet_wrap(n~., ncol = 3, labeller = "label_both") +
  geom_density() +
  labs(y = "p(aX)")
```

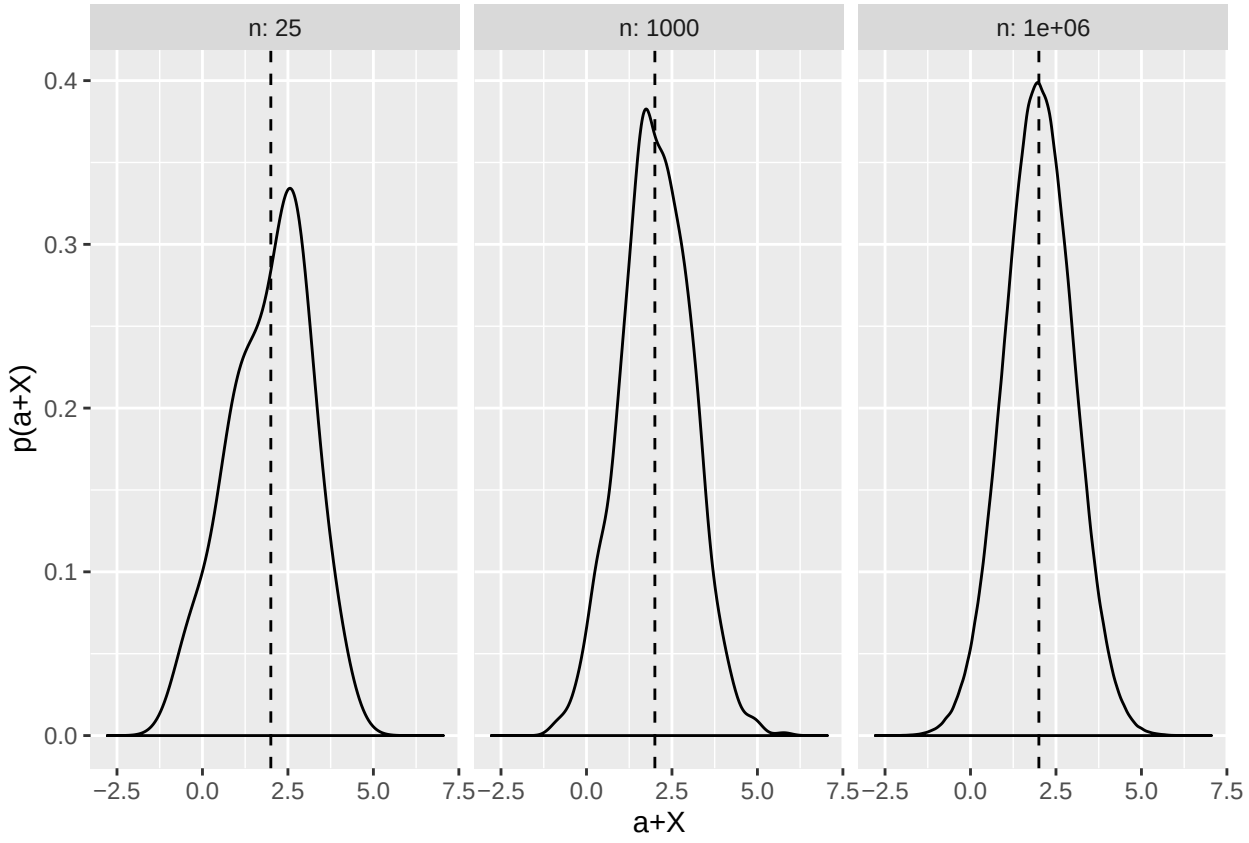


Here we have defined two random variables: `X_n` is a standard normal, and `A_n` converges in value to 2. Varying the value of `n`, I take n draws from a standard normal distribution and calculate the value the converging constant A_n. I then generate the product of these two variables. The figure plots the resulting distribution `aX`. We can see that as `n` increases, the distribution becomes increasingly normal, remains centred around 0 and the variance approaches 4 (since 95% of the curve is approximately bounded between $0 \pm 2 \times \sqrt{var(aX)} = 0 \pm 2 \times 2 = 0 \pm 4$).

Similarly, if we add the constant a to a standard distribution, the effect is to shift the distribution in its entirety (since a constant has no variance, it does not ''stretch" the distribution). As A_n converges in probability, therefore, the shift converges on the constant a. Again, we can demonstrate this result in R:

```
library(ggplot2)
set.seed(89)
N <- c(25,1000,1000000)

results <- data.frame(n = as.factor(levels(N)),
```

```
                      X_n = as.numeric(),
                      A_n = as.numeric(),
                      a_plus_X= as.numeric())

for (n in N) {
  X_n <- rnorm(n)
  A_n <- 2 + exp(-n)
  a_plus_X <- A_n + X_n

  results <- rbind(results, cbind(n, X_n, A_n, a_plus_X))

}

ggplot(results, aes(x = a_plus_X)) +
  facet_wrap(n~., ncol = 3, labeller = "label_both") +
  geom_density() +
  geom_vline(xintercept = 2, linetype = "dashed") +
  labs(y = "p(a+X)", x="a+X")
```



As `n` becomes larger, the resulting distribution becomes approximately normal, with variance of 1 and a mean value centred around $0 + a = 2$.

Slutsky's Theorem is so useful precisely because it allows us to combine multiple random variables with known asymptotics, and retain this knowledge i.e. we know what the resultant distribution will converge to assuming $n \to \infty$.

5.3 Proof of Slutsky's Theorem

Despite the intuitive appeal of Slutsky's Theorem, the proof is less straightforward. It relies on the continuous mapping theorem (CMT), which in turns rests on several other theorems such as the Portmanteau Theorem. To avoid the rabbit hole of proving all necessary antecedent theorems, I simply introduce and state the continuous mapping theorem (CMT) here, and then show how this can be used to prove Slutsky's Theorem.

5.3.1 CMT

The continuous mapping theorem states that if there is some random variable such that $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$, so long as g is a continuous function. In approximate terms (which are adequate for our purpose), a continuous function is one in which for a given domain the function can be represented as an single unbroken curve (or hyperplane in many dimensions). For example, consider the graph of $f(x) = x^{-1}$. For the domain $D_+ : \mathbb{R} > 0$, this function is continuous. But for the domain $D_\infty : \mathbb{R}$, the function is discontinuous because the function is undefined when $x = 0$.

In short, CMT states that a continuous function preserves the asymptotic limits of a random variable. More broadly (and again, I do not prove this here), CMT entails that $g(P_n, Q_n, ..., Z_n) \xrightarrow{d} g(P, Q, ..., Z)$ if all $P_n, Q_n, ...$ etc. converge in distribution to $P, Q, ...$ respectively.

5.3.2 Proof using CMT

How does this help prove Slutsky's Theorem? We know by the definitions in Slutsky's Theorem that $X_n \xrightarrow{d} X$ and, by a similar logic, we know that $A_n \xrightarrow{d} a$ (since $A_n \xrightarrow{p} a$, and converging in probability entails converging in distribution). So we can note that the joint vector $(X_n, A_n) \xrightarrow{d} (X, a)$. By CMT, therefore, $g(X_n, A_n) \xrightarrow{d} g(X, a)$. Hence, any continuous function g will preserve the limits of the respective distributions.

Given this result, it is sufficient to note that both addition and multiplication are continuous functions. Again, I do not show this here but the continuity of addition and multiplication (both scalar and vector) can be proved mathematically (for example see one such proof here For an intuitive explanation, think about the diagonal line $y = X$ – any multiplication of that line is still a single, uninterrupted line ($y = aX$) assuming a is a constant. Similarly, adding a constant to the function of a line also yields an uninterrupted line (e.g. $y = X + a$).

Hence, CMT guarantees both parts 1 and 2 of the Theorem. □

5.4 Applications

Slutsky's Theorem is a workhorse theorem that allows researchers to make claims about the limiting distributions of multiple random variables. Instead of being used in applied settings, it typically underpins the modelling strategies used in applied research. For example, Aronow and Samii (2016) consider the problem of weighting multiple regression when the data sample is unrepresentative of the population of interest. In their proofs, they apply Slutsky's Theorem at two different points to prove that their weighted regression estimates converge in probability on the weighted expectation of individual treatment effects, and subsequently, that the same coefficient converges in probability to the true average treatment effect in the population.

5.4.1 Proving the consistency of sample variance, and the normality of the t-statistic

In the remainder of this chapter, I consider applications of both Central Mapping Theorem and Slutsky's Theorem in fundamental statistical proofs. I first show how CMT can be used to prove the consistency of the variance of a random variable, and subsequently how in combination with Slutsky's Theorem this helps prove the normality of a t-statistic. These examples are developed from David Hunter's notes on asymptotic theory that accompany his Penn State course in large-sample theory.

5.4.1.1 Consistency of the sample variance estimator

First, let us define the sample variance (s_n^2) of a sequence of i.i.d random variables drawn from a distribution X with $\mathbb{E}[X] = \mu$ and $var(X) = \sigma^2$ as:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2.$$

We can show that the sample variance formula above is a consistent estimator of the true variance σ^2. That is, as the sequence of i.i.d. random variables $X_1, X_2, ...$ increases in length, the sample variance estimator of that sequence converges in probability to the true variance value σ^2.

We can prove this by redefining s^2 as follows:

$$s_n^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X_n} - \mu)^2\right],$$

which clearly simplifies to the conventional definition of s^2 as first introduced.

From here, we can note using WLLN, that $(\bar{X_n} - \mu) \xrightarrow{p} 0$, and hence that $(\bar{X_n} - \mu)^2 \xrightarrow{p} 0$. Note that this term converges in probability to a constant. Moreover, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{p} \mathbb{E}[X_i - \mu] = var(X) = \sigma^2$, by definition.

Now let us define an arbitrary continuous function $g(A_n, B_n)$. We know by CMT that if $A_n \xrightarrow{p} A, B_n \xrightarrow{p} B$ then $g(A_n, B_n) \xrightarrow{p} g(A, B)$. And hence, using the implications above we know that for any continuous function g that $g(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, (\bar{X_n} - \mu)) \xrightarrow{p} g(\sigma^2, 0)$.

Since subtraction is a continuous function, we therefore know that:

$$\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X_n} - \mu)^2\right] \xrightarrow{p} [\sigma^2 - 0] = \sigma^2.$$

Separately we can intuitively see that $\frac{n}{n-1} \xrightarrow{p} 1$. Hence, by applying CMT again to this converging variable multiplied by the converging limit of the above (since multiplication is a continuous function), we can see that:

$$s_n^2 \xrightarrow{p} 1 \times \sigma^2 = \sigma^2 \quad \square$$

5.4.1.2 Normality of the t-statistic

Let's define a t-statistic as:

$$t_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\hat{\sigma^2}}}$$

By the Central Limit Theorem (CLT, Chapter 5), we know that for a random variable X with mean μ and variance σ^2 that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

We also know from the proof above that if $\hat{\sigma}^2 = s^2$ then $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ – a constant. Given this, we can also note that $\frac{1}{\hat{\sigma^2}} \xrightarrow{p} \frac{1}{\sigma^2}$.

Hence, by Slutsky's Theorem:

$$\sqrt{n}(\bar{X}_n - \mu) \times \frac{1}{\sqrt{\sigma^2}} \xrightarrow{d} N(0, \sigma^2) \times \frac{1}{\sqrt{\sigma^2}} \tag{5.1}$$

$$= \sigma N(0,1) \times \frac{1}{\sigma} \tag{5.2}$$

$$= N(0,1) \quad \square \tag{5.3}$$

One brief aspect of this proof that is noteworthy is that since Slutsky's Theorem rests on the CMT, the application of Slutsky's Theorem requires that the function of the variables g (in this case multiplication) is continuous and defined for the specified domain. Note that $\frac{1}{0}$ is undefined and therefore that the above proof only holds when we assume $\sigma^2 > 0$. Hence why in many statistics textbooks and discussions of model asymptotics, authors note that they must assume a positive, non-zero variance.

Chapter 6

Big Op and little op

6.1 Stochastic order notation

"Big Op" (big oh-pee), or in algebraic terms O_p, is a shorthand means of characterising the convergence in probability of a set of random variables. It directly builds on the same sort of convergence ideas that were discussed in Chapters 5 and 6.

Big Op means that some given random variable is stochastically bounded. If we have some random variable X_n and some constant a_n (where n indexes both sets), then

$$X_n = O_p(a_n)$$

is the same as saying that

$$P(|\frac{X_n}{a_n}| > \delta) < \epsilon, \forall n > N.$$

M and N here are just finite numbers, and ϵ is some arbitrary (small) number. In plain English, O_p means that for a large enough n there is some number (M) such that the probability that the random variable $\frac{X_n}{a_n}$ is larger than that number is essentially zero. It is "bounded in probability" (van der Vaart, 1998, Section 2.2).

"Little op" (litle oh-pee), or o_p, refers to convergence in probability towards zero. $X_n = o_p(1)$ is the same as saying

$$\lim_{n\to\infty}(P|X_n| \geq \epsilon) = 0, \forall \epsilon > 0.$$

By definition of the notation, if $X_n = o_p(a_n)$ then

$$\frac{x_n}{a_n} = o_p(1).$$

In turn, we can therefore express $X_n = o_p(a_n)$ as

$$\lim_{n\to\infty}(P|\frac{X_n}{a_n}| \geq \epsilon) = 0, \forall \epsilon > 0.$$

In other words, $X_n = o_p(a_n)$ if and only if $\frac{X_n}{a_n} \xrightarrow{p} 0$.

6.1.1 Relationship of big-O and little-o

O_p and o_p may seem quite similar, and that's because they are! Another way to express $X_n = O_p(a_n)$, is

$$\forall \epsilon \;\; \exists N_\epsilon, \delta_\epsilon \;\; s.t. \forall n > N_\epsilon, \;\; P(|\frac{X_n}{a_n}| \geq \delta_\epsilon) \leq \epsilon.$$

This restatement makes it clear that the values of δ and N are to be found with respect to ϵ. That is, we only have to find one value of N and δ for each *epsilon*, and these can differ across ϵ's.

Using the same notation, $X_n = o_p(a_n)$ can be expressed as

$$\forall \epsilon, \delta \;\; \exists N_{\epsilon,\delta} \;\; s.t. \forall n > N_{\epsilon,\delta}, \;\; P(|\frac{X_n}{a_n}| \geq \delta) \leq \epsilon.$$

o_p is therefore a more general statement, ranging over all values of ϵ and δ, and hence any combination of those two values. In other words, for any given pair of values for ϵ and δ there must be some N that satisfies the above inequality (assuming $X_n = o_p(a_n)$).

Note also, therefore that $o_p(a_n)$ entails $O_p(a_n)$, but that the inverse is not true. If for all ϵ and δ there is some $N_{\epsilon,\delta}$ that satisfies the inequality, then it must be the case that for all ϵ there exists some δ such that the inequality also holds. But just because for some δ_ϵ the inequality holds, this does not mean that it will hold for all δ.

6.2 Notational shorthand and "arithmetic" properties

Expressions like $X_n = o_p\left(\frac{1}{\sqrt{n}}\right)$ do not contain literal identities. Big and little o are merely shorthand ways of expressing how some random variable converges (either to a bound or zero). Suppose for instance that we know $X_n = o_p(\frac{1}{n})$. We also therefore know that $X_n = o_p(\frac{1}{n^{0.5}})$. Analogously, think about an object accelerating at a rate of at least $10ms^{-2}$ – that car is also accelerating at a rate at least $5ms^{-2}$. But it's not the case that $o_p(\frac{1}{n}) = o_p(\frac{1}{\sqrt{n}})$. For instance a car accelerating at least as fast as $5ms^{-2}$ is not necessarily accelerating at least as fast as $10ms^{-2}$.

Hence, when we use stochastic order notation we should be careful to think of it as implying something rather than making the claim that some random variable or expression involving random variables equals some stochastic order.

That being said, we can note some simple implications of combining O_p and/or o_p terms, including:

- $o_p(1) + o_p(1) = o_p(1)$ – this is straightforward: two terms that both converge to zero at the same rate, collectively converge to zero at that rate. Note this is actually just an application of Continuous Mapping Theorem, since If $X_n = o_p(1), Y_n = o_p(1)$ then $X_n \xrightarrow{p} 0, Y_n \xrightarrow{p} 0$ then the addition of these two terms is a continuous mapping function, and therefore $X_n + Y_n \xrightarrow{p} 0, \; \therefore X_n + Y_n = o_p(1)$.
- $O_p(1) + o_p(1) = O_p(1)$ – a term that is bounded in probability ($O_p(1)$) plus a term converging in probability to zero, is bounded in probability.
- $O_p(1)o_p(1) = o_p(1)$ – a bounded probability multiplied by a term that converges (in the same order) to zero itself converges to zero.
- $o_p(R) = R \times o_p(1)$ – again this is easy to see, since suppose $X_n = o_p(R)$, then $X_n/R = o_p(1)$, and so $X_n = Ro_p(1)$.

Further rules, and intuitive explanations for their validity, can be found in Section 2.2 of van der Vaart (1998). The last rule above, however, is worth dwelling on briefly since it makes clear why we use different rate terms (R) in the little-o operator. Consider two rates $R^{(1)} = \frac{1}{\sqrt{n}}$, $R^{(2)} = \frac{1}{\sqrt[3]{2}}$, and some random variable

$Y_n \xrightarrow{p} 0$, that is $Y_n = o_p(1)$. Given the final rule (and remembering the equals signs should not be read literally), if $X_n^{(1)} = o_p(R^{(1)})$, then

$$X_n^{(1)} = \frac{1}{\sqrt{n}} \times Y_n,$$

and if $X_n^{(2)} = o_p(R^{(2)})$, then

$$X_n^{(2)} = \frac{1}{\sqrt[3]{n}} \times Y_n.$$

For each value of Y_n as n approaches infinity, $X_n^{(1)}$ is smaller $X_n^{(2)}$. In other words, $X_n^{(2)}$ will converge in probably towards zero more slowly. This implication of the notation, again,

6.3 Why is this useful?[1]

A simple (trivial) example of this notation is to consider a sequence of random variables X_n with known $\mathbb{E}[X_n] = X$. We can therefore decompose $X_n = X + o_p(1)$, since we know by the Weak Law of Large Numbers that $X_n \xrightarrow{p} X$. This is useful because, without having to introduce explicit limits into our equations, we know that with a sufficiently large n, the second term of our decomposition converges to zero, and therefore we can (in a hand-wavey fashion) ignore it.

Let's consider a more meaningful example. Suppose now that $X_n \sim N(0, n)$. Using known features of normal distributions, we can rearrange this to

$$\frac{X_n}{\sqrt{n}} \sim N(0, 1).$$

There exists some M such that the probability that a value from $N(0, 1)$ exceeds M is less than $\epsilon > 0$, and therefore

$$X_n = O_p(\sqrt{n}).$$

X_n is also little-op of n since

$$\begin{aligned} \frac{X_n}{n} &\sim N(0, \frac{n}{n^2}) \\ &\sim N(0, \frac{1}{n}) \end{aligned}$$

And so we just need to prove the righthand side above is $o_p(1)$. To do so note that:

$$\begin{aligned} P(|N(0, \frac{1}{n})| > \epsilon) &= P(\frac{1}{\sqrt{n}}|N(0, 1)| > \epsilon) \\ &= P(|N(0, 1)| > \sqrt{n}\epsilon) \xrightarrow{p} 0. \end{aligned}$$

The last follows since $\sqrt{n} \to \infty$, and so the probability that the standard normal is greater than ∞ decreases to zero. Hence $X_n = o_p(n)$.

[1] The first two examples in this section are adapted from Ashesh Rambachan's Asymptotics Review lecture slides, from Harvard Math Camp – Econometrics 2018.

$$\lim_{n\to\infty} P\left(\left|\frac{N(0,\frac{1}{n})}{n}\right| \geq \epsilon\right) = 0 = o_p(1),$$

for all $\epsilon > 0$, and therefore that

$$X_n = o_p(n)$$

.

The big-O, little-o notation captures the complexity of the equation or, equivalently, the rate at which it converges. One way to read $X_n = o_p(a_n)$ is that, for any multiple of j, X_n converges in probability to zero at the rate determined by a_n. So, for example, $o_p(a_n^2)$ converges faster than $o_p(a_n)$, since for some random variable X_n, $\frac{X_n}{a_n^2} < \frac{X_n}{a_n}, n > 1$.

When we want to work out the asymptotic limits of a more complicated equation, where multiple terms are affected by the number of observations, if we have a term that converges in probability to zero at a faster rate than others then we can safely ignore that term.

6.4 Worked Example: Consistency of mean estimators

A parameter is "consistent" if it converges in probability to the true parameter as the number of observations increases. More formally, a parameter estimate $\widehat{\theta}$ is consistent if

$$P(|\widehat{\theta} - \theta| \geq \epsilon) \xrightarrow{p} 0,$$

where θ is the true parameter.

One question we can ask is how fast our consistent parameter estimate converges on the true parameter value. This is an "applied" methods problem to the extent that, as researchers seeking to make an inference about the true parameter, and confronted with potentially many ways of estimating it, we want to choose an efficient estimator i.e. one that gets to the truth quickest!

Let's suppose we want to estimate the population mean of X, i.e. $\bar{X}$. Suppose further we have two potential estimators, the sample mean is $\frac{1}{N}\sum_{i=1}^{N} X_i$ and the median is $X_{(N+1)/2}$, where $N = 2n + 1$ (we'll assume an odd number of observations for the ease of calculation) and X is an ordered sequence from smallest to largest.

We know by the Central Limit Theorem that the sample mean

$$\bar{X}_N \sim \mathcal{N}(\theta, \frac{\sigma^2}{N}),$$

and note that I use $\mathcal{N}$ to denote the normal distribution function, to avoid confusion with the total number of observations N.

Withholding the proof, the large-sample distribution of the median estimator can be expressed approximately[2] as

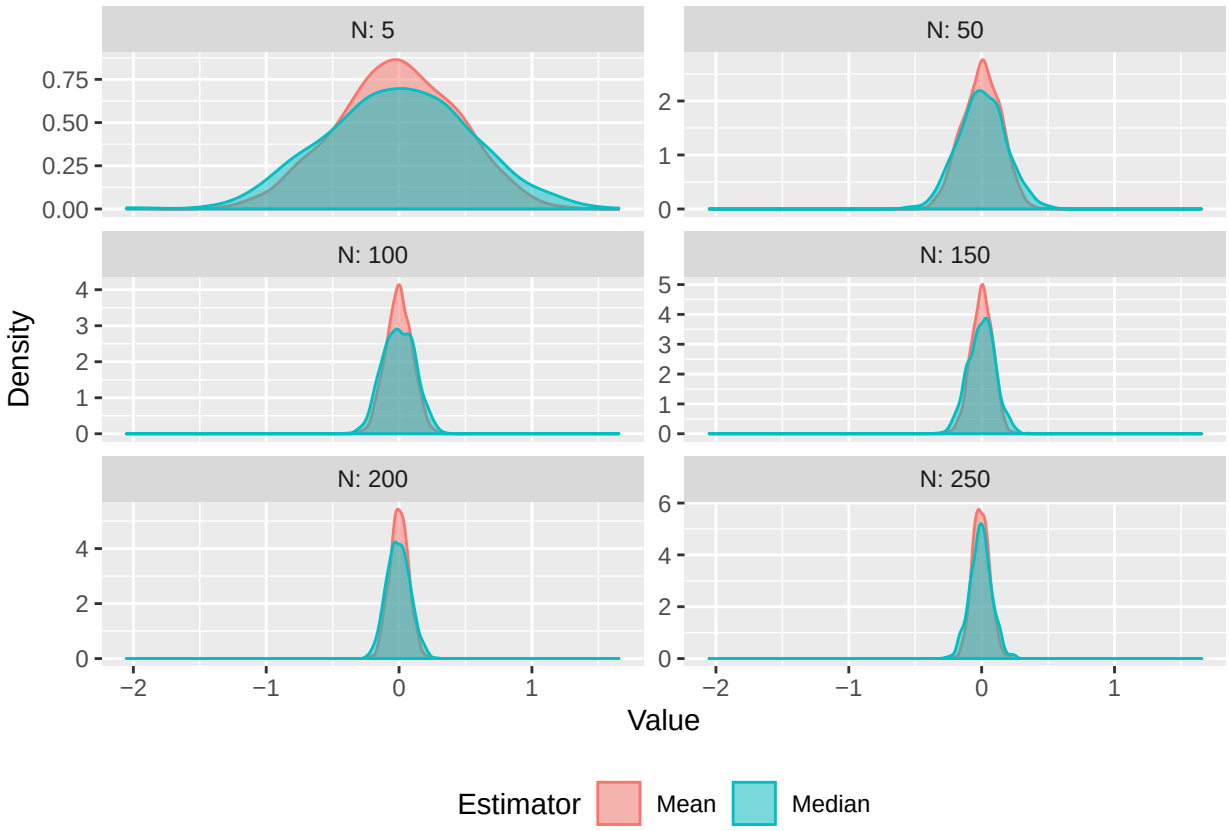
$$\text{Med}(X_1, X_2, ..., X_N) \sim \mathcal{N}(\theta, \frac{\pi\sigma^2}{2N}).$$

[2]See this Wolfram MathWorld post for more information about the exact CLT distribution of sample medians.

How do these estimators perform in practice? Let's first check this via Monte Carlo, by simulating draws of a standard normal distribution with various sizes of N and plotting the resulting distribution of the two estimators:

```
library(tidyverse)
library(ccaPP) # This pkg includes a fast algorithm for the median
```

```
# Compute sample mean and median 1000 times, using N draws from std. normal
rep_sample <- function(N) {
  sample_means <- c()
  sample_medians <- c()
  for (s in 1:1000) {
    sample <- rnorm(N)
    sample_means[s] <- mean(sample)
    sample_medians[s] <- fastMedian(sample)
  }
  return(data.frame(N = N, Mean = sample_means, Median = sample_medians))
}

set.seed(89)
Ns <- c(5,seq(50,250, by = 50)) # A series of sample sizes

# Apply function and collect results, then pivot dataset to make plotting easier
sim_results <- do.call("rbind", lapply(Ns, FUN = function(x) rep_sample(x))) %>%
  pivot_longer(-N, names_to = "Estimator", values_to = "estimate")

ggplot(sim_results, aes(x = estimate, color = Estimator, fill = Estimator)) +
  facet_wrap(~N, ncol = 2, scales = "free_y", labeller = "label_both") +
  geom_density(alpha = 0.5) +
  labs(x = "Value", y = "Density") +
  theme(legend.position = "bottom")
```



Here we can see that for both the mean and median sample estimators, the distribution of parameters is normally distributed around the true mean ($\theta = 0$). The variance of the sample mean distribution, however, shrinks faster than that of the sample median estimator. In other words, the sample mean is more "efficient" (in fact it is the most efficient estimator). Efficiency here captures what we noted mathematically above – that the rate of convergence on the true parameter (i.e. the rate at which the estimator converges on zero) is faster for the sample mean than the median.

Note that both estimators are therefore unbiased (they are centred on θ), normally distributed, and are consistent (the sampling distributions shrink towards the true parameter as N increases), but that the variances shrinks at slightly different rates.

We can quantify this using little-o notation and the behaviour of these estimators with large-samples. First, we can define the estimation errors of the mean and median respectively as

$$\begin{aligned}\psi_{\text{Mean}} &= \hat{\theta} - \theta \\ &= \mathcal{N}(\theta, \frac{\sigma^2}{N}) - \mathcal{N}(\theta, 0) \\ &= \mathcal{N}(0, \frac{\sigma^2}{N}).\end{aligned}$$

Similarly,

$$\begin{aligned}\psi_{\text{Med.}} &= \mathcal{N}(\theta, \frac{\pi\sigma^2}{2N}) - \mathcal{N}(\theta, 0) \\ &= \mathcal{N}(0, \frac{\pi\sigma^2}{2N}).\end{aligned}$$

With both mean and median expressions, we can see that the error of the estimators is centered around zero

(i.e. it is unbiased), and that the dispersion of the error around zero decreases as N increases. Given earlier discussions in this chapter, we can rearrange both to find out their rate of convergence.

For the sample mean:

$$\psi_{\text{Mean}} = \frac{1}{\sqrt{N}}\mathcal{N}(0, \sigma^2)$$
$$\frac{\psi_{\text{Mean}}}{N^{-0.5}} = \mathcal{N}(0, \sigma^2)$$

We know that for a normal distribution, there will be some M_ϵ, N_ϵ, such that $P(|\mathcal{N}(0, \sigma^2)| \geq M_\epsilon) < \epsilon$, and hence:

$$\psi_{\text{Mean}} = O_p(\frac{1}{\sqrt{N}}).$$

Similarly, for the sample median:

$$\begin{aligned}
\psi_{\text{Med.}} &= \mathcal{N}(0, \frac{\pi\sigma^2}{2N}) \\
&= \left(\frac{\pi}{2N}\right)^{0.5}\mathcal{N}(0, \sigma^2) \\
\psi_{\text{Med.}} / \left(\frac{\pi}{2N}\right)^{0.5} &= \mathcal{N}(0, \sigma^2) \\
\psi_{\text{Med.}} &= O_p\left(\left[\frac{\pi}{2N}\right]^{0.5}\right) \\
&= O_p\left(\frac{\sqrt{\pi}}{\sqrt{2N}}\right).
\end{aligned}$$

Now we can see that the big-op of the sample median's estimating error is "slower" (read: larger) than the big-op of the sample mean, meaning that the sample mean converges on the true parameter with fewer observations than the sample median.

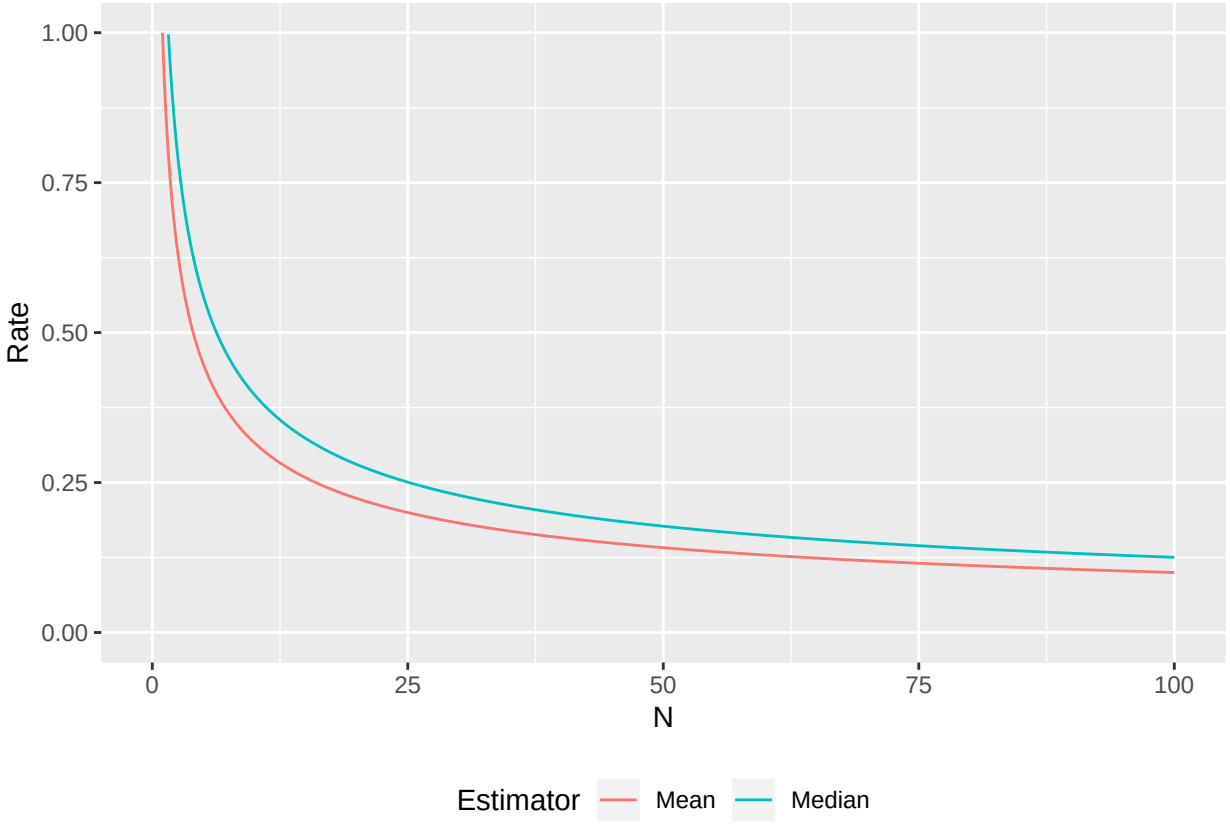
Another, easy way to see the intuition behind this point is to note that at intermediary steps in the above rearrangements:

$$\psi_{\text{Mean}} = \frac{1}{\sqrt{N}}\mathcal{N}(0, \sigma^2)$$
$$\psi_{\text{Med.}} = \frac{\sqrt{\pi}}{\sqrt{2N}}\mathcal{N}(0, \sigma^2),$$

and so, for any sized sample, the estimating error of the median is larger than that of the mean. To visualise this, we can plot the estimation error as a function of N using the rates derived above:

```
N <- seq(0.01,100, by = 0.01)
mean_convergence <- 1/sqrt(N)
median_convergence <- sqrt(pi)/sqrt(2*N)

plot_df <- data.frame(N, Mean = mean_convergence, Median = median_convergence) %>%
  pivot_longer(-N, names_to = "Estimator", values_to = "Rate")

ggplot(plot_df, aes(x = N, y = Rate, color = Estimator)) +
  geom_line() +
  ylim(0,1) +
  theme(legend.position = "bottom")
```

Figure 6.1: Simulated distribution of sample mean and median estimators for different sized samples.



Note that the median rate line is always above the mean line for all N (though not by much) – it therefore has a slower convergence.

Chapter 7

Delta Method

7.1 Delta Method in Plain English

The Delta Method (DM) states that we can approximate the asymptotic behaviour of functions over a random variable, if the random variable is itself asymptotically normal. In practice, this theorem tells us that even if we do not know the expected value and variance of the function $g(X)$ we can still approximate it reasonably. Note that by Central Limit Theorem we know that several important random variables and estimators are asymptotically normal, including the sample mean. We can therefore approximate the mean and variance of some transformation of the sample mean using its variance.

More specifically, suppose that we have some sequence of random variables X_n, such that as $n \to \infty$

$$X_n \sim N(\mu, \frac{\sigma^2}{n}).$$

We can rearrange this statement, to capture that the difference between the random variable and some constant μ converges to a normal distribution around zero, with a variance determined by the number of observations:[1]

$$(X_n - \mu) \sim N(0, \frac{\sigma^2}{n}).$$

Further rearrangement yields

$$(X_n - \mu) \sim \frac{\sigma}{\sqrt{n}} N(0,1)$$
$$\frac{\sqrt{n}(X_n - \mu)}{\sigma} \sim N(0,1),$$

by first moving the finite variance and n terms outside of the normal distribution, and then dividing through.

Given this, if g is some smooth function (i.e. there are no discontinuous jumps in values) then the Delta Method states that:

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \approx N(0,1),$$

where g' is the first derivative of g. Rearranging again, we can see that

[1] There are clear parallels here to how we expressed estimator consistency.

$$g(X_n) \approx N\left(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}\right).$$

Note that the statement above is an approximation because $g(X_n) = g(\mu) + g'(\mu)(\mu \hat{-} \mu + g''(\mu)\frac{(X_n-\mu)^2}{2!} + ...$, i.e. an infinite sum. The Delta Method avoids the infinite regress by ignoring higher order terms (Liu, 2012). I return to this point below in the proof.

DM also generalizes to multidimensional functions, where instead of converging on the standard normal the random variable must converge in distribution to a multivariate normal, and the derivatives of g are replaced with the gradient of g (a vector of all partial derivatives).[^fn_gradient] For the sake of simplicity I do not prove this result here, and instead focus on the univariate case.

$$\nabla g = \begin{bmatrix} \frac{dg}{dx_1} \\ \frac{dg}{dx_2} \\ \vdots \\ \frac{dg}{dx_n} \end{bmatrix}$$

7.2 Proof

Before offering a full proof, we need to know a little bit about Taylor Series and Taylor's Theorem. I briefly outline this concept here, then show how this expansion helps to prove DM.

7.2.1 Taylor's Series and Theorem

Suppose we have some continuous function g that is infinitely differentiable. By that, we mean that we mean some function that is continuous over a domain, and for which there is always some further derivative of the function. Consider the case $g(x) = e^{2x}$,

$$\begin{aligned} g'(x) &= 2e^{2x} \\ g''(x) &= 4e^{2x} \\ g'''(x) &= 8e^{2x} \\ g''''(x) &= 16e^{2x} \\ &... \end{aligned}$$

For any integer k, the kth derivative of $g(x)$ is defined. An interesting non-infinitely differentiable function would be $g(x) = |x|$ where $-\infty < x < \infty$. Here note that when $x > 0$, the first order derivative is 1 (the function is equivalent to x), and similarly at $x < 0$, the first order derivative is -1 (the function is equivalent to $-x$). When $x = 0$, however, the first derivative is undefined – the first derivative jumps discontinuously.

The Taylor Series for an infinitely differentiable function at a given point $x = p$ is an expansion of that function in terms of an infinite sum:

$$g(x) = g(p) + g'(p)(x-p) + \frac{g''(p)}{2!}(x-p)^2 + \frac{g'''(p)}{3!}(x-p)^3 + ...$$

Taylor Series are useful because they allow us to approximate a function at a lower polynomial order, using Taylor's Theorem. This Theorem loosely states that, for a given point $x = p$, we can approximate a continuous and k-times differentiable function to the jth order using the Taylor Series up to the jth derivative. In other words, if we have some continuous differentiable function $g(x)$, its first-order approximation (i.e. its linear approximation) at point p is defined as

$$g(p) + g'(p)(x - p).$$

To make this more concrete, consider the function $g(x) = e^x$. The Taylor Series expansion of g at point $x = 0$ is

$$g(x) = g(0) + g'(0)(x - 0) + \frac{g''(0)}{2!}(x - 0)^2 + \frac{g'''(0)}{3!}(x - 0)^3 + ...$$

So up to the first order, Taylors Theorem states that

$g(x) \approx g(0) + g'(0)(x - 0) = 1 + x,$

which is the line tangent to e^x at $x = 0$. If we consider up to the second order (the quadratic approximation) our fit would be better, and even more so if we included the third, fourth, fifth orders and so on, up until the ∞th order – at which point the Taylor Approximation is the function precisely.

7.2.2 Proof of Delta Method

Given Taylor's Theorem, we know that so long as g is a continuous and derivable up to the kth derivative, where $k \geq 2$, then at the point μ:

$$g(X_n) \approx g(\mu) + g'(\mu)(X_n - \mu).$$

Subtracting $g(\mu)$ we have:

$$(g(X_n) - g(\mu)) \approx g'(\mu)(X_n - \mu).$$

We know by CLT and our assumptions regarding X_n that $(X_n - \mu) \xrightarrow{d} N(0, \frac{\sigma^2}{n})$. Therefore we can rewrite the above as

$$(g(X_n) - g(\mu)) \approx g'(\mu)N(0, \frac{\sigma^2}{n}),$$

Hence, by the properties of normal distributions (multiplying by a constant, adding a constant):

$$g(X_n) \approx N\left(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}\right) \quad \square$$

7.3 Applied example

Bowler et al. (2006) use the DM to provide confidence intervals for predicted probabilities generated from a logistic regression. Their study involves surveying politicians' attitudes toward electoral rule changes. They estimate a logistic model of the support for change on various features of the politicians including whether they won under existing electoral rules or not. To understand how winning under existing rules affects attitudes, they then generate the predicted probabilities for losers and winners separately.

Generating predicted probabilities from a linear regression involves a non-linear transformation of an asymptotically normal parameter (the logistic coefficient), and therefore we must take account of this transformation when variance of the predicted probability.

To generate the predicted probability we use the equation

$$\hat{p} = \frac{e^{(\hat{\alpha}+\hat{\beta}_1 X_1 + \ldots + \beta_n X_n)}}{1 + e^{(\hat{\alpha}+\hat{\beta}_1 X_1 + \ldots + \hat{\beta}_n X_n)}},$$

where $\hat{p}$ is the predicted probability. Estimating the variance around the predicted probability is therefore quite difficult – it involves multiple estimators, and non-linear transformations. But we do know that, assuming i.i.d and correct functional form, the estimating error of the logistic equation is asymptotically multivariate normal on the origin. And so the authors can use DM to calculate 95 percent confidence intervals. In general, the delta method is a useful way of estimating standard and errors and confidence intervals when using (but not limited to) logistic regression and other models involving non-linear transformations of model parameters.

7.4 Alternative strategies

The appeal of the delta method is that it gives an analytic approximation of a function's distribution, using the asymptotic properties of some more (model) parameter. But there are alternative methods to approximating these distributions (and thus standard errors) that do not rely on deriving the order conditions of that function.

One obvious alternative is the bootstrap. For a given transformation of a random variable, calculate the output of the function B times using samples of the same size as the original sample, but with replacement and take either the standard deviation or the a and $1-a$ percentiles of the resultant parameter distribution. This method does not require the user to calculate the derivative of a function. It is a non-parametric alternative that simply approximates the distribution itself, rather than approximates the parameters of a parametric distribution.

The bootstrap is computationally more intensive (requiring B separate samples and calculations etc.) but, on the other hand, is less technical to calculate. Moreover, the Delta Method's approximation is limited analytically by the number of terms considered in the Taylor Series expansion. While the first order Taylor Theorem may be reasonable, it may be imprecise. To improve the precision one has to undertake to find the second, third, fourth etc. order terms (which may be analytically difficult). With bootstrapping, however, you can improve precision simply by taking more samples (increasing B) (King et al., 2000).

Given the ease with which we can acquire and deploy computational resources now, perhaps the delta method is no longer as useful in applied research. But the proof and asymptotic implications remain statistically interesting and worth knowing.

Chapter 8

Frisch-Waugh-Lovell Theorem

8.1 Theorem in plain English

The Frisch-Waugh-Lovell Theorem (FWL; after the initial proof by Frisch and Waugh (1933), and later generalisation by Lovell (1963)) states that:

Any predictor's regression coefficient in a multivariate model is equivalent to the regression coefficient estimated from a bivariate model in which the residualised outcome is regressed on the residualised component of the predictor; where the residuals are taken from models regressing the outcome and the predictor on all other predictors in the multivariate regression (separately).

More formally, assume we have a multivariate regression model with k predictors:

$$\hat{y} = \hat{\beta}_1 x_1 + ...\hat{\beta}_k x_k + \epsilon. \tag{8.1}$$

FWL states that every $\hat{\beta}_j$ in Equation 8.1 is equal to $\hat{\beta}_j^*$, and the residual $\epsilon = \epsilon^*$ in:

$$\epsilon^y = \hat{\beta}_j^* \epsilon^{x_j} + \epsilon^* \tag{8.2}$$

where:

$$\begin{aligned} \epsilon^y &= y - \sum_{k \neq j} \hat{\beta}_k^y x_k \\ \epsilon^{x_j} &= x_j - \sum_{k \neq j} \hat{\beta}_k^{x_j} x_k. \end{aligned} \tag{8.3}$$

where $\hat{\beta}_k^y$ and $\hat{\beta}_k^{x_j}$ are the regression coefficients from two separate regression models of the outcome (omitting x_j) and x_j respectively.

In other words, FWL states that each predictor's coefficient in a multivariate regression explains that variance of y not explained by both the other k-1 predictors' relationship with the outcome and their relationship with that predictor, i.e. the independent effect of x_j.

8.2 Proof

8.2.1 Primer: Projection matrices[1]

We need two important types of projection matrices to understand the linear algebra proof of FWL. First, the prediction matrix that was introduced in Chapter 4:

$$P = X(X'X)^{-1}X'. \tag{8.4}$$

Recall that this matrix, when applied to an outcome vector (y), produces a set of predicted values ($\hat{y}$). Reverse engineering this, note that $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Py$.

Since Py produces the predicted values from a regression on X, we can define its complement, the residual maker:

$$M = I - X(X'X)^{-1}X', \tag{8.5}$$

since $My = y - X(X'X)^{-1}X'y \equiv y - Py \equiv y - X\hat{\beta} \equiv \hat{\epsilon}$, the residuals from regressing Y on X.

Given these definitions, note that M and P are complementary:

$$\begin{aligned} y &= \hat{y} + \hat{\epsilon} \\ &\equiv Py + My \\ Iy &= Py + My \\ Iy &= (P + M)y \\ I &= P + M. \end{aligned} \tag{8.6}$$

With these projection matrices, we can express the FWL claim (which we need to prove) as:

$$\begin{aligned} y &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\epsilon} \\ M_1 y &= M_1 X_2 \hat{\beta}_2 + \hat{\epsilon}, \end{aligned} \tag{8.7}$$

8.2.2 FWL Proof [2]

Let us assume, as in Equation 8.7 that:

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\epsilon}. \tag{8.8}$$

First, we can multiply both sides by the residual maker of X_1:

$$M_1 Y = M_1 X_1 \hat{\beta}_1 + M_1 X_2 \hat{\beta}_2 + M_1 \hat{\epsilon}, \tag{8.9}$$

which first simplifies to:

[1] Citation Based on lecture notes from the University of Oslo's "Econometrics – Modelling and Systems Estimation" course (author attribution unclear), and Davidson et al. (2004).}

[2] Citation: Adapted from York University, Canada's wiki for statistical consulting.

$$M_1 Y = M_1 X_2 \hat{\beta}_2 + M_1 \hat{\epsilon}, \tag{8.10}$$

because $M_1 X_1 \hat{\beta}_1 \equiv (M_1 X_1)\hat{\beta}_1 \equiv 0\hat{\beta}_1 = 0$. In plain English, by definition, all the variance in X_1 is explained by X_1 and therefore a regression of X_1 on itself leaves no part unexplained so $M_1 X_1$ is zero.[3]

Second, we can simplify this equation further because, by the properties of OLS regression, X_1 and ϵ are orthogonal. Therefore the residual of the residuals are the residuals! Hence:

$$M_1 Y = M_1 X_2 \hat{\beta}_2 + \hat{\epsilon} \quad \square.$$

8.2.3 Interesting features/extensions

A couple of interesting features come out of the linear algebra proof:

- FWL also holds for bivariate regression when you first residualise Y and X on a $n \times 1$ vector of 1's (i.e. the constant) – which is like demeaning the outcome and predictor before regressing the two.
- X_1 and X_2 are technically sets of mutually exclusive predictors i.e. X_1 is an $n \times k$ matrix $\{X_1, ..., X_k\}$, and X_2 is an $n \times m$ matrix $\{X_{k+1}, ..., X_{k+m}\}$, where β_1 is a corresponding vector of regression coefficients $\beta_1 = \{\gamma_1, ..., \gamma_k\}$, and likewise $\beta_2 = \{\delta_1, ..., \delta_m\}$, such that:

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 \\ &= X_1\hat{\gamma}_1 + ... + X_k\hat{\gamma}_k + X_{k+1}\hat{\delta}_1 + ... + X_{k+m}\hat{\delta}_m, \end{aligned}$$

Hence the FWL theorem is exceptionally general, applying not only to arbitrarily long coefficient vectors, but also enabling you to back out estimates from any partitioning of the full regression model.

8.3 Coded example

```
set.seed(89)

## Generate random data
df <- data.frame(y = rnorm(1000,2,1.5),
                 x1 = rnorm(1000,1,0.3),
                 x2 = rnorm(1000,1,4))

## Partial regressions

# Residual of y regressed on x1
y_res <- lm(y ~ x1, df)$residuals

# Residual of x2 regressed on x1
x_res <- lm(x2 ~ x1, df)$residuals

resids <- data.frame(y_res, x_res)

## Compare the beta values for x2
```

[3] Algebraically, $M_1 X_1 = (I - X_1(X_1'X_1)^{-1}X_1')X_1 = X_1 - X_1(X_1'X_1)^{-1}X_1'X = X_1 - X_1 I = X_1 - X_1 = 0$.

```r
# Multivariate regression:
summary(lm(y~x1+x2, df))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.451 -1.001 -0.039  1.072  5.320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.33629    0.16427  14.222   <2e-16 ***
## x1          -0.31093    0.15933  -1.952   0.0513 .
## x2           0.02023    0.01270   1.593   0.1116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.535 on 997 degrees of freedom
## Multiple R-squared:  0.006252,Adjusted R-squared:  0.004258
## F-statistic: 3.136 on 2 and 997 DF,  p-value: 0.04388
```

```r
# Partials regression
summary(lm(y_res ~ x_res, resids))
```

```
##
## Call:
## lm(formula = y_res ~ x_res, data = resids)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.451 -1.001 -0.039  1.072  5.320
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.921e-17  4.850e-02   0.000    1.000
## x_res        2.023e-02  1.270e-02   1.593    0.111
##
## Residual standard error: 1.534 on 998 degrees of freedom
## Multiple R-squared:  0.002538,Adjusted R-squared:  0.001538
## F-statistic: 2.539 on 1 and 998 DF,  p-value: 0.1114
```

Note: This isn't an exact demonstration because there is a degrees of freedom of error. The (correct) multivariate regression degrees of freedom is calculated as $N - 3$ since there are three variables. In the partial regression the degrees of freedom is $N - 2$. This latter calculation does not take into account the additional loss of freedom as a result of partialling out X_1.

8.4 Application: Sensitivity analysis

Cinelli and Hazlett (2020) develop a series of tools for researchers to conduct sensitivity analyses on regression models, using an extension of the omitted variable bias framework. To do so, they use FWL to motivate this bias. Suppose that the full regression model is specified as:

$$Y = \hat{\tau}D + X\hat{\beta} + \hat{\gamma}Z + \hat{\epsilon}_{\text{full}}, \tag{8.11}$$

where $\hat{\tau}, \hat{\beta}, \hat{\gamma}$ are estimated regression coefficients, D is the treatment variable, X are observed covariates, and Z are unobserved covariates. Since, Z is unobserved, researchers measure:

$$Y = \hat{\tau}_{\text{Obs.}}D + X\hat{\beta}_{\text{Obs.}} + \epsilon_{\text{Obs}} \tag{8.12}$$

By FWL, we know that $\hat{\tau}_{\text{Obs.}}$ is equivalent to the coefficient of regressing the residualised outcome (with respect to X), on the residualised outcome of D (again with respect to X). Call these two residuals Y_r and D_r.

And recall that the regression model for the final-stage of the partial regressions is bivariate ($Y_r \sim D_r$). Conveniently, a bivariate regression coefficient can be expressed in terms of the covariance between the left-hand and right-hand side variables:[4]

$$\hat{\tau}_{\text{Obs.}} = \frac{cov(D_r, Y_r)}{var(D_r)}. \tag{8.13}$$

Note that given the full regression model in Equation 8.11, the partial outcome Y_r is actually composed of the elements $\hat{\tau}D_r + \hat{\gamma}Z_r$, and so:

$$\hat{\tau}_{\text{Obs.}} = \frac{cov(D_r, \hat{\tau}D_r + \hat{\gamma}Z_r)}{var(D_r)} \tag{8.14}$$

Next, we can expand the covariance using the expectation rule that $cov(A, B + C) = cov(A, B) + cov(A, C)$ and since $\hat{\tau}, \hat{\gamma}$ are scalar, we can move them outside the covariance functions:

$$\hat{\tau}_{\text{Obs.}} = \frac{\hat{\tau}cov(D_r, D_r) + \hat{\gamma}cov(D_r, Z_r)}{var(D_r)} \tag{8.15}$$

Since $cov(A, A) = var(A)$ and therefore:

$$\hat{\tau}_{\text{Obs.}} = \frac{\hat{\tau}var(D_r) + \hat{\gamma}cov(D_r, Z_r)}{var(D_r)} \equiv \hat{\tau} + \hat{\gamma}\frac{cov(D_r, Z_r)}{var(D_r)} \equiv \hat{\tau} + \hat{\gamma}\hat{\delta} \tag{8.16}$$

Frisch-Waugh is so useful because it simplifies a multivariate equation into a bivariate one. While computationally this makes zero difference (unlike in the days of hand computation), here it allows us to use a convenient expression of the bivariate coefficient to show and quantify the bias when you run a regression in the presence of an unobserved confounder. Moreover, note that in Equation 8.14, we implicitly use FWL again since we know that the non-stochastic aspect of Y not explained by X are the residualised components of the full Equation 8.11.

8.4.1 Regressing the partialled-out X on the full Y

In Mostly Harmless Econometrics (MHE; Angrist and Pischke (2009)), the authors note that you also get an identical coefficient to the full regression if you regress the residualised predictor on the non-residualised Y. We can use the OVB framework above to explain this case.

Let's take the full regression model as:

[4] If $y = \hat{\alpha} + \hat{\beta}x + \epsilon$, then by least squares $\hat{\beta} = \frac{cov(x,y)}{var(x)}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

$$Y = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\epsilon}. \tag{8.17}$$

MHE states that:

$$Y = \hat{\beta}_1 M_2 X_1 + \hat{\epsilon}. \tag{8.18}$$

Note that this is just FWL, except we have not also residualised Y. Our aim is to check whether there is any bias in the estimated coefficient from this second equation. As before, since this is a bivariate regression we can express the coefficient as:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{cov(M_2 X_1, Y)}{var(M_2 X_1)} \\
&= \frac{cov(M_2 X_1, \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2)}{var(M_2 X_1)} \\
&= \hat{\beta}_1 \frac{cov(M_2 X_1, X_1)}{var(M_2 X_1)} + \hat{\beta}_2 \frac{cov(M_2 X_1, X_2)}{var(M_2 X_1)} \\
&= \hat{\beta}_1 + \hat{\beta}_2 \times 0 \\
&= \hat{\beta}_1
\end{aligned} \tag{8.19}$$

This follows from two features. First, $cov(M_2 X_1, X_1) = var(M_2 X_1)$. Second, it is clear that $cov(M_2 X_1, X_2) = 0$ because $M_2 X_1$ is X_1 stripped of any variance associated with X_2 and so, by definition, they do not covary. Therefore, we can recover the unbiased regression coefficient using an adapted version of FWL where we do not residualise Y – as stated in MHE.

Chapter 9

Positive Definite Matrices

9.1 Terminology

A $n \times n$ symmetric matrix M is positive definite (PD) if and only if $x'Mx > 0$, for all non-zero $x \in \mathbb{R}^n$. For example, take the 3×3 identity matrix, and a column vector of non-zero real numbers $[a, b, c]$:

$$\begin{aligned}
\begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\
&= \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\
&= a^2 + b^2 + c^2.
\end{aligned}$$

Since by definition a^2, b^2, and c^2 are all greater than zero (even if a, b, or c are negative), their sum is also positive.

A matrix is positive semi-definite (PSD) if and only if $x'Mx \geq 0$ for all non-zero $x \in \mathbb{R}^n$. Note that PSD differs from PD in that the transformation of the matrix is no longer strictly positive.

One known feature of matrices (that will be useful later in this chapter) is that if a matrix is symmetric and idempotent then it will be positive semi-definite. Take some non-zero vector x, and a symmetric, idempotent matrix A. By idempotency we know that $x'Ax = x'AAx$. By symmetry we know that $A' = A$, and therefore:

$$\begin{aligned}
x'Ax &= x'AAx \\
&= x'A'Ax \\
&= (Ax)'Ax \geq 0,
\end{aligned}$$

and hence PSD.[1]

9.1.1 Positivity

Both PD and PSD are concerned with positivity. For scalar values like -2, 5, 89, positivity simply refers to their sign – and we can tell immediately whether the numbers are positive or not. Some functions are also (strictly) positive. Think about $f(x) = x^2 + 1$. For all $x \in \mathbb{R}$, $f(x) \geq 1 > 0$. PD and PSD extend this notion

[1] This short proof is taken from this discussion.

of a positivity to matrices, which is useful when we consider multidimensional optimisation problems or the combination of matrices.

While for abstract matrices like the identity matrix it is easy to verify PD and PSD properties, for more complicated matrices we often require other more complicated methods. For example, we know that a symmetric matrix is PSD if and only if all its eigenvalues are non-negative. The eigenvalue λ is a scalar such that, for a matrix A and non-zero $n \times 1$ vector v, $A \cdot v = \lambda \cdot v$. While I do not explore this further in this chapter, there are methods available for recovering these values from the preceding equation. Further discussion of the full properties of PD and PSD matrices can be found here as well as in print (e.g. Horn and Johnson, 2013, Chapter 7).

9.2 $A - B$ is PSD iff $B^{-1} - A^{-1}$ is PSD

Wooldridge's list of 10 theorems does not actually include a general claim about the importance P(S)D matrices. Instead, he lists a very specific feature of two PD matrices. In plain English, this theorem states that, assuming A and B are both positive definite, $A - B$ is positive semi-definite if and only if the inverse of B minus the inverse of A is positive semi-definite.

Before we prove this theorem, it's worth noting a few points that are immediately intuitive from its statement. Note that the theorem moves from PD matrices to PSD matrices. This is because we are subtracting one matrix from another. While we know A and B are both PD, if they are both equal then $x'(A - B)x$ will equal zero. For example, if $A = B = I_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, then $A - B = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Hence, $x'(A - B)x = 0$ and therefore $A - B$ is PSD, but not PD.

Also note that this theorem only applies to a certain class of matrices, namely those we know to be PD. This hints at the sort of applied relevance this theorem may have. For instance, we know that variance is a strictly positive quantity.

The actual applied relevance of this theorem is not particularly obvious, at least from the claim alone. In his post, Wooldridge notes that he repeatedly uses this fact 'to show the asymptotic efficiency of various estimators.' In his Introductory Economics textbook (2013), for instance, Wooldridge makes use of the properties of PSD matrices in proving that the Gauss-Markov (GM) assumptions ensure that OLS is the best, linear, unbiased estimator (BLUE). And, more generally, PD and PSD matrices are very helpful in optimisation problems (of relevance to machine learning too). Neither appear to be direct applications of this specific, bidirectional theorem. In the remainder of this chapter, therefore, I prove the theorem itself for completeness. I then broaden the discussion to explore how PSD properties are used in Wooldridge's BLUE proof as well as discuss the more general role of PD matrices in optimisation problems.

9.2.1 Proof

The proof of Wooldridge's actual claim is straightforward. In fact, given the symmetry of the proof, we only need to prove one direction (i.e. if $A - B$ is PSD, then $A^{-1} - B^{-1}$ is PSD.)

Let's assume, therefore, that $A - B$ is PSD. Hence:

$$\begin{aligned} x'(A - B)x &\geq 0 \\ x'Ax - xBx &\geq 0 \\ x'Ax &\geq x'Bx \\ Ax &\geq Bx \\ A &\geq B. \end{aligned}$$

Next, we can invert our two matrices while maintaining the inequality:

$$A^{-1}AB^{-1} \geq A^{-1}BB^{-1}$$
$$IB^{-1} \geq A^{-1}I$$
$$B^{-1} \geq A^{-1}.$$

Finally, we can just remultiply both sides of the inequality by our arbitrary non-zero vector:

$$x'B^{-1} \geq x'A^{-1}$$
$$x'B^{-1}x \geq x'A^{-1}x$$
$$x'B^{-1}x - x'A^{-1}x \geq 0$$
$$x'(B^{-1} - A^{-1})x \geq 0.$$

Proving the opposite direction (if $B^{-1} - A^{-1}$ is PSD then $A - B$ is PSD) simply involves replacing A with B^{-1} an B with A^{-1} and vice versa throughout the proof, since $(A^{-1})^{-1} = A$. $\square$

9.3 Applications

9.3.1 OLS as the best linear unbiased estimator (BLUE)

First, let's introduce the four Gauss-Markov assumptions. I only state these briefly, in the interest of space, spending a little more time explaining the rank of a matrix. Collectively, these assumptions guarantee that the linear regression estimates $\widehat{\beta}$ are BLUE (the best linear unbiased estimator of β).

1. The true model is linear such that $y = X\beta + u$, where y is a $n \times 1$ vector, X is a $n \times (k+1)$ matrix, and u is an unobserved $n \times 1$ vector.
2. The rank of X is $(k+1)$ (full-rank), i.e. that there are no linear dependencies among the variables in X. To understand what the rank of matrix denotes, consider the following 3×3 matrix:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{bmatrix}$$

Note that the third row of M_1 is just two times the first column. They are therefore entirely linearly dependent, and so not separable. The number of independent rows (the rank of the matrix) is therefore 2. One way to think about this geometrically, as in Chapter 3, is to plot each row as a vector. The third vector would completely overlap the first, and so in terms of direction we would not be able to discern between them. In terms of the span of these two columns, moreover, there is no point that one can get to using a combination of both that one could not get to by scaling either one of them.

A slightly more complicated rank-deficient (i.e. not full rank) matrix would be:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$$

Here note that the third row is not scalar multiple of either other column. But, it is a linear combination of the other two. If rows 1, 2, and 3 are represented by a, b, and c respectively, then $c = 2a + b$. Again, geometrically, there is no point that the third row vector can take us to which cannot be achieved using only the first two rows.

An example of a matrix with full-rank, i.e. no linear dependencies, would be:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

It is easy to verify that M_1 and M_2 are rank-deficient, whereas M_3 is of full-rank in R:

```
M1 <- matrix(c(1,0,2,0,1,0,0,0,0), ncol = 3)
M1_rank <- qr(M1)$rank

M2 <- matrix(c(1,0,2,0,1,1,0,0,0), ncol = 3)
M2_rank <- qr(M2)$rank

M3 <- matrix(c(1,0,2,0,1,0,0,0,1), ncol = 3)
M3_rank <-qr(M3)$rank

print(paste("M1 rank:", M1_rank))
```

```
## [1] "M1 rank: 2"
```

```
print(paste("M2 rank:", M2_rank))
```

```
## [1] "M2 rank: 2"
```

```
print(paste("M3 rank:", M3_rank))
```

```
## [1] "M3 rank: 3"
```

3. $\mathbb{E}[u|X] = 0$ i.e. that the model has zero conditional mean or, in other words, our average error is zero.
4. $\text{Var}(u_i|X) = \sigma^2, \text{Cov}(u_i, u_j|X) = 0$ for all $i \neq j$, or equivalently that $Var(u|X) = \sigma^2 I_n$. This matrix has diagonal elements all equal to σ^2 and all off-diagonal elements equal to zero.

BLUE states that the regression coefficient vector $\hat{\beta}$ is the best, or lowest variance, estimator of the true β. Wooldridge (2013) has a nice onproof of this claim (p.812). Here I unpack hisi proof in slightly more detail, noting specifically how PD matrices are used.

To begin our proof of BLUE, let us denote any other linear estimator as $\tilde{\beta} = A'y$, where A is some $n \times (k+1)$ matrix consisting of functions of X.

We know by definition that $y = X\beta + u$ and therefore that:

$$\tilde{\beta} = A'(X\beta + u) = A'X\beta + A'u.$$

The conditional expectation of $\tilde{\beta}$ can be expressed as:

$$\mathbb{E}(\tilde{\beta}|X) = A'X\beta + \mathbb{E}(A'u|X),$$

and since A is a function of X, we can move it outside the expectation:

$$\mathbb{E}(\tilde{\beta}|X) = A'X\beta + A'\mathbb{E}(u|X).$$

By the GM assumption no. 3, we know that $\mathbb{E}(u|X) = 0$, therefore:

$$\mathbb{E}(\tilde{\beta}|X) = A'X\beta.$$

Since we are only comparing $\hat{\beta}$ against other unbiased estimators, we know the conditional mean of any other estimator must equal the true parameter, and therefore that

$$A'X\beta = \beta.$$

The only way that this is true is if $A'X = I$. Hence, we can rewrite our estimator as

$$\tilde{\beta} = \beta + A'u.$$

The variance of our estimator $\tilde{\beta}$ then becomes

$$\begin{aligned} Var(\tilde{\beta}|X) &= (\beta - [\beta + A'u])(\beta - [\beta + A'u])' \\ &= (A'u)(A'u)' \\ &= A'uu'A \\ &= A'[\text{Var}(u|X)]A \\ &= \sigma^2 A'A, \end{aligned}$$

since by GM assumption no. 4 the variance of the errors is a constant scalar σ^2.

Hence:

$$\begin{aligned} \text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}|X) &= \sigma^2 A'A - \sigma^2(X'X)^{-1} \\ &= \sigma^2[A'A - (X'X)^{-1}]. \end{aligned}$$

We know that $A'X = I$, and so we can manipulate this expression further:

$$\begin{aligned} \text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}|X) &= \sigma^2[A'A - (X'X)^{-1}] \\ &= \sigma^2[A'A - A'X(X'X)^{-1}X'A] \\ &= \sigma^2 A'[A - X(X'X)^{-1}X'A] \\ &= \sigma^2 A'[I - X(X'X)^{-1}X']A \\ &= \sigma^2 A'MA. \end{aligned}$$

Note that we encountered M in the previous chapter. It is the residual maker, and has the known property of being both symmetric and idempotent. Recall from the first section that we know any symmetric, idempotent matrix is positive semi-definite, and so we know that

$$\text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}|X) \geq 0,$$

and thus that the regression estimator $\hat{\beta}$ is more efficient (hence better) than any other unbiased, linear estimator of β. $\square$

Note that $\hat{\beta}$ and $\tilde{\beta}$ are both $(k+1) \times 1$ vectors. As Wooldridge notes at the end of the proof, for any $(k+1) \times 1$ vector c, we can calculate the scalar $c'\beta$. Think of c as the row vector of the ith observation from X. Then $c'\beta = c'_o\beta_0 + c_1\beta_1 + ... + c_k\beta_k = y_i$. Both $c'\hat{\beta}$ and $c'\tilde{\beta}$ are both unbiased estimators of $c'\beta$. Note as an extension of the proof above that

$$\text{Var}(c'\tilde{\beta}|X) - \text{Var}(c'\hat{\beta}|X) = c'[\text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}|X)]c.$$

We know that $\text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}|X)$ is PSD, and hence by definition that:

$$c'[\text{Var}(\widetilde{\beta}|X) - \text{Var}(\hat{\beta}|X)]c \geq 0,$$

and hence, for any observation in X (call it x_i), and more broadly any linear combination of $\hat{\beta}$, if the GM assumptions hold the estimate $\hat{y_i} = x_i\hat{\beta}$ has the smallest variance of any possible linear, unbiased estimator.

9.3.2 Optimisation problems

Optimisation problems, in essence, are about tweaking some parameter(s) until an objective function is as good as it can be. The objective function summarises some aspect of the model given a potential solution. For example, in OLS, our objection function is defined as $\sum_i (y_i - \hat{y}_i)^2$ – the sum of squared errors. Typically, "as good as it can be" stands for "is minimised" or "is maximised." For example with OLS we seek to minimise the sum of the squared error terms. In a slight extension of this idea, many machine learning models aim to minimise the prediction error on a "hold-out" sample of observations i.e. observations not used to select the model parameters. The objective loss function may be the sum of squares, or it could be the mean squared error, or some more convoluted criteria.

By "tweaking" we mean that the parameter values of the model are adjusted in the hope of generating an even smaller (bigger) value from our objective function. For example, in least absolute shrinkage and selection (LASSO) regression, the goal is to minimise both the squared prediction error (as in OLS) as well as the total size of the coefficient vector. More formally, we can write this objective function as:

$$(y - X\beta)^2 + \lambda||\beta||_1,$$

where λ is some scalar, and $||\beta||_1$ is the sum of the absolute size of the coefficients i.e. $\sum_j |\beta_j|$.

There are two ways to potentially alter the value of the LASSO loss function: we can change the values within the vector β or adjust the value of λ. In fact, iterating through values of λ, we can solve the squared error part of the loss function, and then choose from our many different values of λ which results in the smallest (read: minimised) objective function.

With infinitely many values of λ, we can perfectly identify the optimal model. But we are often constrained into considering only a subset of possible cases. If we are too coarse in terms of which λ values to consider, we may miss out on substantial optimisation.

This problem is not just present in LASSO regression. Any non-parametric model (particularly those common in machine learning) is going to face similar optimisation problems. Fortunately, there are clever ways to reduce the computational intensity of these optimisation problems. Rather than iterating through a range of values (an "exhaustive grid-search") we can instead use our current loss value to adjust our next choice of value for λ (or whatever other parameter we are optimisimng over). This sequential method helps us narrow in on the optimal parameter values without having to necessarily consider may parameter combinations far from the minima.

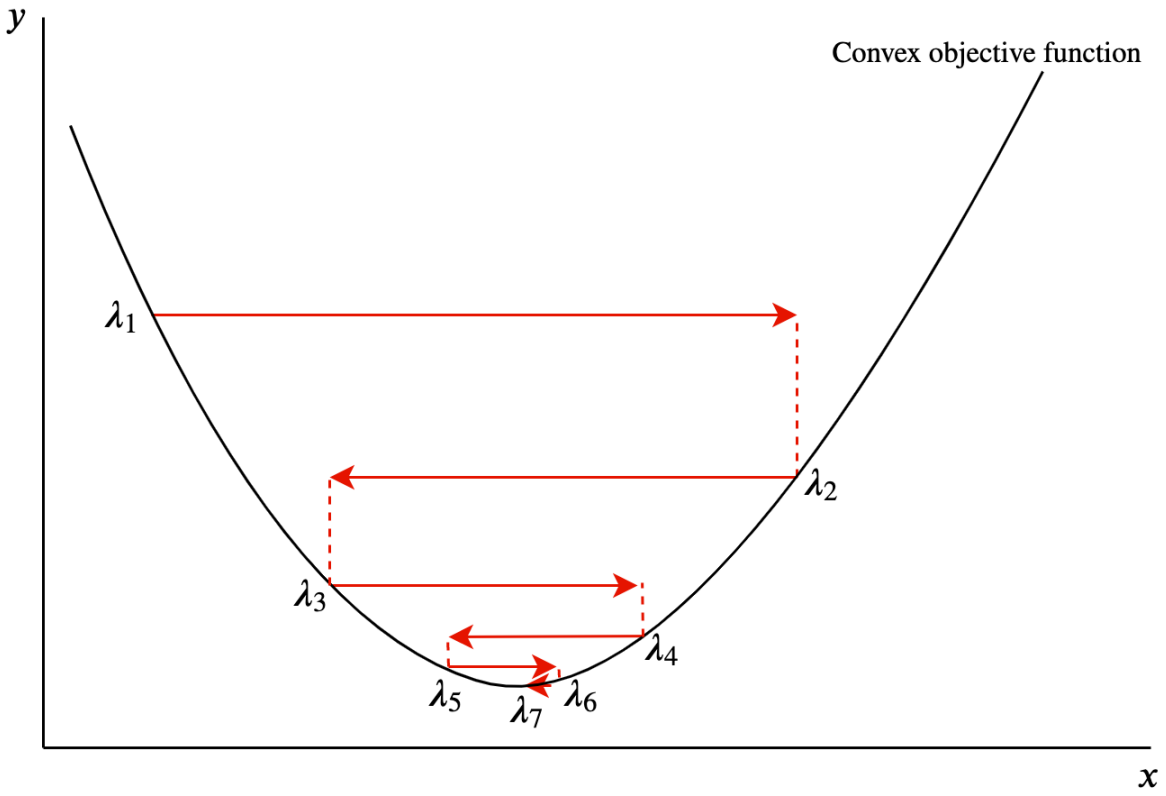
Of course, the natural question is how do we know how to adjust the scalar λ, given our existing value? Should it be increased or decreased? One very useful algorithm is gradient descent (GD), which I will focus on in the remainder of this section. Briefly, the basics of GD are:

1. Take a (random) starting solution to your model
2. Calculate the gradient (i.e. the k-length vector of derivatives) of the loss at that point
3. If the gradient is positive (negative), decrease (increase) your parameter by the gradient value.
4. Repeat 1-3 until you converge on a stable solution.

Consider a quadratic curve in two-dimensions, as in Figure 9.1. If the gradient at a given point is positive, then we know we are on the righthand slope. To move closer to the minimum point of the curve we want to go left, so we move in the negative direction. If the gradient is negative, we are on the lefthand slope and want to move in the positive direction. After every shift I can recalculate the gradient and keep adjusting.

Crucially, these movements are dictated by the absolute size of the gradient. Hence, as I approach the minimum point of the curve, the gradient and therefore the movements will be smaller. In 9.1, we see that each iteration involves not only a move towards the global minima, but also that the movements get smaller with each iteration.

Figure 9.1: Gradient descent procedure in two dimensions.



PD matrices are like the parabola above. Geometrically, they are bowl-shaped and are guaranteed to have a global minimum.[2] Consider rolling a ball on the inside surface of this bowl. It would run up and down the edges (losing height each time) before eventually resting on the bottom of the bowl, i.e. converging on the global minimum. Our algorithm is therefore bound to find the global minimum, and this is obviously a very useful property from an optimisation perspective.

If a matrix is PSD, on the other hand, we are not guaranteed to converge on a global minima. PSD matrices have "saddle points" where the slope is zero in all directions, but are neither (local) minima or maxima in all dimensions. Geometrically, for example, PSD matrices can look like hyperbolic parabaloids (shaped like a Pringles crisp). While there is a point on the surface that is flat in all dimensions, it may be a minima in one dimension, but a maxima in another.

PSD matrices prove more difficult to optimise because we are not guaranteed to converge on that point. At a point just away from the saddle point, we may actually want to move in opposite direction to the gradient dependent on the axis. In other words, the valence of the individual elements of the gradient vector point in different directions. Again, imagine dropping a ball onto the surface of a hyperbolic parabaloid. The ball is likely to pass the saddle point then run off one of the sides: gravity is pulling it down in to a minima in one dimension, but away from a maxima in another. PSD matrices therefore prove trickier to optimise, and can even mean we do not converge on a miniimum loss value. Therefore our stable of basic algorithms like GD like gradient descent are less likely to be effective optimisers.

[2]See these UPenn lecture notes for more details.

9.3.3 Recap

In this final section, we have covered two applications of positive (semi-) definiteness: the proof of OLS as BLUE, and the ease of optimisation when a matrix is PD. There is clearly far more that can be discussed with respect to P(S)D matrices, and this chapter links or cites various resources that can be used to go further.

Bibliography

Angrist, J. D. and Pischke, J.-S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton.

Aronow, P. and Miller, B. (2019). Foundations of Agnostic Statistics. Cambridge University Press.

Aronow, P. M. and Samii, C. (2016). Does regression produce representative estimates of causal effects? American Journal of Political Science, 60(1):250–267.

Bowler, S., Donovan, T., and Karp, J. A. (2006). Why politicians like electoral institutions: Self-interest, values, or ideology? The Journal of Politics, 68(2):434–446.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(1):39–67.

Davidson, R., MacKinnon, J. G., et al. (2004). Econometric theory and methods, volume 5. Oxford University Press New York.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. Econometrica: Journal of the Econometric Society, pages 387–401.

Horn, R. A. H. and Johnson, C. R. (2013). Matrix Analysis. NY, USA, 2nd edition edition.

King, G., Tomz, M., and Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. American Journal of Political Science, 44:341–355.

Lemons, D., Langevin, P., and Gythiel, A. (2002). An introduction to stochastic processes in physics. Johns hopkins paperback. Johns Hopkins University Press. Citation Key: lemons2002introduction tex.lccn: 2001046459.

Liu, X. (2012). Appendix A: The Delta Method, pages 405–406. John Wiley & Sons, Ltd.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association, 58(304):993–1010.

van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference. Springer Texts in Statistics. Springer New York, New York, NY.

Wooldridge, J. M. (2013). Introductory econometrics : a modern approach. Mason, OH, 5th edition edition.