Econometrics Notes

Jason Hall

Fall 2022

1 Linear Algebra

One of the core mathematical objects we find ourselves concerned with are **matrices**. A matrix is a rectangular array of objects.

Take for instance

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

In general, we notate the component on row i in column j by a_{ij}, and we say that a matrix $\mathbf{A}$ is equal to a matrix $\mathbf{B}$ if and only if $a_{ij} = b_{ij}\ \forall i\ \forall j$. In addition, we define the transpose of a matrix $\mathbf{A}$ as a matrix $\mathbf{A}'$ that has the property that $a'_{ij} = a_{ji}\ \forall i\ \forall j$. For instance, using the matrix above, we have

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

We say a matrix $\mathbf{A}$ is symmetric if and only if $\mathbf{A}' = \mathbf{A}$.

Now let $\mathbf{A}$ and $\mathbf{B}$ be two $m \times n$ matrices.[1] Let $\mathbf{C} = \mathbf{A} + \mathbf{B}$. Addition is defined element-wise by

$$c_{ij} = a_{ij} + b_{ij}$$

Now let $\mathbf{A}$ be a $m \times n$ matrix, and let $\mathbf{B}$ be a $n \times k$ matrix.

Then ij component of $\mathbf{C} = \mathbf{AB}$ is given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

One can show (homework problems) that matrix multiplication is well behaved in the sense that it obeys both the associative and distributive properties.

In general, however, matrix multiplication is not commutative.

In general, I think the most productive way to review definitions in Linear Algebra is to recall the (exhaustingly long) list of equivalences in the invertible matrix theorem. Recall that we refer to the group of $n \times n$ invertible matrices with real entries as $GL_n(\mathbb{R})$.

[1] I.e. one with m rows and n columns.

Invertible Matrix Theorem: *Let* $\mathbf{A}$ *be a* $n \times n$ *matrix with entries* $a_{ij} \in \mathbb{R}$*, with associated linear transformation* $T : x \to \mathbf{A}x$*. Then the following are equivalent:*

1. $\mathbf{A}$ *is invertible.*
2. $\mathbf{A}$ *has* n *pivots.*
3. $Nul(\mathbf{A}) = \{0\}$.
4. *The columns of* $\mathbf{A}$ *are linearly independent.*
5. *The columns of* $\mathbf{A}$ *span* $\mathbb{R}^n$.
6. $\mathbf{A}x = b$ *has a unique solution* $\forall b \in \mathbb{R}^n$.
7. T *is invertible.*
8. T *is one-to-one.*
9. T *is onto.*
10. $\mathbf{A}$ *is row-equivalent to the identity matrix* $\mathbf{I}_n$.
11. $nullity(\mathbf{A}) = 0$
12. *The columns of* $\mathbf{A}$ *form a basis for* $\mathbb{R}^n$.
13. $Col(\mathbf{A}) = \mathbb{R}^n$.
14. $dimCol(\mathbf{A}) = n$.
15. $rank(\mathbf{A}) = n$.
16. *The eigenvalues of* $\mathbf{A}$ *are all distinct from* 0.
17. $det(\mathbf{A}) \neq 0$

Proof. Proof omitted. Consult a linear algebra textbook, virtually all have the IMT in one form or another. ■

Let us now proceed through the long list of equivalences and ensure that we understand what each of them mean.

1. Beginning with the first condition, we say that a matrix $\mathbf{A}$ is invertible if there exists a matrix $\mathbf{B}$, such that $\mathbf{AB} = \mathbf{I}_n = \mathbf{BA}$. One can also show (exercise!) that if $\mathbf{B}$ is a left inverse, it is also a right inverse, and vice versa. Further inverses are unique.
2. Recall that a pivot is the first non-zero entry in each row in a matrix in reduced row echelon form.
3. Recall that $Nul(\mathbf{A})$ is the set of $x \in \mathbb{R}^n$ such that $\mathbf{A}x = 0$.
4. Let $\{\mathbf{x}_i\}$ denote the column vectors of $\mathbf{A}$. We say that the columns are linearly independent

if the only solution to the equation $\sum_i^n \lambda_i \mathbf{x}_i = 0$ is the trivial solution where $\lambda_i = 0 \ \forall i \in \{1, 2, \ldots, n\}$.

5. We say that the columns of $\mathbf{A}$ span $\mathbb{R}^n$ if $\forall \mu \in \mathbb{R}^n$, there exists $\{\lambda_i\}_{i=1}^n$ such that $\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mu$.
6. Obvious.
7. We say that a linear transformation is invertible if and only if there exists $S : x \to \mathbf{B}_x$ such that $S(T(x)) = x$. This is basically just the normal restriction that a function is invertible if and only if it is both injective and surjective (i.e. bijective).
8. Same as functions, it means that T is injective (i.e. $f(x_1) = f(x_2) \implies x_1 = x_2$).
9. Same as functions, it means that T is surjective (i.e. $\forall y \ \exists x | f(x) = y$).
10. Consult a textbook for examples of row-reduction.
11. The nullity of a matrix $\mathbf{A}$ is the dimension of its nullspace (also call its kernel).
12. See 5.
13. The columnspace of a matrix $\mathbf{A}$ is the set of vectors attainable by linear combinations of its column vectors.
14. The columns of A span n dimensions.
15. The column (resp. row) rank of a matrix is the dimension of the vector space that is generated by the columns (resp. rows) of $\mathbf{A}$. One can show that column rank and row rank always coincide, hence we simply refer to this as the rank of a matrix $\mathbf{A}$.
16. Eigenvalues are solutions to the characteristic equation $(\mathbf{A} - \lambda \mathbf{I}_n)x = 0$
17. The determinant of a matrix is the product of its eigenvalues.

1.1 Eigenvalues, Eigenvectors, and Eigenspaces

Now we briefly restrict consideration to square matrices (i.e. those with the same number of rows and columns). Let $\mathbf{A}$ be a square matrix. If there exists $x \in \mathbb{R}^n$ and a $\lambda \in \mathbb{R}$ such that $\mathbf{A}x = \lambda x$, then we call λ an **eigenvalue** of $\mathbf{A}$ and x its associated **eigenvector**. An eigenspace of a given eigenvalue is the span of said eigenvalues eigenvectors.

This is the kind of technical definition that doesn't convey much intuition as to what is going on, so instead let us calculate the eigenvalues of a given matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 1 \\ 3 & 3 \end{bmatrix}$$

We want to solve

$$(5 - \lambda)(3 - \lambda) - 3 = 0$$

$$15 - 8\lambda + \lambda^2 - 3 = 0$$

$$\lambda^2 - 8\lambda + 12 = 0$$

$$\lambda = 6 \vee \lambda = 2$$

These are our eigenvalues, but what are our corresponding eigenvectors.
Eigenvectors, by definition, are vectors v_i that solve

$$(\mathbf{A} - \lambda_i \mathbf{I}_n) v_i = \mathbf{0}$$

Beginning with the case where $\lambda = 6$, this resolves to finding solutions to

$$\begin{bmatrix} -1 & 1 \\ 3 & -3 \end{bmatrix} v_i = \mathbf{0}$$

Which one should be able to see implies that $v_1 = [1, 1]'$.
For the case where $\lambda = 2$, we are considering

$$\begin{bmatrix} 3 & 1 \\ 3 & 1 \end{bmatrix} v_i = \mathbf{0}$$

Which implies $v_2 = [1, -3]'$
What happens when we apply $\mathbf{A}$ to an eigenvector?

$$\mathbf{A} = \begin{bmatrix} 5 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}$$

So $\mathbf{A}$ preserves the orientation of the eigenvector v_1 but scales it by a factor of 6, which is the associated eigenvalue.

To put it in other words, eigenvectors are simply those vectors whose directions[2] are preserved under the linear transformation associated with $\mathbf{A}$.

Note that an eigenvalue may have more than one eigenvector, but an eigenvector cannot have more than one eigenvalue. To see the latter, suppose for the sake of contradiction that an eigenvector x is associated with more than one eigenvalue. It follows then $\lambda_1 x = \lambda_2 x$, with $\lambda_1 \neq \lambda_2$. From the prior observation, we have that $(\lambda_1 - \lambda_2)x = 0$. Since $x \neq 0$ by the fact that it is an eigenvector, it follows that $\lambda_1 = \lambda_2$, giving us the desired contradiction.

Now we prove a useful property of eigenvectors.

[2]With the caveat that eigenvalues may be negative in which case, the direction of the vectors in the eigenspace is flipped 180 degrees.

Claim: *Let $\mathbf{A} : V \to V$, and let $[v_1\ v_2\ \dots\ v_n]$ be n eigenvectors with associated eigenvalues $\{\lambda_i\}_{i=1}^n$. Suppose further that $0 \neq \lambda_i \neq \lambda_j\ \forall i\ \forall j$. Then the collection of eigenvectors is linearly independent.*

Proof. Let us denote the transformation associated with $\mathbf{A}$ by T.
Let $S = \{i \in \{1, 2, \dots, n\} | \forall i' \in \{1, \dots, i\} \sum_{l=1}^{i'} a_l v_l = 0 \iff a_l = 0 \forall l\}$.[a]
Clearly $1 \in S$, since $v_1 \neq \mathbf{0}$ as it is an eigenvector.
Suppose that $k \in S$. We wish to show that

$$\sum_{i=1}^{k+1} a_i v_i = 0$$

has only the trivial solution.
To that end, observe

$$0 = T(0) = T(\sum_{i=1}^{k+1} a_i v_i) = \sum_{i=1}^{k+1} T(a_i v_i) = \sum_{i=1}^{k+1} \lambda_i a_i v_i$$

.
Further

$$0 = 0 * \lambda_{k+1} = \lambda_{k+1} \sum_{i=1}^{k+1} a_i v_i = \sum_{i=1}^{k+1} \lambda_{k+1} a_i v_i$$

Equating both expressions, we have

$$\sum_{i=1}^{k+1} \lambda_{k+1} a_i v_i = \sum_{i=1}^{k+1} \lambda_i a_i v_i \iff \sum_{i=1}^{k+1} a_i (\lambda_i - \lambda_{k+1}) v_i = 0$$

Since by assumption $\lambda_i \neq \lambda_{k+1}$, by our inductive hypothesis we know that it must be that $a_i = 0\ \forall i \in \{1, 2, \dots, k\}$.
Hence we have

$$0 = \sum_{i=1}^{k+1} a_i v_i = a_{k+1} v_{k+1}$$

Since v_{k+1} is an eigenvector, we have that it is non-zero, and thus it must be that $a_{k+1} = 0$
Hence $k + 1 \in S$.
By the principle of mathematical induction, we have that $S = \{1, 2, 3, \dots, n\}$

■

[a] Since this is kind of hard to write, this is the set of linearly independent vectors in the collection of eigenvectors. If $1 \in S$, then v_1 is linearly independent. If $2 \in S$, v_1 and v_2 are linearly independent, and so on.

Another very useful property we have is given by the following claim.

Claim: *Let* $\mathbf{A}$ *be a real symmetric matrix. Then the eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof. Let λ_i and λ_j be eigenvalues associated with $\mathbf{A}$, with $\lambda_i \neq \lambda_j$. Denote their associated eigenvectors v_i and v_j.
By the definition of eigenvectors, we have that
$\mathbf{A}v_i = \lambda_i v_i$ and $\mathbf{A}v_j = \lambda_j v_j$

$$\lambda_i v_i' v_j = (\lambda_i v_i)' v_j = (\mathbf{A}v_i)' v_j = v_i' \mathbf{A} v_j = v_i' (\lambda_j v_j) = \lambda_j v_i' v_j$$

Since $\lambda_i \neq \lambda_j$, we have that $v_i' v_j = \mathbf{0}$ ■

Now let us define positive semi-definite matrices. Specifically, we say that a $n \times n$ matrix $\mathbf{A}$ is **positive semi-definite** if

$$x' \mathbf{A} x \geq 0 \; \forall x \in \mathbb{R}^n$$

Claim: *Suppose* $\mathbf{A}$ *is a* $n \times n$ *positive semi-definite matrix. Then all of its eigenvalues are non-negative.*

Proof. Suppose that λ is an eigenvalue of $\mathbf{A}$. It follows then that for the eigenvector v associated with λ,

$$\mathbf{A}v = \lambda v$$

By the fact that $\mathbf{A}$ is positive semi-definite, we have that

$$0 \leq v' \mathbf{A} v = v' \lambda v = \lambda v' v$$

Hence it must be that $\lambda \geq 0$.

■

Claim: *If* $\mathbf{A}$ *is a symmetric matrix, then* $\mathbf{A}$ *has* n *distinct eigenvectors that form an orthonormal basis for* $\mathbb{R}^n$.

Proof. Omitted. See e.g. here., which is the source for many of these propositions. ■

(Simple) Spectral Decomposition Theorem: *Suppose that* $\mathbf{A}$ *is a symmetric* $n \times n$ *matrix. Then* $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$*, with* $\mathbf{V}$ *an orthonormal basis of eigenvectors of* $\mathbf{A}$*, and* Λ *a matrix composed on the diagonal of the corresponding eigenvalues.*

Proof. Let $V = [v_1 \ v_2 \ \dots \ v_n]$ be the n orthonormal vectors of $\mathbf{A}$, whose existence is guaranteed by the prior claim. Let Λ be the diagonal matrix of associated eigenalues.
Observing that V is an orthonormal collection, we have that $v_i'v_i = 1 \ \forall i$ by the normality of v_i, and further that $v_i'v_j = 0 \ \forall i \neq j$, by the orthogonality of the collection, we have that $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_n$. Equivalently, $\mathbf{V}' = \mathbf{V}^{-1}$.
Now we note that

$$\lambda_i \mathbf{V}' v_i = \lambda_i e_i \ \forall i \in \{1, 2, \dots, n\}$$

Hence, we have that

$$\mathbf{V}'\mathbf{A}\mathbf{V} = \mathbf{V}'[\lambda_1 v_1 \ \lambda_2 v_2 \ \dots \ \lambda_n v_n] = [\lambda_1 e_1 \ \lambda_2 e_2 \ \dots \lambda_n v_n] = \Lambda$$

Since $\mathbf{V}$ is invertible and its inverse is its transpose, we have that

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$$

■

This is a special case of something called an eigendecomposition, and it is very useful in this course, primarily because it ensures the existence of the next matrix factorization we are going to discuss.

Square Root Factorization: *Suppose that* $\mathbf{A}$ *is a symmetric, semi-positive definite matrix. Then* $\mathbf{A} = \mathbf{P}'\mathbf{P}$ *for some matrix* $\mathbf{P}$*.*

Proof. From the prior result, we have that

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}'$$

Observing that since Λ is positive semi-definite, each of its diagonal entries must be greater than or equal to zero. Define the matrix formed by taking the square root of each of these entries as $\Lambda^{\frac{1}{2}}$. Then

$$\mathbf{A} = \mathbf{V}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{V}' = \mathbf{V}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}{}'\mathbf{V}' = \mathbf{V}\Lambda^{\frac{1}{2}}(\mathbf{V}\Lambda^{\frac{1}{2}})'$$

Defining $\mathbf{P} = (\mathbf{V}\Lambda^{\frac{1}{2}})'$, the result follows. ■

If you are wondering why we care so much about properties of positive semi-definite matrices, it's because of variance-covariance matrices.

These are two examples of what is called a **Matrix factorization**. There are many more, including e.g. the singular value decomposition, which is a generalization of the eigendecomposition (itself a more general form of the spectral decomposition given above). Another commonly used one is the QR decomposition, which is perhaps most famously used in the estimation of ordinary least squares in virtually all statistical software.

1.2 Other Useful Things

Claim: $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$

Proof. Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times m$ and $m \times n$ matrices respectively.
Observe that

$$Tr(\mathbf{AB}) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = Tr(\mathbf{BA})$$

■

Now, let us observe that by setting $\mathbf{D} \equiv \mathbf{AB}$, we have that $Tr(\mathbf{DC}) = Tr(\mathbf{CD})$. Substituting for $\mathbf{D}$, we further recover that

$$Tr(\mathbf{ABC}) = Tr(\mathbf{CAB})$$

A similar technique will recover that

$$Tr(\mathbf{CAB}) = Tr(\mathbf{BCA})$$

Hence we have

$$Tr(\mathbf{ABC}) = Tr(\mathbf{CAB}) = Tr(\mathbf{BCA})$$

This is what is referred to as the cyclic property of the trace.
Additionally,

$$Tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

$$det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$$

We say a square matrix $\mathbf{B}$ is idempotent if $\mathbf{BB} = \mathbf{B}$.

Claim: *Suppose that* $\mathbf{A}$ *is an idempotent matrix. Then* $\mathbf{A}$ *is invertible if and only if* $\mathbf{A} = \mathbf{I}$.

Proof. Suppose that $\mathbf{A}$ is idemptotent and invertible.
We then have that

$$\mathbf{AA} = \mathbf{A}$$

Applying $\mathbf{A}^{-1}$ to each side, we then have that

$$\mathbf{A}^{-1}\mathbf{AA} = \mathbf{A}^{-1}\mathbf{A}$$

$$\mathbf{A} = \mathbf{I}$$

■

Another important property of positive definite matrices is that

Claim: *Suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric, positive definite, and non-singular matrices. Then* $\mathbf{A} - \mathbf{B}$ *is positive semi definite implies that* $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ *is positive semi-definite.*

Proof. Omitted. ■

Claim: *An idempotent matrix has only eigenvalues* 0 *or* 1*.*

Proof. Let $\mathbf{A}$ be an idempotent matrix. Then $\mathbf{A} = \mathbf{AA}$. Consider $\mathbf{A}v = \lambda v$. Observe that

$$\lambda v = \mathbf{A}v = \mathbf{AA}v = \mathbf{A}\lambda v = \lambda \mathbf{A}v = \lambda^2 v$$

Hence we have that,

$$\mathbf{0} = \lambda^2 v - \lambda v = \lambda(\lambda - 1)v$$

Therefore $\lambda = 0$ or $\lambda = 1$. ■

1.3 Matrix Differentiation

Let us begin by considering the following:

$$\mathbf{y} = f(\mathbf{x})$$

Where $\mathbf{y}$ is a $m \times 1$ vector, and likewise $\mathbf{x}$ is a $n \times 1$ vector.

Then the Jacobian of f is given by

$$J_f(\mathbf{x}) = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} \\ \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} & \cdots & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix}$$

Suppose now that we are interested in a mapping f of the form $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{y}$ is a $m \times 1$ matrix, $\mathbf{A}$ is a $m \times n$ matrix of constants, and $\mathbf{x}$ is a $n \times 1$ matrix.

Then one should immediately see that

$$y_i = \sum_{j=1}^{n} a_{ij} x_j$$

Where I am treating both $\mathbf{y}$ and $\mathbf{x}$ as vectors (and thus suppressing column indices). As a consequence,

$$\frac{\partial y_i}{\partial x_j} = a_{ij}$$

Hence we have that

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$$

Suppose also that we are interested in a mapping f of the form $\alpha = \mathbf{x}' A \mathbf{x}$, where $\alpha \in \mathbb{R}$, $\mathbf{x}$ is a $n \times 1$ matrix and $\mathbf{A}$ is a $n \times n$ matrix.

One can see that

$$\alpha = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

Differentiating with respect to x_k, we get

$$\frac{\partial \alpha}{\partial x} = \sum_{i=1}^{n} a_{ik} x_i + \sum_{j=1}^{n} a_{jk} x_j$$

This implies that

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$$

In the case where $\mathbf{A}$ is symmetric (which we will encounter many times in this course), the derivative becomes

$$\frac{\partial \alpha}{\partial \mathbf{x}} = 2\mathbf{x}'A$$

For other formulas, refer to the matrix cookbook, which you can find here.

1.4 Vector Spaces and Inner Product Spaces

We know what a vector is: it's a mathematical object that has some notion of both magnitude and direction. We can treat points in $\mathbb{R}^n$ as vectors by considering them to "start" at the origin. But suppose we, for instance, want to talk about collections of vectors in the same way we talk about objects in abstract algebra, i.e. as groups or fields. The most immediate way to do this is to talk about **vector spaces**.

Formally, a vector space over a field[3] F is a set V that comes equipped with two binary operations:

1. (Vector Addition): an operation $+$ that takes $\mathbf{v} \in V$ and $\mathbf{w} \in V$ to a third vector in $\mathbf{v} + \mathbf{w} \in V$
2. (Scalar Multiplication): an operation that takes $\alpha \in F$ and $\mathbf{v} \in V$ and returns a new vector $\alpha\mathbf{v} \in V$.

To then be a vector space, we require that $\forall\, \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\alpha, \beta \in F$ the following are true:

[3] A field is a set F that has two binary operations that we can call addition and multiplication. The operations are required to satisfy the field axioms, which are:

1. Associativity of addition and multiplication
2. Commutativity of addition and multiplication
3. Identities for both addition and multiplication
4. Additive inverses
5. Multiplicative inverses (apart from the case when considering 0)
6. Distributivity of multiplication over addition

1. (Associativity of vector addition): $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
2. (Commutativity of vector addition): $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
3. (Additive Identity): $\exists \mathbf{0} \in V$ such that $\forall\ \mathbf{v} \in V$ we have $\mathbf{v} + \mathbf{0} = \mathbf{v}$
4. (Additive Inverses): $\forall\ \mathbf{v} \in V$, there exists $-\mathbf{v}$ such that $\mathbf{v} + -\mathbf{v} = \mathbf{0}$
5. (Compatibility): $\forall\ \alpha, \beta \in F$, we have that $\alpha(\beta\mathbf{v}) = (\alpha\beta)\mathbf{V})$
6. (Multiplicative Identity): Denote by e the multiplicative identity in F. Then $e\mathbf{V} = \mathbf{V}$
7. (Distributivity of scalar multiplication over vector addition): $\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}$
8. (Distributivity of scalar multiplication over field addition): $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$

It is a homework problem later in the course to check that the vector space where $V = \mathbb{R}^n$ (under pointwise addition) and $F = \mathbb{R}$ (under the normal addition and multiplication) constitutes a normed vector space.

If we want to talk about a vector space, it is convenient to represent a vector space through a **basis**. A basis for a vector space V over a field F is a set of vectors $\{\mathbf{v}_i\}_{i=1}^n$ that satisfies two properties:

1. (Linear Independence): Let $\{\mathbf{v}_i\}_{i=1}^k$ with $k \leq n$ be a a subset of the basis. If $\sum_{i=1}^k a_i\mathbf{v}_i = \mathbf{0}$, then $a_i = 0 \forall i \in \{1, 2, \ldots, k\}$
2. (Spanning): Let $\mathbf{u}$ be a vector in V. Then $\mathbf{u}$ can be written as $\sum_{i=1}^n a_i\mathbf{v}_i$ for some collection of $a_i \in F$.

There is a very important result in linear algebra that says that any basis for $\mathbf{V}$ must have the same number of elements. This result is called the Steinitz exchange lemma.

Steinitz Exchange Lemma: *Let U and W be finite subsets of some vector space V. If U is a set of linearly independent vectors, and W spans V, then the following are true:*

1. $|U| \leq |W|$
2. *There is a set $W' \subseteq W$ such that $|W'| = |W| - |U|$, such that $U \cup W'$ spans V.*

Proof. Let us write $W = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \cdots, \mathbf{w}_n\}$, and $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$. Let $S = \{k \in \{0, 1, \ldots, m\} |$ Property 2 holds.$\}$. Trivially, $0 \in S$, since that implies $U = \emptyset$ and W spans V by assumption. Now suppose that $k \in S$. It follows that $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_n\}$ spans V. Let $\mathbf{u}_{k+1}$ be given, and assume that it is not currently in the span of U. Since $U \cup W$ spans V, it follows that there must exist weights a_i such that $\mathbf{u}_{k+1} = \sum_{i=1}^{k} a_i \mathbf{v}_i + \sum_{i=k+1}^{n} a_i \mathbf{w}_i$, with at least one a_i on $\mathbf{w}_i$ non-zero[a] Without loss of generality, take a_{k+1} to be non-zero. It follows then that

$$w_{k+1} = \frac{1}{a_{k+1}} \left(u_{k+1} - \sum_{i=1}^{k} a_i \mathbf{u}_i - \sum_{i=k+2}^{n} a_i \mathbf{w}_i \right)$$

Hence $\mathbf{w}_{k+1}$ is contained in the span of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k+1}, \mathbf{w}_{k+2}, \ldots, \mathbf{w}_n\}$
And the span of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_n\}$ must be a subset of the span of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k+1}, \mathbf{w}_{k+2}, \ldots, \mathbf{w}_n\}$.
Since the former spans V, the result is shown. Hence if $k \in S$, it must be that $k + 1 \in S$, so by principle of mathematical induction $S = \{1, 2, \ldots, m\}$.
That $|U| \leq |W|$ is a function of the fact that W spans V. Suppose that $|U| > |W|$. Then since $|W|$ spans V, it must be that every vector in U can be written as a linear combination of vectors in W. Since $|U| > |W|$, it follows then that one must be able to write at least $n+1$ linearly independent vectors from n linearly independent ones, which is a contradiction. ■

[a] Linear independence of $\mathbf{u}_{k+1}$ from the collection $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ implies that the coefficients on the first k a_i must be zero. If the weights on all of $\mathbf{w}_i$ were zero, then the weights are everywhere zero, which implies that u_{k+1} is linearly independent of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_n\}$, contradicting the spanning assumption.

This result enables us to prove a fundamental result in linear algebra, namely rank-nullity theorem.

Rank-Nullity Theorem: *Let* $\mathbf{T}$ *be a linear map from some finite-dimensional vector space* V *to a vector space* W *over a field* K *(i.e.* $\mathbf{T} : V \to W$*). Then* $rank(\mathbf{T}) + nullity(\mathbf{T}) = dim(V)$*.*

Proof. Recall that $nullity(\mathbf{T}) = dim(Ker(\mathbf{T}))$. Let us begin by showing that the kernel of $\mathbf{T}$ is a linear subspace. Specifically, let $u, v \in Ker(\mathbf{T})$. It follows that then that $\mathbf{T}(u) = \mathbf{0}$ and likewise for v. Let $\alpha \in K$ be given, and observe that then $\mathbf{T}(\alpha\mathbf{u}) = \alpha * \mathbf{T}(\mathbf{u}) = \mathbf{0}$, since $\mathbf{T}$ is a linear map. Likewise,

$$0 = 0 + 0 = \mathbf{T}(\mathbf{u}) + \mathbf{T}(\mathbf{v}) = \mathbf{T}(\mathbf{u} + \mathbf{v})$$

Where again I am making use of the fact that $\mathbf{T}$ is a linear map. Hence $Ker(\mathbf{T})$ is closed under vector addition and scalar multiplication, and thus is a linear subspace. Furthermore, it must have a basis which I will notate as $\{\mathbf{r}_i\}_{i=1}^k$, where k is the dimension of $Ker(\mathbf{T})$
Now observe that since V is finite dimensional, it must have a basis, which we will write as $\{\mathbf{v}_i\}_{i=1}^n$, where n is the dimension of V.[a] By the *Steinitz Exchange Lemma*, we can enlarge $\{r_i\}_{i=1}^k$ with $n-k$ vectors from $\{\mathbf{v}_i\}_{i=1}^n$. Call this new basis $\mathbf{B}$. Observe that $span(\mathbf{T}(\mathbf{B})) = Image(\mathbf{T})$. Further $span(\mathbf{T}(\{\mathbf{r}_i\}_{i=1}^k, \{\mathbf{v}_i\}_{i=k+1}^n)) = span(\{\mathbf{T}(\mathbf{r}_i)\}_{i=1}^k, \mathbf{T}(\{\mathbf{v}_i\}_{i=k+1}^n)) = span(\mathbf{T}(\{\mathbf{v}_{i=k+1}^n\}))$. Hence, $\mathbf{T}(\{\mathbf{v}_i\}_{i=k+1}^n)$ spans $image(\mathbf{T})$.
Now we wish to show that the vectors $\{\mathbf{T}(\mathbf{v}_i)\}_{i=k+1}^n$ are linearly independent. Suppose that

$$\sum_{i=k+1}^{n} a_i \mathbf{T}(\mathbf{v_i}) = \mathbf{0}$$

.
In that case,

$$\sum_{i=k+1}^{n} a_i \mathbf{v}_i \in Ker(\mathbf{T})$$

Suppose that one of the $a_i \neq 0$. It follows then that at least one of those vectors lies in the kernel of $\mathbf{T}$. But observe then $\mathbf{B}$ is no longer a basis as it has at most $n-1$ linearly independent vectors and is n dimensional. Contradiction, so all the $a_i = 0$. Hence, the dimension of the image of $\mathbf{T}$ is exactly $n-k$.
Together, then, we have that

$$rank(\mathbf{T}) + nullity(\mathbf{T}) = dim(Image(\mathbf{T})) + dim(Ker(\mathbf{T})) = (n-k) + k = n = dim(V)$$

Which was what was wanted. ■

[a] Note that if V is infinite dimensional, the result that V has a basis follows from Zorn's Lemma, which is equivalent to the Axiom of Choice.

I now define a **norm**.
Let V be a vector space over $\mathbb{R}$[4]. Then a norm is a function $p : V \to \mathbb{R}$ that satisfies the

[4] The assumption that the field the vector space is defined over is $\mathbb{R}$ is more strict than is necessary, but I will maintain it here as the generalization is precisely what you would expect, and it simplifies the mathematical intuition.

following properties:

1. (Triangle Inequality):$\forall\ \mathbf{u}, \mathbf{v} \in V$, we have that $p(\mathbf{u} + \mathbf{v}) \leq p(\mathbf{u}) + p(\mathbf{v})$
2. (Absolute Homogeneity):$\forall\ \mathbf{v} \in V$ and $\forall\ \alpha \in \mathbb{R}$, we have that $p(\alpha\mathbf{v}) = |\alpha|p(\mathbf{v})$
3. (Positive Definite): $\forall\ \mathbf{v} \in V$, we have that $p(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

Claim: *$p(\cdot)$ is a non-negative function.*

Proof. Let $\mathbf{v}$ be an element of V. Now observe that $p(-\mathbf{v}) = |-1|p(\mathbf{v}) = p(\mathbf{v})$. Now observe that

$$\begin{aligned} p(\mathbf{0}) &= p(\mathbf{v} + -\mathbf{v}) \\ &\leq p(\mathbf{v}) + p(-\mathbf{v}) \\ &= p(\mathbf{v}) + p(\mathbf{v}) \\ &= 2p(\mathbf{v}) \end{aligned}$$

Hence we have that

$$0 \leq p(\mathbf{0}) \leq p(\mathbf{v})$$

And the result is shown. ■

Note that in the above proof we only made use of properties 1 and 2. These two properties alone define what is called a **semi-norm**. It is the presence of property 3, the uniqueness of the 0, the defines a norm. This is a crucial distinction because it implies that some things that we would like to be norms are, in fact, semi-norms. Fortunately, we can often make semi-norms into norms by redefining what constitutes equality.[5]

A normed vector space is, in effect, a vector space with a norm. To check if something is a normed vector space, then, we need to check that 1) the vector space satisfies the vector space axioms, and 2) the norm satisfies the norm axioms.

We are often concerned with a subset of the normed vector spaces that are called **inner product spaces**, which is simply a vector space V over a field F with a map

$$\langle \cdot, \cdot \rangle : V \times V \to F$$

that has the following properties $\forall\ \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\forall\ \alpha, \beta \in F$

1. (Conjugate symmetry): $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$[6]
2. (Linearity in the first argument): $\langle \alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w} \rangle = \alpha\langle \mathbf{u}, \mathbf{w} \rangle + \beta\langle \mathbf{v}, \mathbf{w} \rangle$
3. (Positive Definite): So long as $\mathbf{v} \neq \mathbf{0}$, we have that $\langle \mathbf{v}, \mathbf{v} \rangle > 0$

[5] This is a very, very rough characterization of what we need to do, and I will perhaps write an aside about this at a later time.

[6] In the case where the field the vector space is defined over is $\mathbb{R}$, conjugate symmetry is just symmetry. This is the case that we will be interested in in this course, but it is not the technical definition, so I do not use it here.

Note that I said that the space of inner product spaces is a subset of the space of normed vector spaces. This is because every inner product induces a natural norm. Simply define

$$||\mathbf{v}|| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$$

This is called a **canonical norm**[7] on the inner product space, and they are often the ones that we like to use. You are probably aware of the Cauchy-Schwarz inequality, but you may not know that Cauchy-Schwarz holds for any inner product space. I will prove Cauchy-Schwarz for the special case where the field of interest is $\mathbb{R}$. This gives me that the linearity of the inner product can be upgraded to bilinearity, a useful property to simplify the proof.[8] But in order to prove Cauchy-Schwarz, I first generalize the Pythagorean theorem to an arbitrary inner product space over $\mathbb{R}$[9]

[7]Note that the typical inner product we are concerned with on $\mathbb{R}^n$ is the dot product. You should observe that the induced norm under the dot product is exactly the Euclidean norm. This is why, we often refer to the Euclidean norm as the canonical norm on $\mathbb{R}^n$, or simply *the* canonical norm.

[8]In contrast, if the field of interest were $\mathbb{C}$, then inner products are sesquilinear operators.

[9]This is again more strict than necessary, but it makes the proof more straight-forward because bilinearity is a somewhat cleaner property than sesquilinearity.

Pythagorean Theorem for Inner Product Spaces: *Let V be a finite dimensional vector space over $\mathbb{R}$, equipped with an inner product $\langle \cdot, \cdot \rangle$. Then, if $\{v_i\}$ is a collection of vectors in V such that $\langle v_i, v_j \rangle = 0$ if $i \neq j$, we have that*

$$||\sum_{i=1}^{n} v_i||^2 = \sum_{i=1}^{n} ||v_i||^2$$

where $||\cdot||$ denotes the canonical norm given the inner product.

Proof. Let $\{v_i\}_{i=1}^n$ be a collection of orthogonal vectors in V. Then we have that

$$\begin{aligned} ||\sum_{i=1}^{n} v_i||^2 &= \langle \sum_{i=1}^{n} v_i, \sum_{j=1}^{n} v_j \rangle \\ &= \sum_{i=1}^{n} \langle v_i, \sum_{j=1}^{n} v_j \rangle \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} \langle v_j, v_i \rangle \\ &= \sum_{i=1}^{n} \langle v_i, v_i \rangle \\ &= \sum_{i=1}^{n} ||v_i||^2 \end{aligned}$$

Where the first equality follows from the definition of the canonical norm, the second equality from the linearity in the first argument of the inner product, the third equality from real symmetry and again linearity in the first argument, the fourth from the fact that the vectors are orthogonal in the inner product space, and the fifth from the definition of the canonical norm. ■

With this result, we can now prove the Cauchy-Schwarz inequality, again using the stricter than necessary assumption that the field of interest is $\mathbb{R}$.

Cauchy-Schwarz Inequality: *Let V be an inner product space over $\mathbb{R}$ and denote its inner product with $\langle \cdot, \cdot \rangle$. Then $\forall\ \mathbf{u}, \mathbf{v} \in V$, we have that*

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle$$

Proof. Fix $\mathbf{u}, \mathbf{v} \in V$. There are two cases of interest.
In the first case, at least one of the vectors is the zero vector. In this case, the result holds trivially, because both sides of the inequality are zero.
In the second case, both of the vectors are non-zero. Observe now that we can write

$$\mathbf{u} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v} + \left(\mathbf{u} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v} \right)$$

[a]

For notational convenience, let

$$\mathbf{w} = \left(\mathbf{u} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v} \right)$$

So that we can write

$$\mathbf{u} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v} + \mathbf{w}$$

We now use the Pythagorean theorem to deduce that

$$||\mathbf{u}||^2 = ||\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v}||^2 + ||\mathbf{w}||^2$$

Now note that, by bilinearity of the inner product, we have

$$||\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||^2}\mathbf{v}||^2 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{(||\mathbf{v}||^2)^2}||\mathbf{v}||^2 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{||\mathbf{v}||^2}$$

Substituting back in,

$$||\mathbf{u}||^2 = \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{||\mathbf{v}||^2} + ||\mathbf{w}||^2$$

By the non-negativity of norms, then:

$$||\mathbf{u}||^2 \geq \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{||\mathbf{v}||^2}$$

Which implies that

$$||\mathbf{u}||^2||\mathbf{v}||^2 \geq \langle \mathbf{u}, \mathbf{v} \rangle^2$$

Recalling that the canonical norm is given by $|| \cdot || = \sqrt{\langle \cdot, \cdot \rangle}$, we can rewrite the above as:

$$\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle$$

■

[a] If you are wondering where this is coming from, this is an orthogonal decomposition of $\mathbf{u}$.

You can also observe that you should get a triangle inequality for arbitrary inner product spaces, and the proof generalizes exactly in the way you would expect. This is, in fact, a homework problem later in the semester.

I mentioned earlier that not all normed vector spaces are inner product spaces.[10] What determines if a normed vector space is induced by some inner product is something called the Parellelogram law. All inner product spaces satisfy the Parellogram law, and it turns out that if a normed vector space satisfies the Parellogram law, then the norm is induced by some inner product. I will prove that all inner product spaces satisfy the Parellogram law, but I will not prove the other claim.

Claim: *Let V be an inner product space over $\mathbb{R}$ and denote its inner product with $\langle \cdot, \cdot \rangle$. Then for all $\mathbf{u}, \mathbf{v} \in V$, we have that*

$$\langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle + \langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v} \rangle = 2\langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{v}, \mathbf{v} \rangle$$

Proof. The proof is almost entirely algebraic manipulation.

$$\begin{aligned} \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle &= \langle \mathbf{u}, \mathbf{u}+\mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u}+\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \end{aligned}$$

$$\begin{aligned} \langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v} \rangle &= \langle \mathbf{u}, \mathbf{u}-\mathbf{v} \rangle - \langle \mathbf{v}, \mathbf{u}-\mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle - 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \end{aligned}$$

Adding the two equations above then, we see that $2\langle \mathbf{u}, \mathbf{v} \rangle$ cancels, leaving us with $2\langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{v}, \mathbf{v} \rangle$, which was what was wanted. ■

I now state a very famous (indeed perhaps the most famous) result from the theory of Hilbert spaces: the Hilbert Projection Theorem.

[10] Also if you are keeping track of the kind of spaces that inner product spaces are (beyond just inner product spaces of course): inner products define a canonical norm, and hence are normed vector spaces. Since every norm induces a metric via $d(x, y) = ||x - y||$, we have that every inner product space is also an induced metric space. If the inner product space viewed as a metric space is complete in the sense that all Cauchy sequences converge, we say that it is a **Hilbert space**. It turns out that all L^p spaces are complete, so $L^2(X, \mathcal{X}, \mu)$ is a Hilbert space. This is important for a few reasons but figures prominently in the modern treatment of conditional expectation, which relies on projections in Hilbert spaces. Finally, since metrics induce topologies, we see that every inner product space is also a topological space, where the topology is the one induced by that same metric.

Hilbert Projection Theorem: *Let $\mathcal{H}$ be a Hilbert space, $\mathcal{M}$ be a closed subspace of $\mathcal{H}$, and y be an element of $\mathcal{H}$. Then*

- *There exists a unique element $\hat{\mu} \in \mathcal{M}$ such that $||y - \hat{\mu}|| = \inf_{\mu \in \mathcal{M}} ||y - \mu||$*[a]
- *$\hat{\mu} \in \mathcal{M}$ and $||y - \hat{\mu}|| = \inf_{\mu \in \mathcal{M}} ||y - \mu||$ if and only if $\hat{\mu} \in \mathcal{M}$ and $(y - \hat{\mu}) \in \mathcal{M}^{\perp}$, where $\mathcal{M}^{\perp}$ is the orthogonal complement to $\mathcal{M}$ in $\mathcal{H}$.*

Proof. ■

[a] Neither the existence nor uniqueness of the projection rely on the fact that $\mathcal{M}$ is a subspace. Indeed we can weaken this first part to simply the requirement that $\mathcal{M}$ is a closed and convex subset of $\mathcal{H}$. This weakening (and the accompanying restriction that the $\mathcal{M}$ be a subspace in the second part of the theorem), constitutes the most general form of the Hilbert Projection theorem.

Claim: *Let $\mathcal{H}$ be a Hilbert space. Let $\mathcal{M}$ be a Hilbert subspace, and denote its orthogonal complement by $\mathcal{M}^{\perp}$. Then $\mathcal{M} \cap \mathcal{M}^{\perp} = \{\mathbf{0}\}$.*

Proof. Let $\mathcal{M}$ be given as above. Let $\mathbf{v} \in \mathcal{M} \cap \mathcal{M}^{\perp}$. Then $\mathbf{v} \in \mathcal{M}$ and $\mathbf{v} \in \mathcal{M}^{\perp}$. Then $\langle \mathbf{v}, \mathbf{v} \rangle = 0$. But then $||\mathbf{v}|| = 0$, which holds if and only if $\mathbf{v} = \mathbf{0}$. ■

Claim: *Let $\mathcal{H}$ be a Hilbert space. Let $\mathcal{M}$ be a finite dimensional closed Hilbert subspace of $\mathcal{H}$, and denote its orthogonal complement as $\mathcal{M}^{\perp}$. Then the following is true:*

$$\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}$$

Proof. Let $\{\mathbf{e}_i\}_{i=1}^k$ be an orthonormal basis for $\mathcal{M}$. Let $\mathbf{h} \in \mathcal{H}$ be given. Define $\mathbf{r} = \sum_{i=1}^k \langle \mathbf{h}, \mathbf{e}_i \rangle \mathbf{e_i}$.[a] Observe now that for all $j \in \{1, 2, \ldots, k\}$

$$\begin{aligned}
\langle \mathbf{h} - \mathbf{r}, \mathbf{e}_j \rangle &= \langle \mathbf{h}, \mathbf{e}_j \rangle - \langle \mathbf{r}, \mathbf{e}_j \rangle \\
&= \langle \mathbf{h}, \mathbf{e}_j \rangle - \langle \langle \mathbf{h}, \mathbf{e}_j \rangle \mathbf{e}_j, \mathbf{e}_j \rangle \\
&= \langle \mathbf{h}, \mathbf{e}_j \rangle - \langle \mathbf{h}, \mathbf{e}_j \rangle \langle \mathbf{e}_j, \mathbf{e}_j \rangle \\
&= \langle \mathbf{h}, \mathbf{e}_j \rangle - \langle \mathbf{h}, \mathbf{e}_j \rangle \\
&= 0
\end{aligned}$$

Hence $\mathbf{h} - \mathbf{r} \perp \mathcal{M}$.
Observing now that $(\mathbf{h} - \mathbf{r}) + \mathbf{r} = \mathbf{h}$, we have the desired result.

■

[a] This is the projection.

Claim: *Let $\mathcal{H}$ be a Hilbert space. Let $\mathcal{M}$ be a finite dimensional closed Hilbert subspace of $\mathcal{H}$, and denote its orthogonal complement as $\mathcal{M}^{\perp}$. Then the following is true:*

$$dim(\mathcal{H}) = dim(\mathcal{M}) + dim(\mathcal{M}^{\perp})$$

Proof. Let $\mathcal{H}, \mathcal{M}$ be given. Suppose without loss of generality that $dim(\mathcal{H}) = n$ and that $dim(\mathcal{M}) = m$, with $m < n$. Let $\{\mathbf{v}_i\}_{i=1}^{m}$ be an orthonormal basis for $\mathcal{M}$. Collect the basis vectors as column vectors in a matrix and call this matrix $\mathbf{A}$. By construction then $dim(\mathcal{M}) = rank(\mathbf{A}) = m$. Now observe that the column vectors of $\mathbf{A}'$ constitute a basis for $Ker(\mathbf{A}') = \mathcal{M}^{\perp}$.
Now, recall that $rank(\mathbf{A}) = rank(\mathbf{A}')$.
By the rank-nullity theorem then $rank(\mathbf{A}') + nullity(\mathbf{A}') = n$ ■

Triangle Inequality in Hilbert Spaces: *Let $\mathcal{H}$ be a Hilbert space.*[a] *Let the inner product $\mathcal{H}$ be given by $\langle \cdot, \cdot \rangle$. Let $|| \cdot ||$ denote the induced norm. Then $\forall \mathbf{u}, \mathbf{v} \in \mathcal{H}$, we have that*

$$||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$$

Proof.

$$\begin{aligned} ||\mathbf{u} + \mathbf{v}||^2 &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle \\ &= ||\mathbf{u}||^2 + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + ||\mathbf{v}||^2 \\ &= ||\mathbf{u}||^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + ||\mathbf{v}||^2 \\ &\leq ||\mathbf{u}||^2 + 2|\langle \mathbf{u}, \mathbf{v} \rangle| + ||\mathbf{v}||^2 \\ &\leq ||\mathbf{u}||^2 + 2||\mathbf{u}||||\mathbf{v}|| + ||\mathbf{v}||^2 \\ &= (||\mathbf{u}|| + ||\mathbf{v}||)^2 \end{aligned}$$

Taking square roots on each side delivers the desired result. ■

[a] Over field $\mathbb{R}$.

It is natural to wonder why we might care about Hilbert spaces. One of the reasons is that Hilbert spaces are much much more general than $\mathbb{R}^n$. Famously, to solve many of the more important problems in partial differential equations, one must work in Sobolev spaces, which are also Hilbert spaces. The magic of the projection theorem, and of many of the results in the theory of Hilbert spaces in general, is that our basic geometric intuitions from $\mathbb{R}^2$ carry through into *very* abstract mathematical spaces. To demonstrate how broad the collection of Hilbert spaces is, I consider a specific example directly relevant to what we are learning in class.

1.5 The Space of Square-Integrable Random Variables

Let us fix a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the space of random variables X from this space to $(\mathbb{R}, \mathcal{B})$ that satisfy the following two conditions:

1. (Mean Zero)

$$\int_{\Omega} X d\mathbb{P} = 0$$

2. (Square Integrable)

$$\int_\Omega |X|^2 d\mathbb{P} < \infty$$

Call this space $L^2(\Omega)_0$[11]

[11] Again note that we are going to work with "functions" as representatives of their almost everywhere equivalence classes.

Claim: *Covariance is an inner product on this space.*

Proof. Let X and Y be elements of $L^2(\Omega)_0$. Observe that the covariance of X and Y is given by

$$\mathbb{E}[X - \mathbb{E}[X]]\mathbb{E}[Y - \mathbb{E}[Y]]$$

Observe also that by property 1, we have that $\mathbb{E}[X] = X$ and the same for Y, so the above simplifies to $\mathbb{E}[XY]$. Simple inspection verifies that $\mathbb{E}[XY] = \mathbb{E}[YX]$.
Now we verify linearity. Let X,Y,and Z be given, with $\alpha, \beta \in \mathbb{R}$ given as well.
By the above, we have that

$$cov(X, Y) = \mathbb{E}[XY] = \int_\Omega XY d\mathbb{P}$$

A similar argument establishes that

$$cov(\alpha X + \beta Z, Y) = \mathbb{E}[(\alpha X + \beta Z)Y] = \int_\Omega (\alpha X + \beta Z)Y d\mathbb{P} = \int_\Omega \alpha XY + \beta ZY d\mathbb{P}$$

Exploiting linearity of integrals,[a] we then have that

$$\int_\Omega \alpha XY + \beta ZY d\mathbb{P} = \alpha \int_\Omega XY d\mathbb{P} + \beta \int_\Omega ZY d\mathbb{P} = \alpha cov(X, Y) + \beta cov(Z, Y)$$

Now we need to prove that the covariance is positive definite in this environment. To that end let $X \in L^2(\Omega)_0$ be such that $X \neq 0$[b] Since X is a random variable, X is an $\mathcal{F}$-measurable function. Partitioning X into the portions where it is greater than 0 and those where it is less than 0, and calling them X_+ and X_- respectively, we have that there exist simple functions S_+ and S_- that minorize X_+ and $-(X_-)$ respectively. Since $X \neq 0$, there must exist S_+ and S_- not equal to zero.[c]
But then observe that $S \equiv \max(S_+, S_-)$ minorizes $|X|$, and likewise $S^2 \equiv \max(S_+^2, S_-^2)$ minorizes $|X|^2$. Now, exploiting the monotonicity of unsigned integrals, we have that

$$0 < \int_\Omega S^2 d\mathbb{P} \leq \int_\Omega |X|^2 d\mathbb{P}$$

Thus we have that covariance is positive definite, and since it satisfies all three properties, an inner product on this space. ■

[a] This is proved in the measure theory notes.
[b] This means that X differs from 0 on a set of measure greater than 0.
[c] If both are zero, then the only simple function that minorizes X would be 0, which contradicts our assumption that $X \neq 0$. If only one were positive that would imply that the expectation is greater than zero, contradicting the presence of X in $L^2(\Omega)_0$.

Now I have, in some sense, put the carriage before the horse here. Recall that we require that an inner product space be a *vector space* over some field F, before we endow it with an inner product. But how do we know that the space of square integrable functions is a vector space? What does it mean for a space of functions to be a vector space? It is not immediately obvious that this defines a vector space, but one can, with a bit of work, show that it verifies the axioms above. Doing this is in fact a good exercise.

I will also note that the way that I defined an inner product can be "fixed" to compensate for the fact that we want to consider random variables that do not have mean zero. The way that this is done is with the canonical inner product on $L^2(\Omega, \mathcal{F}, \mathbb{P})$, which is given by

$$\langle X, Y \rangle \equiv \int_\Omega XY \, d\mathbb{P}$$

. It is not a coincidence that this is what covariance in the case above reduces to in the case when $\mathbb{E}[X] = 0$. This is exactly the condition that is required to make covariance an inner product. Observe also that the induced norm in this formulation is exactly the L^2 norm.

Now we shall work with this space of functions. As before we are going to restrict our consideration to functions that are square integrable, but no longer are we going to use covariance as our inner product, rather we are going to use the inner product above.

Claim: $Cov(X,Y)^2 \leq Var(X)Var(Y)$

Proof. Let $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$.

$$\begin{aligned} Cov(X,Y)^2 &= \left(\int_\Omega (X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) d\mathbb{P} \right)^2 \\ &\leq \int_\Omega (X - \mathbb{E}[X])^2 d\mathbb{P} \int_\Omega (Y - \mathbb{E}[Y])^2 d\mathbb{P} \\ &= Var(X)Var(Y) \end{aligned}$$

■

2 Null Hypothesis Statistical Testing

Suppose that we want to make statements about a population. Say, for instance, something like the claim people with larger aims tend to be less supportive of redistribution[12] The problem with making claims like this is that we in general do not observe the whole population. Although in theory we could certainly go up to every person in the world, measure the size of their arms and then ask them what they think about redistribution, this is obviously not feasible. As a result, in general we limit our consideration to samples.

Null hypothesis statistical testing is one way - indeed the traditional way - that we go from looking at samples to making comments about population parameters.[13]

Traditional NHST starts from two claims:

1. **The Null Hypothesis** which we often abbreviate H_0, which represents the claim that there is no change or effect or relationship depending on your context. If we are running an experiment, the null says that there is no relationship between the independent variable and the dependent variable.

2. **The Alternative Hypothesis** which we abbreviate H_A or H_1. It says that there is a relationship.

[12] I am not making this up, someone actually did a study like this.

[13] That is not to say that it is the only way. Some very prominent statisticians, including e.g. Andrew Gelman criticize this approach, and indeed this has been a somewhat controversial topics since at least the seventies.

To decide between the two, we make use of a **test statistic**, which has the important property that it's distribution, given that the null is true, is calculable. This allows us to determine the relative likelihood of observing the data given the distribution implied by the null. Loosely, if it is sufficiently unlikely that we would see the observed data conditional on the DGP implied by the null, we reject the null.

Mathematically, we can write both the null and the alternative hypothesis in the following manner:

$$H_0 : \theta = \theta_0$$

Where θ_0 is the assumed "truth."

$$H_A : \{\theta \in \Theta | \theta \neq \theta_0\}$$

Where again θ_0 is the assumed "truth."

$T(\{y_i, x_i\}_{i=1}^n)$ is a test statistic with a distribution F_{H_0} that is derived from the assumed "truth" under the null.

We set a threshold α that basically is our cutoff for a test statistic being from the distribution implied by the null.

Now we want to talk about errors. There are two kinds of errors that we are concerned about:

1. Failing to reject the null even though the null was not true. This is called a **Type II Error**, and we denote the probability of this happening as β.

2. Rejecting the null even though the null was true. This is called a **Type I Error**, and we denote the probability of this α. Note the probability of doing this is exactly the threshold value we have selected, hence why we say it is α.

Generally, there is a tension because if we pick a lower α as our threshold, the corresponding β goes up.

One way to (sort of) reconcile the two is to fix some value for α and then select among our test statistics for the one that maximizes the power, where the **Statistical Power** of a test is $1 - \beta$.

How do we do this? There are a few ways:

1. Get more data.

2. Select a different test statistic.

3. Raise α (don't do this)

3 Moment Generating Functions

Suppose that we have a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on this triple. One way to talk about X is to talk about its distribution measure $\mu_X(\cdot)$, or more commonly, its associated cumulative distribution function $F_X(\cdot)$. Another way, if it exists, is to talk about the probability density function f associated with $\mu_X(\cdot)$.[14] A third way that we can sometimes do is to talk about the moment generating function of a random variable X, or less precisely, its moments.

The $\mathbf{n^{th}}$ **uncentered moment** of a random variable is given by $\mathbb{E}[X^n]$.

[14] Or rather a representative of the equivalence class of functions that when integrated recover the CDF.

The **n^{th} centered moment** of a random variable is given by $\mathbb{E}[(X - \mathbb{E}[X])^n]$

At least for the centered moments, these should look familiar to you. Consider the second centered moment of a random variable X. Simply plugging two into the formula above, we recover

$$\mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X]) = \mathbb{E}[(X^2-2X\mathbb{E}[X]-(\mathbb{E}[x])^2] = \mathbb{E}[X^2]-(\mathbb{E}[X])^2 = \mathbb{V}[X]$$

The second centered moment is the variance of a distribution. Similarly, the third centered moment is the **skew**[15] of the distribution, and the fourth the **kurtosis**.[16] This I think provides some intuition as to what moments are: in some sense they are just parameters that tell us things about the shape of the distribution.

You should notice that in general it is going to be painful to find moments by calculating $\mathbb{E}[X^n]$. Fortunately, for many of the distributions we work with, we have access to what is called a **Moment Generating Function**, which we denote by $M_X(t)$.

More specifically, we say that a moment generating function $M_x(t)$ is given by

$$M_x(t) = \mathbb{E}[e^{tX}]$$

Provided there $\exists \epsilon$ such that $\forall t \in B(0, \epsilon)$[17] where this expectation exists. There is a close relationship between moment generating functions and characteristic functions of random variables, but we will not discuss this in depth other than to say that whereas the characteristic function always exists, the moment generating function does not.[18]

One might naturally wonder how do we get the moments from the moment generating function, and the answer lies in the use of a taylor expansion of e^{tX}. Specifically recall that that taylor expansion of e^x about the origin is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Plugging the taylor expansion for e^{tx}, which is very similar to above (simply replace x with tx), we have that

$$\mathbb{E}[e^{tX}] = \mathbb{E}[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots]$$

Exploiting the linearity of the expectation operator, we can then rewrite the above as:

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + \frac{t^2\mathbb{E}[X^2]}{2!} + \frac{t^3\mathbb{E}[X^3]}{3!} + \dots$$

Notice that within each of the separate terms in the above, we have an uncentered moment of X. To recover the n^{th} moment of the random variable, we simply differentatiate the above expression n times, and evaluate the expression at $t = 0$.

So to find the first moment, we would take the derivative once and evaluate the derivative at $t = 0$, leaving us with

[15] The skew is approximately a measure of how "lopsided" a distribution is. A symmetric distribution will always have a third central moment equal to zero.

[16] The kurtosis is essentially a measure of how fat the tails of the distribution are.

[17] This is an open ball of radius ϵ centered around the origin.

[18] For an example where the moment generating function does not exist (but the characteristic function does), consider a Cauchy random variable.

$$\mathbb{E}[X] + 0 * \mathbb{E}[X^2] + \frac{0^2 * \mathbb{E}[X^3]}{2!} + \dots$$

As this is somewhat abstract, let us consider an example involving a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

Recall that a normal random variable has a density given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

By the law of the unconscious statistician then, we can write

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Simplifying this expression, we then recover

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{2\sigma^2 tx + 2x\mu - x^2 - \mu^2}{2\sigma^2}}$$

Consider the exponentiated expression:

$$2\sigma^2 tx + 2x\mu - x^2 - \mu^2$$

We can rewrite this as

$$-(x^2 - 2x(\mu + t\sigma^2)) - (\mu + t\sigma^2)^2 + (\mu + 2t\sigma^2)^2 - \mu^2$$

Observe that this is equivalent to

$$-(x^2 - 2x(\mu + t\sigma^2)) + (\mu + t\sigma^2)^2) + (\mu + t\sigma^2)^2 - \mu^2$$

$$-(x - (\mu + 2t\sigma^2))^2 + \mu^2 + 2t\mu\sigma^2 + t^2\sigma^4 - \mu^2$$

$$-(x - (\mu + 2t\sigma^2))^2 + t\sigma^2(2\mu + t\sigma^2)$$

Substituting this in, we have that

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-(\mu+2t\sigma^2))^2 + t\sigma^2(2\mu+t\sigma^2)}{2\sigma^2}} dx$$

$$= e^{\frac{2t\mu + t\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-(\mu+2t\sigma^2))^2}{2\sigma^2}} dx$$

Observing that expression on the right of the integral is the density function of a normal random variable with $\mu' = \mu + t\sigma^2$, and variance σ^2, the expression in the integral integrates to unity, leaving us with.

$$\mathbb{E}[e^{tX}] = e^{t\mu + \frac{t^2\sigma^2}{2}}$$

Moment generating functions are useful in quite a few ways.

For example, suppose that we have random variables X_1 and X_2 and X_1

We will now derive some facts about one of the most important of the basic distributions that one learns, the Chi-squared distribution.

Claim: *Suppose that* $X \sim \mathcal{N}(0,1)$*. Then* $Y = X^2 \sim \chi_1^2(0)$*.*

Proof. Observe that $P(Y \leq y) = 0$ for all $y \in (-\infty, 0)$. For $y \geq 0$, we have that

$$P(Y \leq y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y})$$

Denote the cumulative distribution function of the standard gaussian at $x \in \mathbb{R}$ as $\Phi(x)$. Then we have that

$$P(-\sqrt{y} < X < \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$$

Exploiting the symmetry of the standard Gaussian, we then have that

$$\Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

Substitituing in for $\Phi(\cdot)$, we then have that

$$2\Phi(\sqrt{y}) - 1 = 2\int_{-\infty}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt - 1 = \frac{\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{\sqrt{y}} e^{\frac{-t^2}{2}} dt - 1$$

Differentiating with respect to y, we recover that

$$f_y(y) = \frac{1}{\sqrt{2}\sqrt{\pi}} e^{-\frac{y}{2}} \frac{1}{\sqrt{y}}$$

Recalling that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we have that

$$f_y(y) = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} e^{-\frac{y}{2}} \frac{1}{\sqrt{y}}$$

Which is exactly the PDF of a $\chi_1^2(0)$. ■

We will now use this fact to derive the moment generating function of a chi-squared.

Claim: *The moment generating function of a random variable $Y \sim \chi^2(1)$ is given by*

$$M_Y(t) = \frac{1}{\sqrt{1-2t}}$$

Proof. By the above, we have that $Y = X^2$, where $X \sim \mathcal{N}(0,1)$. The moment generating function of Y is thus equivalent to that of X^2, and is given by

$$\mathbb{E}[e^{tX^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{x^2}{2}} dx$$

Observing that

$$e^{tx^2} e^{-\frac{x^2}{2}} = e^{\frac{2tx^2 - x^2}{2}} = e^{\frac{x^2(2t-1)}{2}} = e^{\frac{1-2t}{2}(-x^2)}$$

It turns out that

$$\int_{-\infty}^{\infty} e^{\frac{x^2(2t-1)}{2}} dx = \frac{\sqrt{2\pi}}{\sqrt{(1-2t)}}$$

[a]

Hence

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2\pi}}{\sqrt{(1-2t)}} = \frac{1}{\sqrt{1-2t}}$$

■

[a] This is because we have written the expression as a standard Gaussian integral.

Now noting from our earlier discussion of the properties of the sums of random variables, we have that the sum of independent random variables has a moment generating function equal to their product. Applying this to a sum of k squared standard normal random variables, we then have that $Z = \sum_{i=1}^{k} X^2$ has a moment generating function given by

$$M_z(t) = (\frac{1}{\sqrt{1-2t}})^k = (1-2t)^{\frac{-k}{2}}$$

This is unsurprisingly a $\chi^2_k(0)$ random variable.

Note that a Chi-Squared with k degrees of freedom here is denoted by $\chi^2_k(0)$. This is because what we are directly referring to is a *central* χ^2 with k degrees of freedom. The zero here denotes that non-centrality parameter of the Chi-Squared distribution at issue is zero. You will often see the centrality of the χ^2 distribution assumed, but there is at least one case where we are going to see non-central Chi-Squared variables later, so it is worth noting that non-central distributions exist.

3.2 Moment Problems

One might be tempted to assume based on the earlier claim that if two functions' moments are the same, then they are the same distribution. This claim, however, is not true. Consider a random

variable X that is distributed with the log standard normal distribution (i.e. $X \sim \log(\mathcal{N}(0,1))$). X then has a density function given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{ln(x)^2}{2}}$$

This function has moments in the specific sense that $\mathbb{E}[X^n]$ exists, and in general, the n^{th} moment of the distribution of a standard log-normal is going to be:

$$\mathbb{E}[X^n] = e^{\frac{n^2}{2}}$$

[19]

Note that the moment generating function of the log normal distribution *does not exist in a neighborhood of 0*. This is in some sense a crucial observation, because we know that the moment generating function *does determine uniquely a distribution*. To see that the moments generated by the above do not uniquely determine the distribution, denote the density of the standard log normal by $f_X(x)$, and consider the density $g(x)$ given by:

$$g(x) = f_X(x)[1 + \sin(2\pi \ln(x))]$$

What you will find is that these two densities both define an identical series of moments, but the two are not identical almost everywhere (with respect to the Lebesque measure on $\mathbb{R}$).

The question then naturally becomes, are there sufficient conditions that we can impose on the random variables so that the sequence of moments - and not the moment generating function itself - uniquely defines the distribution. The problem of going from a sequence of moments to a distribution and the random variable that defines such a sequence is called a **Moment Problem**.

[19] The general log-normal will have moments given by $\mathbb{E}[X^n] = e^{n\mu + \frac{n^2\sigma^2}{2}}$

4 Quadratic Forms of Random Variables

Suppose that X is a multivariate distribution with vector of means μ_X and variance covariance matrix Σ_X. That is:

$$X \sim (\mu_X, \Sigma_X)$$

Define $Z = \mathbf{A}X$, with $\mathbf{A}$ conformable.

Claim: $\mathbb{E}[Z] = A\mu_X$

Proof.

$$\mathbb{E}[Z] = \mathbb{E}[\mathbf{A}X] = \mathbf{A}\mathbb{E}[X] = A\mu_X$$

■

Claim: $\mathbb{V}[Z] = \mathbf{A}\Sigma_X\mathbf{A}'$

Proof.

$$\begin{aligned}
\mathbb{V}[Z] &= \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'] \\
&= \mathbb{E}[(\mathbf{A}X - \mathbb{E}[\mathbf{A}X])(\mathbf{A}X - \mathbb{E}[\mathbf{A}X])'] \\
&= \mathbb{E}[\mathbf{A}(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\mathbf{A}'] \\
&= \mathbf{A}\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']\mathbf{A}' \\
&= \mathbf{A}\Sigma_X\mathbf{A}'
\end{aligned}$$

■

Now let us make a more stringent distributional assumption, namely that

$$X \sim \mathcal{N}(\mu_X, \sigma^2\mathbb{I}_n)$$

Now let $\mathbf{A}$ be a $n \times n$ symmetric and idempotent matrix[20] with $rank(\mathbf{A}) = k$

[20] This implies that $\mathbf{A}$ is a projection matrix.

Claim:

$$\frac{X'\mathbf{A}X}{\sigma^2} \sim \chi^2_k(\frac{\mu'_X\mathbf{A}\mu_X}{\sigma^2})$$

Proof. Let us begin by observing that $\mathbf{A}$ is a symmetric matrix and thus has a spectral decomposition so that $A = \mathbf{V}\Lambda\mathbf{V}'$. Since $\mathbf{A}$ is idempotent, it has only zero or one as eigenvalues, and since it is rank k, we have k occurrences of $\lambda = 1$. Without loss of generality, reorganize Λ and $\mathbf{V}$ so that the first k diagonal entries are $\lambda = 1$ (and the corresponding eigenvectors are also in the right place), and the remaining $n - k$ diagonal entries are zero. Hence we have

$$\begin{aligned}
\frac{X'\mathbf{A}X}{\sigma^2} &= \frac{X'\mathbf{V}\Lambda\mathbf{V}'X}{\sigma^2} \\
&= \frac{(\mathbf{V}'X)'\Lambda(\mathbf{V}'X)}{\sigma^2} \\
&= \frac{Y'\Lambda Y}{\sigma^2} \\
&= \sum_{i=1}^{k} \frac{Y_i^2}{\sigma^2}
\end{aligned}$$

Now observe that $\frac{Y_i^2}{\sigma^2} \sim \chi^2_1((\frac{v'_i\mu}{\sigma})^2)$. Hence the sum of k of them will be a Chi-Squared distribution with k degrees of freedom and non-centrality parameter $\frac{\mu'_X\mathbf{A}\mu_X}{\sigma^2}$.[a] ■

[a] Observe that if $\mathbf{A}\mu_X = 0$, then

$$\mu'_X\mathbf{A}\mu_X = \mu'_X\mathbf{A}'\mathbf{A}\mu_X = (\mathbf{A}\mu_X)'(\mathbf{A}\mu_X) = 0$$

Claim:

$$\frac{\epsilon'\mathbf{M}\epsilon}{\sigma^2} \sim \chi^2_{n-k}(0)$$

Proof. Observe that $\mathbf{M}$ is a symmetric matrix, and thus admits a spectral decomposition, which again we will notate as $\mathbf{V}\Lambda\mathbf{V}'$, which we will exploit here. Also since it is idempotent, it has either 0 or 1 as its eigenvalues. Without loss of generality, I can reorganize the first $n-k$ diagonal entries in Λ so that they consist of the $n-k$ entries of $\lambda = 1$, and reorganize the linearly independent columns of $\mathbf{V}$ so that they properly correspond as well.

$$\begin{aligned}
\frac{\epsilon'\mathbf{M}\epsilon}{\sigma^2} &= \frac{\epsilon'\mathbf{V}\Lambda\mathbf{V}'\epsilon}{\sigma^2} \\
&= \frac{(\mathbf{V}'\epsilon)'\Lambda(\mathbf{V}'\epsilon)}{\sigma^2} \\
&= \frac{Y'\Lambda Y}{\sigma^2} \\
&= \sum_{i=1}^{n-k} \frac{Y_i^2}{\sigma^2}
\end{aligned}$$

Now observe that $\frac{Y_i^2}{\sigma^2} \sim \chi^2_1(0)$.[a] Hence the sum of $n-k$ of them will be a Chi-Squared distribution with $n-k$ degrees of freedom and non-centrality parameter 0. Hence we have that

$$\frac{\epsilon'\mathbf{M}\epsilon}{\sigma^2} \sim \chi^2_{n-k}(0)$$

■

[a] To see this, observe that the j-th entry in $\mathbf{V}'\epsilon$ is given by $v_j \cdot \epsilon = \sum_{i=1}^n \epsilon_i v_{ji}$. Taking expectations on both sides, it is obvious that the expectation of the sum is 0. Taking variances on both sides, we have that $Var(v_j \cdot \epsilon) = Var(\sum_{i=1}^n \epsilon_i v_{ij}) = \sum_{i=1}^n v_{ji}^2 Var(\epsilon_i)$, where the fact that we can interchange the variance operator and the sum without consideration of covariance follows from the independence of the ϵ_i. Observing that $Var(\epsilon_i) = \sigma^2$, we then have that $\sum_{i=1}^n v_{ji}^2 Var(\epsilon_i) = \sigma^2 \sum_{i=1}^n v_{ji}^2 = \sigma^2$, where the last equality follows from the fact that V is an orthonormal collection of eigenvectors, and hence $\sum_{i=1}^n v_{ji}^2 = 1$. Normalizing each ϵ_i by σ delivers that $\frac{Y_i}{\sigma} \sim \mathcal{N}(0,1)$.

5 Generalized Least Squares

5.1 An example

Recall the traditional OLS assumptions:

1. (Model Linearity): $y_i = \mathbf{X}_i\beta + \epsilon_i$
2. (Strict Exogeneity): $\mathbb{E}[\epsilon_i|\mathbf{X}] = 0$
3. (Full Rank): $\mathbb{P}(rank(\mathbb{X}) = k) = 1$
4. (Independence and no Heteroskedasticity or Autocorrelation): $\mathbb{E}[\epsilon\epsilon'|\mathbf{X}] = \sigma^2\mathbf{I}_n$

Today we are interested in thinking about what happens when we relax assumption 4, so that we still have independent errors, but they are no longer identically distributed in a particular way. To motivate this problem, consider the following example. You have data on 2954 counties in the United States in the form of average household income and the population of each of the counties. You for whatever reason think that the only thing that affect coronavirus morbidity is the average income of a household (this is of course completely insane, but let's just assume it for educational purposes). You want to evaluate the effect of average income on say coronavirus morbidity.[21]

Denote by y_{ij} the morbidity rate of household j in county i, denote by x_{ij} the household income of household j in county i, and further let's say that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Let's assume that assumptions 1-4[22], hold in the granular dataset, where we observe each of the individual households' income and morbidity.

This means that our generating model is given by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

The problem in our setting of course is that we don't observe individual households, we observe county averages, whether of the morbidity rate,[23] or of household income. That means that we want to run the following regression:

$$y_i = \hat{\beta} + \hat{\beta}_1 \bar{x}_i + e_i$$

The question here is if the naive ordinary least squares estimator here is our best linear unbiased estimator. In order for that to be true, we need to show that the four assumptions above are satisfied, so then we can invoke Gauss-Markov.

Claim: *The aggregated model satisfies the linearity assumption.*

Proof.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

Hence

$$\frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} = \beta_0 + \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i} + \frac{\sum_{j=1}^{N_i} \epsilon_i}{N_i}$$

Which further can be written as

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{\epsilon}_i$$

So $\bar{y}_i$ is linear in $\bar{x}_i$, which was what was wanted. ■

[21] Morbidity is the rate of disease in a population. Mortality is the percentage of people who die.

[22] *A*1.1 to *A*1.4 in Hayashi

[23] As another added wrinkle, assume that all households are the same size, so that our aggregation of the morbidity rate is the simple arithmetic mean of the morbidity rates of the households.

Claim: *The aggregated model satisfies the strict exogeneity assumption.*

Proof. Begin by observing that if $\mathbb{E}[\epsilon_{ij}|X] = 0$, then $\mathbb{E}[\epsilon_{ij}|\bar{X}] = 0$.[a]
Now observe that $\sum_{j=1}^{N_i} \epsilon_{ij} = N_i \bar{\epsilon}_i$
Taking conditional expectations on both sides, we have that:

$$\mathbb{E}[\bar{\epsilon}_i|\bar{X}] = \mathbb{E}[\frac{\sum_{j=1}^{N_i} \epsilon_{ij}}{N_i}|\bar{X}] = \frac{\sum_{j=1}^{N_i} \mathbb{E}[\epsilon_{ij}|\bar{X}]}{N_i} = 0$$

Where the first equality follows from the definition of $\bar{\epsilon}_i$, the second from the linearity of the expectation operator, and the last from strict exogeneity in the granular model. ■

[a] In general $\mathbb{E}[Y|X] = \mathbb{E}[Y|g(X)]$ so long as $g(\cdot)$ is measurable and injective.

Just take it for granted that it satisfies the full rank assumption. In practice you just can look at the correlation matrix and see instantly if you have a problem, so it's worth worrying much about this.

All that remains is the spherical error assumption, but this one, unfortunately, does not hold.

Claim: *The aggregated model's errors have variance-covariance matrix $V \neq \sigma^2 \mathbb{I}_n$.*

Proof. From our assumptions on the granular model, we know that

$$\mathbb{E}[\epsilon_{ij}\epsilon_{ik}] = 0 \forall j \neq k$$

We also know that

$$\mathbb{E}[\epsilon_{ij}^2|\bar{X}] = \sigma^2$$

$$\begin{aligned}
\mathbb{E}[\bar{\epsilon}_i^2|\bar{X}] &= \mathbb{E}\left[\left(\frac{\sum_{j=1}^{N_i} \epsilon_{ij}}{N_i}\right)^2 |\bar{X}\right] \\
&= \frac{1}{N_i^2}\mathbb{E}\left[(\sum_{j=1}^{N_i} \epsilon_{ij})^2|\bar{X}\right] \\
&= \frac{1}{N_i^2}\mathbb{E}\left[\sum_{j=1}^{N_i} \epsilon_{ij}^2|\bar{X}\right] \\
&= \frac{N_i}{N_i^2}\mathbb{E}\left[\epsilon_{ij}^2|\bar{X}\right] \\
&= \frac{\sigma^2}{N_i}
\end{aligned}$$

Where the third equality follows from the fact that ϵ_{ij}

What does this mean? Well for one thing it means that our error distribution has a variance-covariance matrix given by:

$$\begin{bmatrix} \frac{\sigma^2}{N_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{\sigma^2}{N_2} & 0 & \dots & 0 \\ 0 & 0 & \frac{\sigma^2}{N_3} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{\sigma^2}{N_{2954}} \end{bmatrix}$$

But another thing that it means is the OLS estimator is no longer BLUE. We can do better, and seeing how we can do better is in fact pretty easy. We know the theoretical difference between what we would like the variance covariance matrix to be to apply Gauss Markov, and what it actually is. We observe the population of each of the counties, so why don't we just reweight the observations to place more emphasis on those observations that come from a bigger sample. This is the essence of the idea behind weighted least squares, but in order to do so, we have to review a bit of theory.

5.2 Generalized Least Squares: Theory

Let us maintain assumptions 1 through 3 above, and now let us consider a weakening of the spherical error assumption to the assumption below:

4b.) $\mathbb{E}[\epsilon\epsilon'|\mathbf{X}] = \sigma^2\mathbf{V}(\mathbf{X})$ with $\mathbf{V}(\mathbf{X})$ known.[24]

Obviously the Gauss-Markov result is not going to hold in this setting, but the motivation provided by the above example gives us an indication as to how we can get the best linear unbiased estimator in this setting. The game, in effect, is to transform the above model into one that *does* satisfy the assumptions of Gauss-Markov. The estimator implied by that set of transformations will then therefore be the BLUE.

To derive such an estimator, let us begin by observing that $\mathbf{V}$ is a positive definite matrix. Hence it's inverse is also positive definite and thus admits a square root decomposition. Therefore we can write

$$\mathbf{V}^{-1} = \mathbf{P}'\mathbf{P}$$

Using this, let us now define the following:

$$\tilde{\mathbf{y}} = \mathbf{P}y, \tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}, \tilde{\boldsymbol{\epsilon}} = \mathbf{P}\boldsymbol{\epsilon}$$

I now claim that this transformed model satisfies all of the assumptions of Gauss-Markov.

[24] I will notate this simply $\mathbf{V}$ from now on.

Claim: *The transformed model satisfies the linearity assumption if the original model satisfies assumptions* 1 *and* 4*b*.

Proof. If the true DGP satisfies the linearity assumption, then:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

If assumption 4*b* is satisfied, then $\mathbf{V}$ admits a square root decomposition that has full rank.[a] Calling such a square root decomposition $\mathbf{P}$, we have that

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\epsilon}$$

Applying our definitions above, we have:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

So linearity is satisfied.

■

[a] The full rank assumption is critical here.

Claim: *The transformed model satisfies the strict exogeneity assumption given that the original model satisfies assumptions* 2 *and* 4*b*.

Proof. Again, we have that $\mathbf{V}$ has a full rank square root decomposition, so now observe that

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\epsilon}}|\tilde{\mathbf{X}}] &= \mathbb{E}[\mathbf{P}\boldsymbol{\epsilon}|\mathbf{X}] \\ &= \mathbf{P}\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] \\ &= \mathbf{P}\mathbf{0} \\ &= \mathbf{0}\end{aligned}$$

■

Claim: *The transformed model satisfies the rank condition given that the original model satisfies assumptions* 3 *and* 4*b*.

Proof. Observe that assumption 4*b*. guarantees that $rank(\mathbf{P}) = n$. Under assumption 3 on the original model $rank(\mathbf{X}) = k$. By the invertible matrix theorem, $\mathbf{P}$ is invertible, which implies that

$$rank(\mathbf{P}\mathbf{X}) = \min\{rank(\mathbf{P}), rank(\mathbf{X})\} = \min\{n, k\} = k$$

[a]

■

[a] Note that you need to know how to prove that for arbitrary conformable matrices $\mathbf{A}$ and $\mathbf{B}$,

$$rank(\mathbf{A}\mathbf{B}) \leq \min\{rank(\mathbf{A}), rank(\mathbf{B})\}$$

Claim: *The transformed model has spherical errors.*

Proof.

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}'|\tilde{\mathbf{X}}] &= \mathbb{E}[\tilde{\boldsymbol{\epsilon}}\tilde{\boldsymbol{\epsilon}}'|\mathbf{X}] \\
&= \mathbb{E}[\mathbf{P}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{P}'|\mathbf{X}] \\
&= \mathbf{P}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}]\mathbf{P}' \\
&= \mathbf{P}\sigma^2\mathbf{V}\mathbf{P}' \\
&= \sigma^2\mathbf{P}\mathbf{V}\mathbf{P}' \\
&= \sigma^2\mathbf{P}(\mathbf{V}^{-1})^{-1}\mathbf{P}' \\
&= \sigma^2\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}' \\
&= \sigma^2\mathbf{P}\mathbf{P}^{-1}\mathbf{P}'^{-1}\mathbf{P}' \\
&= \sigma^2\mathbf{I}_n
\end{aligned}$$

■

Therefore we have shown that the transformed model satisfies the assumptions of the Gauss-Markov theorem,[25] and hence the traditional "OLS" estimator is optimal in the sense that it is BLUE.

What is the optimal estimator in this case then?

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\
&= ((\mathbf{P}\mathbf{X})'\mathbf{P}\mathbf{X})^{-1}(\mathbf{P}\mathbf{X})\mathbf{P}\mathbf{y} \\
&= (\mathbf{X}\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y} \\
&= (\mathbf{X}\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{y} \\
&= (\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}
\end{aligned}$$

Notice now what happens if we restore assumption 4. Then $\mathbf{V}^{-1}$ is just $\mathbf{I}_n$ and we recover the traditional OLS estimator. It is in this sense that we have generalized the least squares estimator. Similarly to how we derived the GLS estimator, we can also derive the finite sample properties of the GLS estimator, but I leave this as an exercise to the reader.

5.3 Back to Weighted Least Squares

With that theory in hand, we can now observe that the proper weighting matrix is this one:

$$\mathbf{W} = \begin{bmatrix} \sqrt{N_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{N_2} & 0 & \dots & 0 \\ 0 & 0 & \sqrt{N_3} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sqrt{N_{2954}} \end{bmatrix}$$

Mechanically then, we simply regress $\mathbf{W}\mathbf{y}$ on $\mathbf{W}\mathbf{X}$, and then we recover the best estimate for $\boldsymbol{\beta}$.

[25] Sometimes people will refer to the proof that GLS is BLUE in this setting as Aitken's theorem, but it is really just Gauss-Markov.

5.4 A Word of Caution on GLS

Note that the fact that we know the variance covariance matrix $\mathbf{V}(\mathbf{X})$ is doing a lot of work here. Suppose that if instead of knowing the variance covariance matrix, we were trying to estimate it. This approach is called feasible generalized least squares, and it is not commonly used.

To see why, observe that we need to estimate the variance-covariance matrix in its entirety, which means estimating $cov(\epsilon_i, \epsilon_j)\ \forall\ i, j \in \{1, 2, \ldots, n\}$. Exploiting the symmetry of variance-covariance matrices, we then have to estimate

$$\sum_{i=1}^{n}[n-(i-1)] = \frac{1}{2}n(n+1)$$

parameters.

This is, to put it mildly, a herculean task. The historical way that this was done is that one would make use of a consistent[26] estimator. One would then use the implied residuals to generate some consistent estimate of the variance-covariance matrix, which you then use to perform GLS.

OLS also has some nice large sample properties that GLS does not, which I will not discuss but is discussed in Hayashi if interested.

In practice, what one does is use standard errors that have been modified to account for the violation of the no heteroskedasticity or autocorrelation assumption. For heteroskedasticity, these are **White Standard Errors**, and for autocorrelation, these are primarily **Newey-West Standard Errors** or some improvement of them (e.g. the standard errors proposed by Andrews (1991)).

6 Metric Spaces, Big Oh, and Little Oh

7 The Typewriter Sequence

Our first task is to prove that for each n, there exists exactly one positive integer k_n satisfying the following condition:

$$\frac{(k_n-1)k_n}{2} < n \leq \frac{k_n(k_n+1)}{2}$$

A natural impulse here is to do induction, but it is not clear what the inductive set would be. Rather, we know that we want a minimal integer such that the sum is greater than or equal to n. The fact that we know we want a *minimal* one is indicative that we want to make use of the well-ordering of the positive integers. Once you realize this, the proof becomes very obvious.

Let S be defined by $\{k \in \mathbb{N} | n - \sum_{i=1}^{k} i \leq 0\}$.

[26] We have not discussed this, but we say that an estimator $\hat{\theta}$ is consistent for θ if $\operatorname{plim}\hat{\theta} = \theta$. What a plim exactly is is something that we will discuss later on.

Claim: *S is nonempty.*

Proof. Suppose that S is empty. Then it must be that case that $\forall k \in \mathbb{N}$ we have that

$$n > \sum_{i=1}^{k} i$$

But then n is clearly a maximal integer, which is absurd. Hence it must be that S is nonempty. ■

Now, by the Well-Ordering Principle, it must be that S has a least element, call it k_n. Since $k_n \in S$, it must be that $n \leq \frac{k_n(k_n+1)}{2}$

Claim: $\frac{(k_n-1)k_n}{2} < n$

Proof. Suppose not, so that $\frac{(k_n-1)k_n}{2} \geq n$. Then immediately it follows that $k_{n-1} \in S$. But notice that k_n is the least element of S by the WOP. Contradiction, so we have that $\frac{(k_n-1)k_n}{2} < n$. ■

What remains is to show that there is only one k_n that satisfies both inequalities at once. To that end suppose that $k_n + 1$ satisfied both inequalities, then it must be that $\frac{k_n(k_n+1)}{2} < n$, but then $k_n \notin S$, which is a contradiction. Obviously then, if $k_n + 1$ does not satisfy both, no strictly larger k will either.

The second part of this question asks you to show that $j_n \equiv$

8 Ergodicity and Time Series

It is common when dealing with macroeconomics to consider sequences of random variables over time - the level of GDP in a country, the inflation rate, and so on. Suppose that you believe that the evolution of these variables is derived according to some underlying data-generating process. The question becomes, can we estimate the parameters of this process? The answer to this question is not simple, because unlike in, say, the panel data setting, we do not observe a plethora of possible realizations of the data-generating process, and associated covariates. Rather, when considering time series of random variables, we observe only one path. We would like to use the one path that we see to make statements about the underlying parameters, but that is in general problematic. The ergodic theorem, and ergodic theory more generally, is a result that allows us to make make progress on this problem. These notes draw heavily from chapter 3 of Hayashi. I have tried to collect concepts in one place with theorems, and provide a bit more intuition as to what ergodicity is doing, but the mathematical content is from Hayashi.

8.1 Preliminaries

Let $\{X_t\}_{t \in T}$ be some collection of random variables from some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ to some space S (typically for our purposes the reals). The set T is called the index set, and we typically associate it with say, the natural numbers, or the integers, and say that it represents time. The

set S is called the **state space**. The entire collection $\{X_t\}_{t\in T}$ is called a **stochastic process**. A sequence of realizations of $\{X_t\}_{t\in T}$ is called a **path**.[27]

For the time being, let us restrict our consideration to the time interpretation of stochastic processes. There are two different but related concepts that we often want to think about in this context. The first is the concept of strict stationarity, and the second is the concept of covariance stationarity.

We say that a process $\{X_t\}$ is **(strictly) stationary** if $\forall t \in T$, the joint distribution of $(X_t, X_{t-1}, X_{t-2}, \dots)$ does not depend on t. In other words, say (X_1, X_5) must have the same joint distribution as (X_5, X_9) and so on. In a strictly stationary sequence, what matters is the relative difference between observations, not where they are in time absolutely. Perhaps the most famous eXample of a stationary process is a sequence of independent and identically distributed draws.[28]

In general, applying a $\mathcal{F}$-measurable function to a stationary process will result in another stationary process.

We say that a process is **covariance-stationary** if

1. $\mathbb{E}[X_t]$ does not depend on t
2. $Cov(X_t, X_{t-j})$ exists, is finite, and depends only on j and not on t

The j-th order autocovariance, which we write as $\mathbf{\Gamma}_j$, is defined as

$$\mathbf{\Gamma}_j \equiv Cov(\mathbf{X}_t, \mathbf{X}_{t-j})$$

[29]

As an example, consider the covariance stationary bivariate case, where $\mathbf{X}_t = [X_{1,t}|X_{2,t}]$, where $X_{i,t}$ is mean zero $\forall\, i\, t$. In this case, we have that the zero-order autocovariance is given by:

$$\mathbb{E}\begin{bmatrix} X_{1,t}^2 & X_{1,t}X_{2,t} \\ X_{2,t}X_{1,t} & X_{2,t}^2 \end{bmatrix}$$

The first order autocovariance is given by:

$$\mathbb{E}\begin{bmatrix} X_{1,t}X_{1,t-1} & X_{1,t}X_{2,t-1} \\ X_{2,t}X_{1,t-1} & X_{2,t}X_{2,t-1} \end{bmatrix}$$

One should also see that if $\{X_t\}_{t\in T}$ is covariance stationary, then

$$\mathbf{\Gamma}_j = \mathbf{\Gamma}'_{-j}$$

Note that strict stationary processes are *not* necessarily covariance stationary. A strict stationary process is covariance stationary if it also has finite variances and covariances.

[27] Sometimes you will see it referred to as a sample function.

[28] To see this observe that the marginal distribution of each individual innovation is an identical distribution, call it $F_X(\cdot)$. Since the draws are independent, the joint distribution is the product of the marginals, and the marginals are identical.=

[29] Note that this is a covariance matrix, as $\mathbf{X}_t$ may be a vector random variable.

LLN for Covariance Stationary Processes: *If $\{X_t\}$ is a covariance stationary process with $\mathbb{E}[X_t] = \mu$, $Cov(X_t, X_{t-j}) = r_j$, and $\lim_{j\to\infty} r_j = 0$ then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to_{M.S.} \mu$$

In other words, given a covariance stationary process where the past covariances are sufficiently weak, we have that the time average is equal to the population mean.

8.2 Unit Roots

A **unit root** process can be written in the following recursive manner:

1. Let $u_t \sim iid(0, \sigma^2)$
2. $\epsilon_0 = u_0$
3. $\epsilon_t = \epsilon_{t-1} + u_t$

One should immediately see two things. The first is that $\mathbb{E}[\epsilon_t] = 0$, as ϵ_t is the sum of t mean zero independent random-variables. The second is that $V(\epsilon_t) = \sum_{i=1}^{t} \sigma^2 = t\sigma^2$. As $t \to \infty$, this implies that the variance of this ϵ_t goes off to infinity, so this process is not covariance stationary (recall that we required a process to have finite variances and covariances to be covariance stationary).

We say that a process $\{X_t\}_{t\in T}$ is **difference stationary** if, when we take first differences, the resulting sequence is strictly stationary. To put it more formally, consider $\{X_t\}_{t\in\mathbb{Z}}$. This sequence is difference stationary if $\{X_t - X_{t-1}\}_{t\in\mathbb{Z}}$ is strictly stationary.

This is important because some people like to argue that certain aggregates like GDP have unit roots, and thus we first difference them and consider growth rates.

8.3 Ergodicity

What does ergodicity really mean? The mathematical definition of ergodicity is given in terms of measure theory: let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure-space, and let T be an $(\Omega, \mathcal{F}) - (\Omega, \mathcal{F})$ measurable morphism (i.e. a measurable self-map).[30] Suppose that T has the property that for all $E \in \mathcal{F}$ such that $T^{-1}(E) = E$, we have either that $\mathbb{P}(E) = 0$ or $\mathbb{P}(E) = 1$. Then we say that the collection $(\Omega, \mathcal{F}, \mathbb{P}, T)$ is an ergodic system. The perhaps more intuitive way (but less precise) way to see ergodicity is to imagine a state space. Drop a particle at some point in the space, and let it move throughout the state space according to some process. An ergodic set is one in which, if the ball gets into that set, it a)never leaves that set again, and b)traverses the whole set over and over again an infinite number of times as t goes to infinity.

Another way to imagine ergodicity is to imagine that you flip a coin an infinite number of times, with the aim of determining whether or not the coin is fair. Suppose that you could flip the coin an infinite number of times all at once somehow.[31] If you could do this, you would know,

[30] Alternatively, one could write that we require that $T : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple isomorphism.

[31] Or you could observe the infinite number of possible realizations of the outcome of the coin flip in all the other parallel universes if you believe in that sort of thing.

precisely what the probability of the outcome of the coin flip being heads would be. Ergodicity in some sense says that rather than observing all of those infinite coinflips simultaneously, we can observe a sequence of coin flips every second say, and use that to deduce whether or not the coin is fair.

With this kind of loose idea of what ergodicity means, we can now introduce the ergodic theorem.

The Ergodic Theorem: *Suppose that $\{X_t\}_{t\in T}$ is a strictly stationary process. Then the following are equivalent:*

1. *$\{X_t\}_{t\in T}$ is ergodic.*
2. *For every k and every function ϕ of $k+1$ variables:*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\phi(X_t,\ldots,X_{t+k})=\mathbb{E}[\phi(X_t,\ldots,X_{t+k})]$$

if the expectation on the right hand side exists.

One obvious application of the ergodic theorem is when $k=0$, and $\phi(X_t)=X_t$. Then the theorem simply says that if the process is strictly stationary and ergodic, then

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}X_t=\mathbb{E}[X_t]$$

In other words, given an ergodic process and a sufficiently long path, we can recover the true expectation of the underlying process. To put it in the context of the earlier analogy, given an infinite period of time, the way the ball bounces around the ergodic set tells us everything we could want to know about the set itself. We also have a law of large numbers for ergodic sequences of random variables.

Claim: *Suppose that $\{X_t\}$ is a stationary and ergodic sequence of random variables having finite mean. Then*

$$\bar{X}_n \to_{a.s.} \mu$$

8.4 Martingales

Imagine a game where if you win, you get back twice the amount that you bet, where if you lose, you lose everything you bet and that the odds of winning in this game are such that with $p=0.5$ you win, and with symmetric probability you lose. You can play this game as many times as you want, and the probability of you winning does not depend on the past history of your wins and losses (i.e. this is a sequence of independent and identically distributed draws).

If one does not have a liquidity constraint[32], then one can engage in the following betting strategy: if you win, you leave the game. If you lose, you play again, but double your bet. The idea behind this betting strategy is that you will eventually recover all your losses plus a win

[32] Or a time constraint.

however much you bet eventually (since the probability of a n fair coin flips all coming up heads goes to zero as $n \to \infty$). This betting strategy is called a martingale.

In the context of probability theory, we say $\{X_t\}$ is a **martingale** if the following condition holds:

$$\mathbb{E}[X_{t+1}|X_t, X_{t-1}, \dots] = X_t \forall t \in T$$

To put it in words, a martingale is a process where the best guess of what happens tomorrow is what is happening today. Nothing further in the past tells you anything about how the sequence is going to evolve.

Similarly to how we defined martingales, we can also define submartingales and supermartingales. We say that a sequence $\{X_t\}$ is a **supermartingale** if

$$\mathbb{E}[X_{t+1}|X_t, X_{t-1}, \dots] \leq X_t$$

Similarly, a sequence is a **submartingale** if

$$\mathbb{E}[X_{t+1}|X_t, X_{t-1}, \dots] \geq X_t$$

Perhaps the most famous example of martingales is the so-called **random walk**. A random walk is a stocastic process $\{X_n\}$ where $X_n = \sum_{i=1}^{n} g_i$, and $g_i \sim iid(0, \sigma^2)$.[33]

Claim: *A random walk is a martingale.*

Proof.

$$\begin{aligned}
\mathbb{E}[X_{t+1}|X_t, X_{t-1}, \dots] &= \mathbb{E}[X_t + g_{t+1}|X_t, X_{t-1}, \dots] \\
&= \mathbb{E}[X_t|X_t, X_{t-1}, \dots] + \mathbb{E}[g_{t+1}|X_t, X_{t-1}, \dots] \\
&= X_t + \mathbb{E}[g_{t+1}|X_t, X_{t-1}, \dots] \\
&= X_t
\end{aligned}$$

The first equality follows from the definition of a random walk. The second equality follows from the linearity of conditional expectation. The third inequality follows from the fact that X_t is in the conditioned information set. The fourth equality follows from the fact that g_{t+1}

Another important class of stochastic processes is what are called martingale difference sequences. A **martingale difference sequence** is a stochastic process $\{g_t\}$ such that for all t, we have that $\mathbb{E}[g_{t+1}|g_t, g_{t-1}, \dots] = 0$[34].

To see why it is called a martingale difference sequence, observe that if you sum the shocks over time, you generate a martingale itself. Martingale difference sequences have the important property that $Cov(g_t, g_{t-j}) = 0 \ \forall t \ \forall j \neq 0$. That is to say that an MDS exhibits no serial correlation.

We also have an important central limit theorem for martingale difference sequences.

[33] You sometimes see this g_i called an independent white noise term.

[34] If you have a finite time stochastic process, so that T is say $\{1, 2, \dots, n\}$, the condition must hold for all $t \geq 2$.

Claim: *Let* $\{\mathbf{g}_i\}$ *be a vector martingale difference sequence that is stationary and ergodic with* $\mathbb{E}[\mathbf{g}_i\mathbf{g}_i'] = \mathbf{\Sigma}$*, and define* $\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i$*. Then*

$$\sqrt{n}\bar{\mathbf{g}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}_i \to_d \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

9 Large Sample Properties of OLS

Again, this is mostly (almost entirely really) a regurgitation of Hayashi.

Recall that when dealing with finite samples, we had a set of assumptions made for OLS. These were:

1. (Model Linearity): $y_i = \mathbf{x}_i'\beta + \epsilon_i$
2. (Strict Exogeneity): $\mathbb{E}[\epsilon_i|\mathbf{X}] = 0$
3. (Full Rank): $\mathbb{P}(rank(\mathbb{X}) = k) = 1$
4. (Independence and no Heteroskedasticity or Autocorrelation): $\mathbb{E}[\epsilon\epsilon'|\mathbf{X}] = \sigma^2\mathbf{I}_n$
5. (Normal Errors): The distribution of $\boldsymbol{\epsilon}$ conditional on $\mathbf{X}$ is jointly normal.

When we move from the world of finite samples to the world of large samples, we are no longer concerned with making statements about the joint distribution of $(y, \mathbf{X})$, but rather we are concerned with making statements about the stochastic process that determines the joint distribution of our finite sample. If we can say something about this stochastic process (what we call the **data generating process**), then we can by construction say something about the joint distribution of our sample.[35]

However, because we were concerned before with making statements about joint distributions, and now we want to make statements about stochastic processes more generally, we need to reformulate the OLS assumptions. The large-sample OLS assumptions are given by:

1. (Linearity):
$$y_i = \mathbf{x}_i'\beta + \epsilon_i \; \forall i \in \{1, 2, \ldots, n\}$$
where $\mathbf{x}_i$ and β are K-dimensional vectors of regressors and coefficients respectively, ϵ_i is the unobserved error, and y_i is the observed outcome.
2. (Ergodic Stationarity): The $(K+1)$ dimensional vector stochastic process $\{y_i, \boldsymbol{x}_i\}$ is jointly stationary and ergodic.
3. (Predetermined regressors): All regressors are orthogonal to the contemporaneous error term, so that $\mathbb{E}[x_{ik}\epsilon_i] = 0 \; \forall k \in \{1, 2, \ldots, K\}$. Let $\mathbf{g}_i \equiv \mathbf{x}_i \cdot \epsilon_i$. This means that we can write the above in matrix notation as
$$\mathbb{E}[\mathbf{g}_i] = \mathbf{0}$$
4. (Rank condition): The $K \times K$ matrix $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i']$ is non-singular. Denote this matrix by $\mathbf{\Sigma}_{\boldsymbol{xx}}$

[35] At least when the process is stationary.

5. ($\mathbf{g}_i$ is a MDS with finite second moments): $\{\mathbf{g}_i\}$ is a martingale difference sequence. The matrix of cross moments, $\mathbb{E}[\mathbf{g}_i\mathbf{g}_i']$ is nonsingular. Writing the asymptotic variance of $\sqrt{n}\bar{\mathbf{g}}$ as $\mathbf{S}$, we have by assumption 2.2 and the ergodic stationary martingale differences CLT that

$$\mathbf{S} = \mathbb{E}[\mathbf{g}_i\mathbf{g}_i']$$

There are a couple of things to note. The first is that although we often think of stochastic processes in the context of time series data, processes need only be defined over an index set. This helps I think make it more clear that we are not restricting ourselves to time series data in this formulation, because a random sample, i.e. a sequence of *i.i.d.* draws is an ergodic stationary process.

The second thing to note is that assumption three is strictly weaker than $\mathbb{E}[\epsilon_i|\mathbf{x}_i] = 0$. To see this note that:

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_i)\epsilon_i] &= \mathbb{E}[\mathbb{E}[f(\mathbf{x}_i)\epsilon_i|\mathbf{x}_i]] \\ &= \mathbb{E}[f(\mathbf{x}_i)\mathbb{E}[\epsilon_i|\mathbf{x}_i]] \\ &= \mathbb{E}[f(\mathbf{x}_i)0] \\ &= 0\end{aligned}$$

Predetermined regressors is also in general a weaker assumption than strict exogeneity, because strict exogeneity in this context would imply not only that the error term is orthogonal to the current regressors, but also that it is orthogonal to all future regressors.[36] In the case of an $AR(1)$ process for example, this is not going to hold, so it is good that we don't require it.

9.1 OLS Asymptotic Distribution

To characterize the asymptotic distribution of the OLS estimator, we are going to show three things:

1. Under assumptions 1-4, we have that $\hat{\beta}$ is consistent for β
2. Under assumptions 1,2,4, and 5, we have that

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}(0, Avar(\hat{\beta})) \text{ as } n \to \infty$$

   where

$$Avar(\hat{\beta}) = \mathbf{\Sigma}_{xx}^{-1}\mathbf{S}\mathbf{\Sigma}_{xx}^{-1}$$

3. Suppose that there is a consistent estimator $\hat{\mathbf{S}}$ of $\mathbf{S}$.[37] Under assumption 2 then, $Avar(\hat{\beta})$ is consistently estimated by

$$\widehat{Avar(\hat{\beta})} = \mathbf{S}_{xx}^{-1}\hat{\mathbf{S}}\mathbf{S}_{xx}^{-1}$$

   Where $\mathbf{S}_{xx} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X}$

[36] Or, if you like, it would imply an individual's error term is orthogonal to the regressors of all other individuals.
[37] $\mathbf{S} = \mathbb{E}[\mathbf{g}_i\mathbf{g}_i']$

Before doing anything else, we rewrite the distance between $\hat{\beta}$ and β in terms of sample means.

$$\begin{aligned}
\hat{\beta} - \beta &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y - \beta \\
&= \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \boldsymbol{\epsilon}) - \beta \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} - \beta \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\
&= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \cdot \epsilon_i\right) \\
&= \mathbf{S}_{xx}^{-1}\bar{\mathbf{g}}
\end{aligned}$$

Claim: *Under assumptions 1 to 4, we have that $\hat{\beta}$ is consistent for β.*

Proof. By assumption 2, we have that $\{y, \mathbf{x}_i\}$ is jointly stationary and ergodic, we have that $\{\mathbf{x}_i\}$ is stationary and ergodic. Since measurable functions of stationary and ergodic sets are stationary and ergodic, we thus have that $\{\mathbf{x}_i\mathbf{x}_i'\}$ is stationary and ergodic. This then implies that $\operatorname{plim}\mathbf{S}_{xx} = \boldsymbol{\Sigma}_{xx}$. Assumption 4 delivers us that $\boldsymbol{\Sigma}_{xx}$ is invertible, so we have that

$$\operatorname{plim}\mathbf{S}_{xx}^{-1} = \boldsymbol{\Sigma}_{xx}^{-1}$$

By the continuous mapping theorem. Similarly we know that $\operatorname{plim}\bar{\mathbf{g}} = \mathbf{E}[\mathbf{g}_i] = \mathbf{0}$ by assumption 3. Again applying the continuous mapping theorem, we have that

$$\operatorname{plim}\mathbf{S}_{xx}^{-1}\bar{\mathbf{g}} = \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{0} = \mathbf{0}$$

Since this quantity is equal to $\hat{\beta} - \beta$, we have that $\operatorname{plim}\hat{\beta} = \beta$, which was what was wanted. ■

Claim: *Under assumptions 1,2,4, and 5, we have that*

$$\sqrt{n}(\hat{\beta} - \beta) \to_d \mathcal{N}(0, Avar(\hat{\beta})) \text{ as } n \to \infty$$

where

$$Avar(\hat{\beta}) = \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}$$

Proof. First observe that $\sqrt{n}(\hat{\beta} - \beta) = \mathbf{S}_{xx}^{-1}(\sqrt{n}\bar{\mathbf{g}})$. By assumption 5, we have that $\sqrt{n}\bar{\mathbf{g}} \to_D \mathcal{N}(0, \mathbf{S})$. Since $\mathbf{S}_{xx}^{-1}$ is conformable with $\sqrt{n}\bar{\mathbf{g}}$, using properties of convergence in distribution, we have that $\sqrt{n}(\hat{\beta} - \beta) \to_D \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}(\boldsymbol{\Sigma}_{xx}^{-1})')$.[a] Then, since $\boldsymbol{\Sigma}_{xx}^{-1}$ is symmetric, we recover that:

$$Avar(\hat{\beta}) = \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}$$

Which was what was wanted. ■

[a] Since $\mathbf{S}_{xx}^{-1}$ converges to $\boldsymbol{\Sigma}_{xx}^{-1}$.

The last thing we need to need to show is that given some consistent estimator $\hat{\mathbf{S}}$ for $\mathbf{S}$, we have that $\mathbf{S}_{xx}^{-1}\hat{\mathbf{S}}\mathbf{S}_{xx}^{-1}$ is consistent for $\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}$. This is fairly trivial: observe that $\operatorname{plim}\mathbf{S}_{xx}^{-1}\hat{\mathbf{S}}\mathbf{S}_{xx}^{-1} = (\operatorname{plim}\mathbf{S}_{xx}^{-1})(\operatorname{plim}\hat{\mathbf{S}})(\operatorname{plim}\mathbf{S}_{xx}^{-1})$. Invoking ergodic stationarity delivers that $\operatorname{plim}\mathbf{S}_{xx}^{-1} = \boldsymbol{\Sigma}_{xx}^{-1}$, which, combined with the assumption that $\hat{\mathbf{S}}$ is consistent for $\mathbf{S}$, tells us that the whole plim is in fact $\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}$, which was what was wanted.

The key wrinkle is of course finding that consistent estimator $\hat{\mathbf{S}}$, which is a non-trivial task. This will be discussed later on, but for now we will prove one more result about the estimating $\mathbf{E}[\epsilon_i^2]$.

Claim: *Let e_i be the OLS residual. Under assumptions 1 through 4, we have that*

$$s^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n} e_i^2 \to_p \mathbb{E}[\epsilon_i^2]$$

so long as the expectation exists and is finite.

Proof. The proof is almost entirely algebraic manipulation, starting from the observation that $e_i = \epsilon_i - \mathbf{x}_i'(\hat{\beta} - \beta)$. This then implies $e_i^2 = \epsilon_i^2 - 2(\hat{\beta} - \beta)'\mathbf{x}_i\epsilon_i + (\hat{\beta} - \beta)'\mathbf{x}_i\mathbf{x}_i'(\hat{\beta} - \beta)$. Summing over i, then we get that

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - 2(\hat{\beta} - \beta)'\bar{\mathbf{g}} + (\hat{\beta} - \beta)'\mathbf{S}_{xx}(\hat{\beta} - \beta)$$

Taking plims on the left and right hand side, and then distributing them, we need only consider

1. $\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$
2. $\operatorname{plim} 2(\hat{\beta} - \beta)'\bar{\mathbf{g}}$
3. $\operatorname{plim}(\hat{\beta} - \beta)'\mathbf{S}_{xx}(\hat{\beta} - \beta)$

By ergodic stationarity 1. is equal to $\mathbb{E}[\epsilon_i^2]$, so all that remains is to show that 2. and 3. are equal to zero.
To see that 2. is equal to 0, observe that

$$\operatorname{plim}(\hat{\beta} - \beta)'\bar{\mathbf{g}} = (\operatorname{plim}(\hat{\beta} - \beta))'(\operatorname{plim}\bar{\mathbf{g}})$$

Both of these terms are equal to $\mathbf{0}$, as $\hat{\beta}$ is consistent for β, and by assumption 3, $\mathbb{E}[\mathbf{g}_i] = \mathbf{0}$, so obviously this term is equal to zero.
To see that 3. is equal to 0, again we apply properties of plims to separate the product into $(\operatorname{plim}\hat{\beta} - \beta)'(\operatorname{plim}\mathbf{S}_{xx})(\operatorname{plim}\hat{\beta} - \beta)$. The first and third are equal to zero by the consistency of $\hat{\beta}$ for β. The second probability limit is equal to $\mathbf{\Sigma}_{xx}$, which is finite by assumption 4. The two combined deliver that the product is equal to 0.
Hence we have that

$$\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \operatorname{plim}\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 = \mathbb{E}[\epsilon_i^2]$$

Which was what was wanted. ■

9.2 The Problem of S

If you want to test regression coefficients for significance in the traditional way, you need a t-statistic. In this world, the t-ratio we use, which we call the robust t-ratio, is given by

$$t_k = \frac{\sqrt{n}(\hat{\beta}_k - \bar{\beta}_k)}{\sqrt{\widehat{Avar}(\hat{\beta}_k)}} = \frac{\hat{\beta}_k - \bar{\beta}_k}{SE^*(\hat{\beta}_k)}$$

Where $\bar{\beta}_k$ is the assumed truth, and

$$SE^*(\hat{\beta}_k) = \sqrt{\frac{1}{n}(\mathbf{S}_{xx}^{-1}\hat{S}\mathbf{S}_{xx}^{-1})_{kk}}$$

This standard error, which is different from the standard error in Chapter 1,[38] is called **White's standard error** or **heteroskedasticity-consistent standard error**, where the latter name is due to the fact that we have not assumed that the error term is conditionally homoskedastic.

The proof that this statistic is distributed asympotically unit normal is covered in detail in Hayashi, so please read pages 118-120.

Again recall that all of these results are based on the presupposition that there exists a consistent estimator $\hat{\mathbf{S}}$ for $\mathbf{S}$.[39] If no such estimator exists, then we are in a lot of trouble. Given the discussion above, a natural idea is to use the residuals from our OLS regression, since we know that a properly adjusted residual sum of squares will converge in probability to $\mathbb{E}[\epsilon_i^2]$.

So let us proceed with that in our mind. For ease of exposition, consider only the simple one variable case. As with most things in Hayashi, the generalization to multivariate regression conveys little more intuition and is massively more painful algebraically.

We know from the above that

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - 2(\hat{\beta}-\beta)'\bar{\mathbf{g}} + (\hat{\beta}-\beta)'\mathbf{S}_{xx}(\hat{\beta}-\beta)$$

Substituting back out for $\bar{\mathbf{g}}$ and $\mathbf{S}_{xx}$,we can rewrite this as:

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 - 2(\hat{\beta}-\beta)\frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i + (\hat{\beta}-\beta)^2\frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Recalling that we are interested in $\mathbb{E}[\epsilon_i^2 x_i^2]$,[40] we now multiply both sides by x_i^2, to recover that

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i^2 = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 x_i^2 - 2(\hat{\beta}-\beta)\frac{1}{n}\sum_{i=1}^{n} x_i^3\epsilon_i + (\hat{\beta}-\beta)^2\frac{1}{n}\sum_{i=1}^{n} x_i^4$$

But here we have a problem, because we do not know that the latter two terms on the right hand side are finite. Indeed we have assumed only finite second moments so far. Here we are faced with something that looks like a fourth moment. So what do we do, we assume what we need to make the result go through.

This assumption is as follows:

6. $\mathbb{E}[(x_{ik}x_{ij})^2]$ exists and is finite for all $k, j \in \{1, 2, \ldots, K\}$.

As a result of this equation, we have that $(\hat{\beta}-\beta)^2\frac{1}{n}\sum_{i=1}^{n} x_i^4 \to_p 0$, since by ergodic stationarity, we have that $\frac{1}{n}\sum_{i=1}^{n} x_i^4 \to E[x_i^4]$, which we have just assumed to be finite, and $\hat{\beta}$ is consistent for β. To complete the proof, now we need only that $2(\hat{\beta}-\beta)\frac{1}{n}\sum_{i=1}^{n} x_i^3\epsilon_i$ goes away in the limit. So let us prove that now:

[38] The standard error of OLS in the finite sample world was given by $\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}}$, where $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$. $\mathbf{X}'\mathbf{X}$

[39] Again, recall that $\mathbf{S} = \mathbb{E}[\mathbf{g}_i\mathbf{g}_i'] = \mathbb{E}[\epsilon_i^2\mathbf{x}_i\mathbf{x}_i']$.

[40] In this environment, $\mathbf{x}_i\mathbf{x}_i' = x_i^2$, since we have only one regressor.

Claim: $2(\hat{\beta}-\beta)\frac{1}{n}\sum_{i=1}^{n}x_i^3\epsilon_i \to_p 0$

Proof. We immediately know that $\hat{\beta}$ is consistent for β, which implies that the only thing we really need to show is that $\frac{1}{n}\sum_{i=1}^{n}x_i^3\epsilon_i$ converges in probability to something finite. To that end, observe that Cauchy-Schwartz implies that

$$\mathbb{E}[|x_i^3\epsilon_i|] \leq \sqrt{\mathbb{E}[x_i^2\epsilon_i^2]\mathbb{E}[x_i^4]}$$

Since we know that both terms under the square root are finite (one by assumption 5 and other by equation 6), we also know that $\mathbb{E}[|x_i^3\epsilon_i|]$ is finite, and thus that $\mathbb{E}[x_i^3\epsilon_i]$ is also finite. Now we need only invoke ergodic stationarity so that

$$\frac{1}{n}\sum_{i=1}^{n}x_i^3\epsilon_i \to_p \mathbb{E}[x_i^3\epsilon_i]$$

To get that

$$2(\hat{\beta}-\beta)\frac{1}{n}\sum_{i=1}^{n}x_i^3\epsilon_i \to_p 0$$

Which was what was wanted. ■

That these terms go to zero in the plim then delivers us that

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}e_i^2x_i^2 = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2x_i^2 = \mathbf{S}$$

So our candidate estimator for $\hat{S}$, $\frac{1}{n}\sum_{i=1}^{n}e_i^2x_i^2$ is consistent for $\mathbf{S}$.
In general, we use a degree of freedom adjustment and estimate $\mathbf{S}$ with

$$\hat{\mathbf{S}} = \frac{1}{n-K}\sum_{i=1}^{n}\epsilon_i^2\mathbf{x}_i\mathbf{x}_i'$$

10 Wold Representations

In general, when we think about time series data, the first thing we think of is autoregressive proceses of order p ($AR(p)$) and moving average processes of order q ($MA(q)$).

One particular $AR(p)$ process has the following form:

$$x_t = \phi_1x_{t-1} + \phi_2x_{t-2} + \cdots + \phi_px_{t-p} + \epsilon_t$$

Where $\epsilon_t \sim WN(0,\sigma^2)$ (ϵ_t is an independent white noise process), and $\mathbb{E}[x_t] = \mathbb{E}[x_s] = 0 \forall\ s,t$.
If we me relax the mean zero assumption $\mathbb{E}[x_t] = \mathbb{E}[x_s] = \mu\ \forall s,t$, we can rewrite the above as

$$x_t - \mu = \phi_1(x_{t-1}-\mu) + \cdots + \phi_p(x_{t-p}-\mu) + \epsilon_t$$

Doing a little algebra, we see that

$$x_t = \delta + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \epsilon_t$$

Where $\delta = \mu(1 - \sum_{i=1}^{p} \phi_i)$

This is the familiar form of the AR process that most of you have probably seen before.

Moving average processes, on the other hand take the form

$$\epsilon_t = \mu + \nu_t + \theta_1 \nu_{t-1} + \cdots + \theta_q \nu_{t-q}$$

Where $\nu \sim WN(0, \sigma^2)$

10.1 AR(1) Representation

The two are related to each other by a very important result called **Wold's Representation Theorem**. Loosely, Wold's theorem says that every covariance stationary time series is the sum of a purely deterministic time series and a purely stochastic one. In lieu of a proof, which you can find online if interested, I will simply show the how to build a Wold decomposition of an AR(1).

Let $\eta \sim iid(0, \sigma_\eta^2)$

Now $x_0 = \eta_0$,$x_1 = \rho\eta_0 + \eta_1$,$\rho x_2 = \rho^2\eta_0 + \rho\eta_1\eta_2$, and so on.

One should see that, if we start indexing from 0, that $x_T = \sum_{i=0}^{T} \rho^{T-i}\eta_i$

Taking the limit as $T \to \infty$ delivers that

$$x_t = \sum_{s=0}^{\infty} \rho^s \eta_{t-s}$$

This is called a Wold Decomposition of an $AR(1)$.

We can use this to derive properties of the $AR(1)$, like variance, etc.

So for example, we immediately see that

$$\mathbb{E}[x_t] = \mathbb{E}[\sum_{s=0}^{\infty} \rho^s \eta_{t-s}] = \sum_{s=0}^{\infty} \rho^s \mathbb{E}[\eta_{t-s}]$$

. This then tells us that the unconditional expectation is equal to zero.

To calculate the variance, we do something similar and note that

$$\mathbb{E}[\sum_{s=0}^{\infty} (\rho^s \eta_{t-s})^2] = \sum_{s=0}^{\infty} \rho^{2s} \mathbb{E}[(\eta_{t-s})^2]$$

[41]

This then simplifies to

$$\sum_{s=0}^{\infty} \rho^{2s}\sigma_\eta^2 = \sigma_\eta^2 \sum_{s=0}^{\infty} \rho^{2s}$$

We then see that $Var(x_t) = \frac{\sigma_\eta^2}{1-\rho^2}$ if this sum converges, and ∞ if it does not.

[41] The cross terms drop because η is a white noise process, so $\eta_t \perp \eta_j \forall j, t$

11 Probability Models

Suppose that you are interested in predicting whether or not something will happen, like say whether or not a mortgage application will be denied. You might think that the decision to accept or reject a mortgage application can only consider some subset of applicant criterias, and you would be right.[42] Now suppose that we are interested in understanding whether or not there are racial differences in the approval or denial of mortgages.

The hypothetical dataset that we have is a set of application characteristics, including the applicant's race, their income, the size of the loan, etc.. We observe as an outcome whether the loan is approved or denied.[43] Suppose that we were to run the following reqression.

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + \nu Race_i + u_i$$

Where y_i is a binary that takes the value 1 if the application was denied, and 0 otherwise, $\mathbf{X}_i$ is a vector of application characteristics, and $Race_i$ is the race of the applicant (for exposition, suppose that this is a binary variable that takes the value 1 if the individual is black, and 0 otherwise). This is a **linear probability model.** Now, if the model is correctly specified, then each coefficient can be interpreted as reflective of an increase in probability if that covariate increases. For example ν reflects the increase in denial probability that a black person faces, holding all other covariates constant. If this coefficient is significantly different from 0, then one could interpret that as evidence of discrimination.[44]

The somewhat obvious problem with linear probability models is that they can predict probabilities greater than 1, which obviously does not comply with our intuitions about how probabilities work. You can take the probability as $\mathbb{P}(\mathbf{X}, Race) = \min\{\max\{\mathbf{X}_i\hat{\boldsymbol{\beta}}+\nu Race, 0\}, 1\}$, but then probabilities are not smooth, which is not something we like. To resolve this problem, we have two kind of canonical models, which are **probit** and **logit** models. The two models are very similar. We will go through the probit model first.

11.1 A Brief Digression

One of the most common ways of estimating parameters is via the method of maximum likelihood estimation. The core insight of maximum likelihood is that, under the supposition that your model is correctly specified, each choice of parameters will generate a probability that you observe the given sample. Your objective then is simply to find the parameter values that maximize the joint probability of observing the sample.

To write this formally, let $\mathbf{y}$ be some data sample, and suppose that we *a priori* think that the distribution comes from some underlying family of distributions $\mathbf{f}(\cdot\,;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^n$. If we assume that each realization is independent and identically distributed, then we can write the joint distribution $\mathbf{f}$ as the product of the (identical) marginals $f(y_i\,;\boldsymbol{\theta})$.

[42] See the Equal Credit Opportunity Act, 15 U.S.C. 1691 *et seq.*, which you can find here. You have the usual restrictions against discrimination on the basis of race, color, religion, national origin, sex, marital status, and age, among certain other criteria.

[43] You can imagine that one might be interested in the closely related question of whether or not the terms that otherwise identical people of different races get for the same loan vary.

[44] The literature refers to this as *differential treatment discrimination*. For an old review see this article from the Urban Institute. For a newer article, see this Federal Reserve Board discussion paper. There is quite a lot of work on this and related topics.

The conceptual leap is here: moving from treating the (observed) sample as a function of the (unobserved) parameters to treating the (unobserved) parameters as a function of the (observed) sample. That is to say that, we define the likelihood of a given observation y_i by

$$\mathcal{L}_i(\boldsymbol{\theta}) = \mathcal{L}_i(\boldsymbol{\theta}\,; y_i) = f(y_i\,; \boldsymbol{\theta})$$

The joint likelihood (which we typically just call the likelihood) is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta})$$

Now recall that monotonic transformations by definition preserve maximizers, so instead of maximizing the likelihood, we can maximize the natural log of the likelihood.

$$l(\boldsymbol{\theta}) = ln \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} ln(\mathcal{L}_i(\boldsymbol{\theta}))$$

I will not derive properties of the MLE, but it will be discussed in the 4th mini, and one of the most famous results in statistics is that asymptotically the MLE attains the Cramer-Rao lower bound, so that Maximum Likelihood Estimators are efficient.[45]

11.2 Probit

The idea behind both a probit and a logit is similar. We are interested in the probability that the outcome is equal to 1.

For the probit, the most common motivation is that of a latent variable. Suppose that $\epsilon \sim \mathcal{N}(0,1)$. Now suppose that we observe that $Y = 1$ if and only if $\mathbf{X}\boldsymbol{\beta} + \epsilon > 0$. We are then interested in $\mathbb{P}(\epsilon > -\mathbf{X}\beta)$. By the symmetry of the normal distribution, we know that this is equal to $\mathbb{P}(\epsilon < \mathbf{X}\beta)$, which since ϵ is standard normal we can simply write as $\Phi(\mathbf{X}\beta)$.

Here is the question, how do we estimate β, if we don't know it. This is not obvious, since Φ is a non-linear function.

The most common approach is to use maximum likelihood estimation. To do maximum likelihood estimation, we first need to think carefully about what the likelihood is going to be. We

[45] This is related to, but distinct from the question of whether MLE is UMVUE - uniformly minimum variance unbiased estimator. If a set of complete sufficient statistics exists, then the MLE is UMVUE. This is a consequence of the Lehman-Scheffe theorem. If, however, such a set of statistics does not exist, then the MLE is inadmissible. We say that an estimator $\hat{\theta}$ is **inadmissible** if there exists $\tilde{\theta}$ such that $\forall \theta \in \Theta\ \mathbb{E}[l(\hat{\theta}, \theta)] \geq \mathbb{E}[l(\tilde{\theta}, \theta)]$, with equality strict for some θ.

As an example of a case where MLE is not MVUE: the MLE of the variance of a normal distribution is given by:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We know that the minimum variance unbiased estimator for this quantity is given by:

$$\hat{\sigma}^2_{MVUE} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In general, if an estimator is efficient and unbiased, then it is MVUE. The converse is not true, i.e. an MVUE need not be efficient.

have done most of the legwork above. The likelihood of observing that $y_i = 1$, given a particular guess of parameters $\boldsymbol{\beta}$ is simply $\Phi(\mathbf{X}_i\boldsymbol{\beta})$. The likelihood of observing that $y_i = 0$ given that same guess of parameters is $1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})$. Together, this implies that we can write the likelihood as:

$$\mathcal{L}(\boldsymbol{\beta}\,;\mathbf{X},\mathbf{y}) = \prod_{i=1}^{n}(1-\Phi(\mathbf{X}_i\boldsymbol{\beta}))^{1-y_i}(\Phi(\mathbf{X}_i\boldsymbol{\beta}))^{y_i}$$

which implies the normal log likelihood of

$$l(\boldsymbol{\beta}\,;\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n}(1-y_i)ln(1-\Phi(\mathbf{X}_i\boldsymbol{\beta})) + y_i ln(\Phi(\mathbf{X}_i\boldsymbol{\beta}))$$

We then put this in an optimization program and solve numerically for $\boldsymbol{\beta}$.

11.3 Logit

The core difference between the logit and the probit is the distribution that we assumed that the latent shock ϵ_i follows. In probit, we assume that the shock follows a standard normal distribution, whereas in logit we assume that the shock follows a standard logistic distribution. This is

$$F(x;0,1) = \frac{1}{1+e^{-x}}$$

This means that

$$p_i \equiv \mathbb{P}(y_i = 1|\mathbf{X}_i) = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$$

Similarly, we have that

$$1 - p_i \equiv \mathbb{P}(y_i = 0|\mathbf{X}_i) = 1 - \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}} = \frac{e^{-\mathbf{X}_i\boldsymbol{\beta}}}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$$

Dividing p_i by $1 - p_i$, we have that

$$\frac{p_i}{1-p_i} = \frac{1}{e^{-\mathbf{X}_i\boldsymbol{\beta}}}$$

Apply $ln()$ to each side, so that we have

$$ln(\frac{p_i}{1-p_i}) = \mathbf{X}_i\beta$$

Now we will make a small observation that is useful, namely that

$$p_i = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}} = \frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}$$

Similarly, we have

$$1 - p_i = 1 - \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}} = \frac{1}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}$$

This expression of probabilities in terms of the logit is the more common way of formulating the MLE.

With expressions for the probability that $y_i = 1$ and 0, we can now turn to the problem of formulating the likelihood. Given a $\boldsymbol{\beta} \in \mathcal{B}$, the likelihood is given by

$$\mathcal{L}(\boldsymbol{\beta}\,;\mathbf{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)^{1-y_i}$$

Let's take the log of the likelihood of a single observation and then simplify.

$$ln\left(\left(\frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)^{1-y_i}\right)$$

Using properties of logs, we have that we can rewrite this as

$$y_i ln\left(\frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right) + (1-y_i) ln\left(\frac{1}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)$$

Grouping terms that are premultiplied by y_i, we have

$$y_i\left[ln\left(\frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right) - ln\left(\frac{1}{1+e^{\mathbf{X}_i\boldsymbol{\beta}}}\right)\right] + ln\left(\frac{1}{1+e^{\mathbf{X}_i\beta}}\right)$$

Simplifying, we have

$$l(\beta;\ y_i) = y_i\mathbf{X}_i\beta - ln(1+e^{\mathbf{X}_i\beta})$$

The aggregate log likelihood is then

$$l(\beta;\ \mathbf{y}) = \sum_{i=1}^{n} y_i\mathbf{X}_i\beta - ln(1+e^{\mathbf{X}_i\beta})$$

To find $\hat{\beta}$, we numerically solve the following problem:

$$\max_{\beta\in\mathcal{B}} \sum_{i=1}^{n} y_i\mathbf{X}_i\beta - ln(1+e^{\mathbf{X}_i\beta})$$

12 Heckman Correction

These notes are heavily based on notes provided by Andrew Goodman Bacon at the Minneapolis Fed.

Suppose that we are interested in estimating the increase in say yearly earnings from college estimation. We have a sample for a bunch of workers that includes their years of experience at their job, their intelligence, the socioeconomic status of their parents, and so on. Now suppose we were to simply regress their wage on a bunch of covariates and a dummy that is equal to 1 if they attended college and zero if not. That is, we want to run the regression

$$y_i = \mathbf{X}_i\hat{\beta} + \hat{\gamma}D_i + e_i$$

The question is, does $\hat{\gamma}$ recover the true effect of going to college? The answer is no, and the reason is basically *we are ignoring the fact that people choose to go to college or not for a reason.* We call this a selection problem. People who choose to go to college *select* into treatment.

Ok, so we have a problem. How do we fix it? Heckman's solution is roughly as follows: first we model the selection process, and then we use the output from that model as a covariate in our OLS regression. This is the general idea of a Heckman correction, so it is good to keep in mind. But before we get to actually implementing the Heckman correction, we have to take a slight detour.

12.1 The Roy Model

Suppose that an individual i is forced to pick from two alternatives a and b. The utility agent i gets from picking option a is $\mu_a + \epsilon_{i,a}$, and the utility that they get from option b is $\mu_b + \epsilon_{i,b}$, where μ_j is some common level of utility across the population, and $\epsilon_{i,j}$ can be thought of as some idiosyncratic shock to the utility that agent i gets from choosing j. Obviously, agent i picks a if $\mu_a + \epsilon_{i,a} > \mu_b + \epsilon_{i,b}$. Now let us make a particular distributional assumption here: namely suppose that

$$\begin{bmatrix}\epsilon_a \\ \epsilon_b\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2\end{bmatrix}\right)$$

In words we would say that the ϵs are jointly normal with a correlation of ρ. This is an important assumption that it is worth discussing in detail. The first thing to note is that this assumption is not equivalent to assuming that each of the marginals is normal. The sum of two correlated normal distributions **is not** in general, normal. This strictly stronger assumption buys us a lot:

1. The sum of linear combinations of the marginals is normal. For instance,

$$\epsilon_a - \epsilon_b \sim \mathcal{N}(0, \sigma_a^2 + \sigma_b^2 - 2\rho\sigma_a\sigma_b)$$

2. $\mathbb{E}[\epsilon_j | \epsilon_j > c] = \sigma_j \frac{\phi(c/\sigma_j)}{1-\Phi(c/\sigma_j)}$. The quantity $\frac{\phi(c/\sigma_j)}{1-\Phi(c/\sigma_j)}$ is called the Inverse Mills ratio, and it is the ratio of the marginal and the survival function.

3. $\mathbb{E}[\epsilon_j | \epsilon_k > c] = \rho\frac{\sigma_j}{\sigma_k}\left(\sigma_k \frac{\phi(c/\sigma_k)}{1-\Phi(c/\sigma_k)}\right) = \frac{\rho\sigma_j\sigma_k}{\sigma_k^2}\left(\sigma_k \frac{\phi(c/\sigma_k)}{1-\Phi(c/\sigma_k)}\right)$

Now suppose that choosing one option forces you to incur a cost but the other does not. For instance, suppose you are choosing to go to college. The decision to go to college has a bunch of constituent costs:

- You are leaving your family behind, which may or may not be a psychological cost, depending on your relationship with them I suppose.
- You are foregoing the earnings you would've earned from working while you are in college.
- You have to pay the actual pecuniary costs of going to college.
- Certainly other less obvious things.

So given an interest rate r, you choose to go to college if your increase in lifetime earnings is greater than what you forgo, which we can simply say is the number of years in school times the market rate.[46] Hence you choose to go to college if

$$(\epsilon_c - \epsilon_h) > sr - (\mu_c - \mu_h)$$

In words, we would say that you choose to go to college if your internal return $\mu_c + \epsilon_c - \mu_h - \epsilon_h$ surpasses the cost of college sr.

So what is the expected earnings for college grads in this world? It is:

$$\mathbb{E}[ln(y_{c,i})|College] = \mu_c + \mathbb{E}[\epsilon_{c,i}|\epsilon_{c,i} - \epsilon_{h,i} > sr - (\mu_c - \mu_h)]$$

The distribution of $\epsilon_{c,i} - \epsilon_{h,i} \sim \mathcal{N}(0, \sigma_c^2 + \sigma_h^2 - 2\rho\sigma_c\sigma_h)$. Call this variance σ_ϵ^2

We also need to know the covariance of $\epsilon_{c,i}$ with $\epsilon_{c,i} - \epsilon_{h,i}$. This is $\sigma_c^2 - \rho\sigma_c\sigma_h$.

From our claims about the jointly normal R.V.s above, we know that this is true:

$$\mathbb{E}[\epsilon_{c,i}|\epsilon_{c,i} - \epsilon_{h,i} > sr - (\mu_c - \mu_h)] = \frac{(\sigma_c^2 - \rho\sigma_c\sigma_h)}{\sigma_\epsilon} \frac{\phi((sr - (\mu_c - \mu_h))/\sigma_\epsilon)}{1 - \Phi((sr - (\mu_c - \mu_h))/\sigma_\epsilon)}$$

12.2 Heckman Correction

This suggests that we need to first model the selection process, and then use whatever we get out of that to help restrain β. This first step is called estimating the selection equation. The second is called estimating an outcome equation. Specifying the selection equation amounts to specifying the reduced form things that matter to peoples decisions when making the choice. Their parent's household income for instance might matter, cultural attitudes toward college education might matter, and so on. The outcome equation is typically a simple mincer style wage regression.

More formally, suppose that selection is governed by a process of the following form:

$$D_i = \mathbf{1}\{\mathbf{W}_i\boldsymbol{\gamma} + u_i > 0\}$$

And suppose that outcomes - observed wages - are governed by the following process:

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$$

Where again, u_i and ϵ_i are jointly normal with correlation coefficient ρ and respective variances σ_s^2 and σ_o^2

The expected wage of college graduates in the sample is going to be:

$$\mathbb{E}[y_i|\mathbf{X}, D_i = 1] = \mathbf{X}_i\boldsymbol{\beta} + \mathbb{E}[\epsilon_i|\mathbf{X}_i, u_i > -\mathbf{W}_i\boldsymbol{\gamma}]$$

We know from the discussion above that

$$\mathbb{E}[\epsilon_i|u_i > -c] = \frac{\rho\sigma_o\sigma_s}{\sigma_s^2}\mathbb{E}\left[\frac{\epsilon_i}{\sigma_s}|\frac{u_i}{\sigma_s} > -\frac{c}{\sigma_s}\right] = \frac{\rho\sigma_o\sigma_s}{\sigma_s^2}\frac{\phi(-c/\sigma_s)}{1 - \Phi(-c/\sigma_s)}$$

[46] The percentage increase in lifetime earnings is approximately $ln(y_c) - ln(y_h)$. To recover the actual amount of earnings, you can multiply this by the high school salary, so that your left hand side reads $y_h[ln(y_c) - ln(y_h)]$. This is the amount of extra income you would earn by going to college. Discount this back to the future (imagine you only save for retirement), so that the left hand side reads $\frac{y_h[ln(y_c) - ln(y_h)]}{r}$ How much do you forgo? You forgo s years of schooling, where you earn y_{hi}. Hence you lose $y_h s$

By the symmetry of the unit normal, we have that $\phi(-\alpha) = \phi(\alpha)$, and that $\Phi(-\alpha) = 1 - \Phi(\alpha)$, so we can further rewrite the above as:

$$\frac{\rho\sigma_o\sigma_s}{\sigma_s^2}\frac{\phi(c/\sigma_s)}{\Phi(c/\sigma_s)}$$

To see what we have to do, we should substitute back in for c, so that we recover:

$$\frac{\rho\sigma_o\sigma_s}{\sigma_s^2}\frac{\phi(\mathbf{W}_i\boldsymbol{\gamma}/\sigma_s)}{\Phi(\mathbf{W}_i\boldsymbol{\gamma}/\sigma_s)}$$

If we were to regress the wages that we see y on $\mathbf{X}$ and this quantity $\frac{\phi(\mathbf{W}_i\boldsymbol{\gamma}/\sigma_s)}{\Phi(\mathbf{W}_i\boldsymbol{\gamma}/\sigma_s)}$, we would recover the true census coefficient (which would be the causal quantity, in the absence of any other endogeneity). The problem is that we don't observe $\boldsymbol{\gamma}$, so we have to estimate it. This is not that big a deal, since what we can do is simply estimate a probit regression to get $\hat{\boldsymbol{\gamma}}$.

So, the Heckman correction proceeds in two steps:

1. First we estimate a probit model to get an estimate of $\boldsymbol{\gamma}$. We then calculate the IMR using this estimate (so we calculate $\frac{\phi(\mathbf{W}_i\hat{\boldsymbol{\gamma}}/\sigma_s)}{\Phi(\mathbf{W}_i\hat{\boldsymbol{\gamma}}/\sigma_s)}$)
2. Run the following regression:

$$y_i = \mathbf{X}_i\beta + \delta\frac{\phi(\mathbf{W}_i\hat{\boldsymbol{\gamma}}/\sigma_s)}{\Phi(\mathbf{W}_i\hat{\boldsymbol{\gamma}}/\sigma_s)} + \xi_i$$

Note that the coefficient $\hat{\delta}$ is consistent for $\rho\frac{\sigma_o}{\sigma_s}$. σ_o and σ_s are strictly positive quantities (they are the square root of variances), so the sign of $\hat{\delta}$ is informative about the sign of ρ. More directly, the significance of $\hat{\delta}$ is informative about whether or not there is selection!