Advanced Econometrics I

Jürgen Meinecke

Week 1 Lecture

Research School of Economics, Australian National University

Welcome

Welcome Advanced Econometrics I

This is a PhD level course in econometric theory

The course makes heavy use of the following mathematical tools:

- linear algebra
- multivariate calculus
- concepts in analysis (real and functional)

If you don't feel familiar with these, then this course will be extremely demanding

Staff

You can seek help on matters *academic* from

- your friendly lecturer (me): Juergen Meinecke
- your amazing tutor: Shu Hu

Be nice to us!

Course Website

I'm not using Canvas very much (with few exceptions)

I've set up a public course website that contains pretty much everything you need to know

Let's take a look:

`https://juergenmeinecke.github.io/EMET8014`

Roadmap

Announcements

Vector Spaces, Hilbert Spaces, Projections

- Vector Spaces
- Normed Spaces, Banach Spaces
- Inner Product Spaces, Hilbert Spaces
- Projection Theorem
- Linear Projections in L_2

Definition (Vector Space)

A **real vector space** is a triple $(\mathcal{V}, +, \cdot)$, in which $\mathcal{V}$ is a set, and $+$ and $\cdot$ are binary operations such that, for any two elements, $X, Y \in \mathcal{V}$ and scalar $\lambda \in \mathbb{R}$:

$$X + Y \in \mathcal{V} \qquad \text{(closure under additivity)}$$

$$\lambda \cdot X \in \mathcal{V} \qquad \text{(closure under scalar product)}$$

(Note: instead of writing $\lambda \cdot X$ we typically just write λX)

Big picture:
We define a notion of addition between any two elements of $\mathcal{V}$, and we define a notion of multiplication between a constant and an element of $\mathcal{V}$

These operations do not take us out of the vector space

With addition and multiplication there are some typical sensible ‘requirements’:

Let $X, Y, Z \in \mathcal{V}$, and $\lambda, \mu \in \mathbb{R}$

- addition
  - (i) commutativity: $X + Y = Y + X$
  - (ii) associativity: $(X + Y) + Z = X + (Y + Z)$
  - (iii) $\mathcal{V}$ contains a unique element 0 such that $X + 0 = X$
  - (iv) $\mathcal{V}$ contains a unique element $-X$ such that $X + (-X) = 0$
- multiplication
  - (i) distributivity: $\lambda \cdot (X + Y) = \lambda \cdot X + \lambda \cdot Y$
  - (ii) distributivity: $(\lambda + \mu) \cdot X = \lambda \cdot X + \mu \cdot X$
  - (iii) associativity: $\lambda \cdot (\mu \cdot X) = (\lambda \cdot \mu) \cdot X$
  - (iv) $1 \cdot X = X$

The perhaps most intuitive illustration of a real vector space:

Example (Euclidian Space $(\mathbb{R}^n, +, \cdot)$)

- *elements are quite literally vectors or arrows*
- $X := (x_1, \ldots, x_n)'$ *and* $Y := (y_1, \ldots, y_n)'$
- *define* $X + Y := (x_1 + y_1, \ldots, x_n + y_n)'$
- *define* $\lambda \cdot X := (\lambda x_1, \ldots, \lambda x_n)'$
- *let* $n = 2$*,* $X = (24, 7)'$ *and* $Y = (18, 2)'$*,*
  then $X + Y = (42, 9)'$

When $X \in \mathcal{V}$, I refer to X as an "*element*" of $\mathcal{V}$

Some books use "*vector*", one could also say "*point*"

A less intuitive example...

Example (The Space of Continuous Functions)

Denote by $C[a,b]$ the space of all real valued univariate and continuous functions on a closed interval $[a,b]$.

- *each $X \in C[a,b]$ is a function $X : [a,b] \to \mathbb{R}$*
- *the points or elements of the space are functions*
- *let $t \in [a,b]$ and write $X(t)$ for the function value at t*
- *define $(X+Y)(t) := X(t) + Y(t)$*
- *define $(\lambda \cdot X)(t) := \lambda \cdot X(t)$*
- *let $[a,b] = [2,3]$, $X(t) = 2 \cdot t$ and $Y(t) = 1 + 5 \cdot t$, then $(X+Y)(t) = 1 + 7 \cdot t$*

Vector spaces of functions are very important, more examples:

- space of differentiable functions
- space of functions that are integrable (random variables live in this space)

Roadmap

Each element of a vector space can be given a ‘length’:

Definition (Norm)

A **norm** on a real vector space $\mathcal{V}$ is a real valued function, denoted by $\|\cdot\|$, on $\mathcal{V}$ with the properties

(i) $\|X\| \geq 0$

(ii) $\|X\| = 0 \Leftrightarrow X = 0$

(iii) $\|\lambda \cdot X\| = |\lambda| \cdot \|X\|$

(iv) triangle inequality: $\|X + Y\| \leq \|X\| + \|Y\|$

where X and Y are in $\mathcal{V}$, and λ is a real constant

Any function $\|\cdot\| : \mathcal{V} \to \mathbb{R}_+$ that satisfies these properties is a norm

Here are useful examples...

Examples of norms

Example (Euclidian Space $(\mathbb{R}^n, +, \cdot)$)

- *recall that elements are quite literally vectors or arrows*
- $X := (x_1, \dots, x_n)'$
- $\|X\| := \sqrt{x_1^2 + \cdots + x_n^2}$

Example (The Space of Continuous Functions)

- *recall that each $X \in C[a, b]$ is a function $X : [a, b] \to \mathbb{R}$*
- $\|X\| := \max_{t \in [a,b]} |X(t)|$

Definition (Normed Space)

A **normed space** $\mathcal{M}$ is a vector space endowed with a norm $\|\cdot\|$.

The norm *induces* a metric, a notion of distance between two elements

Definition (Metric)

Given a normed space $\mathcal{M}$ and $X, Y \in \mathcal{M}$, the **metric** is defined by $\|X - Y\|$.

A notion of distance between elements is crucial for the understanding of limiting behavior of sequences inside the vector space

For example, does a sequence $\{X_n, n = 1, 2, \ldots\}$ of elements of a normed space get "close" to a point?

The metric is key in determining what we mean by "closeness"

Definition (Convergence in Norm)

A sequence $\{X_n, n = 1, 2, \ldots\}$ of elements of a normed space $\mathcal{M}$ is said to **converge in norm** (or simply **converge**) to $X \in \mathcal{M}$ if for every $\varepsilon > 0$ there is an N_ε such that $\|X_n - X\| < \varepsilon$ for every $n > N_\varepsilon$.

This definition requires explicit knowledge of the limit X

In optimization settings (such as the ones we work with in econometrics), often times we do not know these limits

We're looking for an alternative way to characterize convergence

We want to consider sequences of elements of a normed space $\mathcal{M}$ that become closer and closer to each other

That's the idea behind **Cauchy sequences**:

Definition (Cauchy Sequence)

A sequence $\{X_n, n = 1, 2, \ldots\}$ of elements of a normed space $\mathcal{M}$ is said to be a **Cauchy sequence** if for every $\varepsilon > 0$ there is an N_ε such that $\|X_m - X_n\| < \varepsilon$ for every $m, n > N_\varepsilon$.

Idea: elements 'out there' become arbitrarily close to each other

This is captured in the so-called *Cauchy criterion* $\|X_m - X_n\| < \varepsilon$

It 'feels' like Cauchy sequences do converge!

Careful: only the converse is true!

If a sequence converges, then it must by Cauchy

But sadly: not every Cauchy sequence converges (problem set 1)

Notwithstanding, we would like to deal with sequences that rely exclusively on the Cauchy criterion

In most applications, the Cauchy criterion will be practical to check

It provides a feasible test in applications without the reliance on an explicit limit

Long story short:
We want to work on spaces in which the Cauchy criterion is necessary and sufficient for convergence

Definition (Complete Space)

A normed space $\mathcal{M}$ is **complete** if every Cauchy sequence of elements of $\mathcal{M}$ converges to an element of $\mathcal{M}$, that is, every Cauchy sequence in $\mathcal{M}$ has a limit which is an element of $\mathcal{M}$.

We like complete spaces because we can use the convenient Cauchy criterion to safely consider limits of elements within the space

Definition (Banach Space)

A **Banach space** $\mathcal{B}$ is a complete normed space, that is, a normed space in which every Cauchy sequence $\{X_n, n = 1, 2, \ldots\}$ converges in norm to some element $X \in \mathcal{B}$.

In Banach spaces we can safely play with length, distance of elements and sequences and limits of sequences of elements

But something is still missing...

Roadmap

Given a vector space or Banach space, we can add elements and multiply them by scalars

We can also measure their length and distance via the metric

We would, in addition, like notions of

- multiplication between elements of a space
- angle, or *orthogonality*, or perpendicularity between elements of a space

The inner product comes to the rescue

Definition (Inner Product)

An **inner product** on a vector space $\mathcal{V}$ is a mapping, denoted by $\langle \cdot, \cdot \rangle$, of $\mathcal{V} \times \mathcal{V}$ into $\mathbb{R}$ such that

$$
\begin{aligned}
\langle X, Y \rangle &= \langle Y, X \rangle && \text{(commutativity)} \\
\langle X + Y, Z \rangle &= \langle X, Z \rangle + \langle Y, Z \rangle && \text{(distributivity)} \\
\langle \lambda X, Y \rangle &= \lambda \langle X, Y \rangle && \\
\langle X, X \rangle &\geq 0 && \text{(positive semi-definiteness)} \\
\langle X, X \rangle = 0 &\iff X = 0 && \text{(point separating)}
\end{aligned}
$$

where X, Y and Z are in $\mathcal{V}$, and λ is a real constant.

Examples of inner products

Example (Euclidian Space $(\mathbb{R}^n, +, \cdot)$)

- *recall that elements are quite literally vectors or arrows*
- $X := (x_1, \ldots, x_n)'$ *and* $Y := (y_1, \ldots, y_n)'$
- $\langle X, Y \rangle := x_1 y_1 + \cdots x_n y_n$

Example (The Space of Continuous Functions)

- *let* X *and* Y *be real-valued functions on* $[a, b]$
- $\langle X, Y \rangle := \int_a^b X(t) Y(t) dt$

Inner products lend themselves naturally to the creation of a norm

Definition (Induced Norm)

Let $\mathcal{M}$ be a normed space. The **norm induced by the inner product** is $\|X\| := \sqrt{\langle X, X \rangle}$, for any $X \in \mathcal{M}$.

Likewise there is a metric $\|X - Y\|$ induced by the inner product

Definition (Inner Product Space)

An **inner product space** is a vector space endowed with an inner product $\langle \cdot, \cdot \rangle$.

Definition (Hilbert Space)

A **Hilbert space** is a complete inner product space.

It is clear that completeness is with respect to the norm induced by the inner product

It follows that all Hilbert spaces are Banach spaces
(but not all Banach spaces are Hilbert spaces)

Why do we like Hilbert spaces?
They are the subset of the Banach spaces that behave very similarly to Euclidian space, while being considerably more general

A lot of intuition from Euclidian space carries over to Hilbert space (case in point: triangle inequality)

Equipped with the inner product, we can now define the notion of angle between elements of an inner product space

Definition (Orthogonality)

Two elements X, Y of a Hilbert space are **orthogonal** if $\langle X, Y \rangle = 0$.

We write $X \perp Y$.

Think of vectors that are perpendicular
(this is another case in point of Euclidean geometry carries over to the more general Hilbert space setting)

Roadmap

Hilbert spaces are useful for econometrics because they offer us powerful tools to address the following optimization problem:

- given an element Y in a Hilbert space $\mathcal{H}$,
- and a subspace $\mathcal{S}$ of $\mathcal{H}$,
- find the element $\widehat{Y} \in \mathcal{S}$ closest to Y in the sense that $\|Y - \widehat{Y}\|$ is minimal

Key questions

- is there such an element $\widehat{Y}$?
- is it unique?
- what is it, or how can it be characterized?

The projection theorem answers these questions

Definition (Subspace)

A subset $\mathcal{S}$ of a Hilbert space $\mathcal{H}$ is called a **subspace** of $\mathcal{H}$ if $\mathcal{S}$ itself is a vector space.

We will focus on *complete* subspaces of Hilbert spaces:
subspaces that contain all of their limit points
(i.e., addition and multiplication don't take us out of the subspace)

Theorem (Projection Theorem)

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{S}$ be a complete subspace of $\mathcal{H}$.

(i) *For any element $Y \in \mathcal{H}$ there is a unique element $\widehat{Y} \in \mathcal{S}$ such that $\|Y - \widehat{Y}\| \leq \|Y - s\|$ for all $s \in \mathcal{S}$.*

(ii) *$\widehat{Y} \in \mathcal{S}$ is the unique minimizer if and only if $Y - \widehat{Y} \perp \mathcal{S}$.*

The element $\widehat{Y}$ is called the **orthogonal** projection of Y onto $\mathcal{S}$, also denoted $\mathbb{P}_{\mathcal{S}} Y$

$\mathbb{P}_{\mathcal{S}}$ is the projection operator of $\mathcal{H}$ onto $\mathcal{S}$

Existence of a unique minimizer sounds great, but how to obtain $\widehat{Y}$?

Roadmap

Before we come up with a constructive result for $\hat{Y}$, let's first take a deep breath and consider our Hilbert space situation

When we consider regression equations such as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u,$$

the Hilbert space setting is this:

- Y is an outcome variable (eg, earnings)
- $X_1, \ldots, X_K$ are explanatory variables (schooling, age, etc)

We regard $Y, X_1, \ldots, X_K$ as random variables on the probability space $(\Omega, \mathscr{F}, P)$

We would like them to be well behaved random variables...

We would like them to be from the following collection:

Definition (The Space $L_2(\Omega, \mathscr{F}, P)$)

$L_2(\Omega, \mathscr{F}, P)$ is the set of all random variables with finite second moments.

I'll refer to $L_2(\Omega, \mathscr{F}, P)$ simply as L_2

The condition $\mathrm{E}\left(Y^2\right) < \infty$ is sometimes referred to as finite second moment or X being square integrable; it implies that $\mathrm{Var}\,(Y) < \infty$

L_2 is a huge set, but is it a Hilbert space?

Need an inner product: let $\langle X, Y \rangle := \mathrm{E}(X \cdot Y)$

Let $X, Y \in L_2$ with $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, then $\langle X, Y \rangle = \mathrm{Cov}(X, Y)$

In other words, the inner product we're using here is related to the familiar notion of covariance

Proposition

The space L_2 with $\langle X, Y \rangle = E(X \cdot Y)$ is a Hilbert space.

Translating this into the language of the projection theorem

If I run a regression of Y on X_1 only, then the subspace used for the projection is called $\text{sp}(X_1)$

That's the subspace spanned by X_1:
the set of all random variables that you can build 'linearly' from X_1

If I run a regression of Y on X_1 and X_2, then the subspace used for the projection is called $\text{sp}(X_1, X_2)$

That's the subspace spanned by X_1 and X_2:
the set of all random variables that you can build from the linear combos of X_1 and X_2

If I run a regression of Y on $X_1, \ldots, X_K$, then the subspace used for the projection is called $\text{sp}(X_1, \ldots, X_K)$

That's the subspace spanned by $X_1, \ldots, X_K$:
the set of all random variables that you can build from the linear combos of $X_1, \ldots, X_K$

Formal definition

Definition (Span)

Let $X_1, \dots, X_K$ be elements of a vector space. The **span** of $X_1, \dots, X_K$ is the set of all linear combinations $b_1 X_1 + \cdots + b_K X_K$, where the b_i are real numbers.

Notice:
While the finite collection $X_1, \dots, X_K$ are not a subspace, their span is a subspace by construction

Fun fact:
if $\mathcal{V}$ is a vector space, and if $X_1, \dots, X_K$ are elements of $\mathcal{V}$ then $\text{sp}(X_1, \dots, X_K)$ is the smallest subspace of $\mathcal{V}$ containing each X_k, $k = 1, \dots, K$

Obtaining the projection $\widehat{Y}$ of Y on sp$(X_1, \dots, X_K)$ turns out to be straightforward

We need to use so-called orthonormal bases

Think of them, tentatively, as modified versions of $X_1, \dots, X_K$

Theorem

Let $X_1, \dots, X_K$ and Y be elements from a Hilbert space. The projection of Y on sp$(X_1, \dots, X_K)$ is

$$\widehat{Y} = \mathbb{P}_{sp(X)} Y = \sum_{i=1}^{K} \langle \widetilde{X}_i, Y \rangle \cdot \widetilde{X}_i,$$

where $\widetilde{X}_1, \dots, \widetilde{X}_K$ is an orthonormal basis for sp(X).

This gives us a constructive method for obtaining $\widehat{Y}$

But first let's look at what orthonormal bases are...

In econometrics, we are usually interested in the subspace spanned by 'regressors' $X_1, \dots, X_K$

Collect them all in the vector $X = (X_1, \dots, X_K)$

Theorem (Existence of Orthonormal Basis (Gram-Schmidt))

There exists is a collection $\widetilde{X}_1, \dots, \widetilde{X}_K$ such that

(i) $\langle \widetilde{X}_j, \widetilde{X}_l \rangle = \begin{cases} 0 & \textit{for } j \neq l \\ 1 & \textit{for } j = l. \end{cases}$

(ii) $sp(\widetilde{X}) = sp(X)$ *for* $\widetilde{X} := (\widetilde{X}_1, \dots, \widetilde{X}_K)$.

The collection $\widetilde{X}_1, \dots, \widetilde{X}_K$ is called an **orthonormal basis** for sp(X)

Why do we consider orthonormal bases?

Recall our theorem from two slides ago:

The projection of Y on $\mathrm{sp}(X_1, \dots, X_K)$ is

$$\widehat{Y} = \mathbb{P}_{\mathrm{sp}(X)} Y = \sum_{i=1}^{K} \langle \widetilde{X}_i, Y \rangle \cdot \widetilde{X}_i,$$

Also remember: we're using $\langle \widetilde{X}_i, Y \rangle \coloneqq \mathrm{E}(\widetilde{X}_i \cdot Y)$

Now, let's work this out explicitly, step-by-step

We go slow and set $K = 1$

That is, we only have one regressor X_1 to predict Y

Going back to the question: What is $\widehat{Y}$ equal to?

Answer, of course, depends on choice of subspace to project on

In the current example we have sp(X_1)
(so only one random variable)

Let's make the problem even simpler and pick $X_1 = 1$
(the degenerate rv that is almost surely equal to 1)

What is the orthonormal basis for sp(1)?

Easy: X_1 already is an orthonormal basis (because $\mathsf{E}(1 \cdot 1) = 1$), so:

$$\widehat{Y} = \mathbb{P}_1 Y = \sum_{i=1}^{K} \langle \widetilde{X}_i, Y \rangle \cdot \widetilde{X}_i = \sum_{i=1}^{1} \langle \widetilde{X}_i, Y \rangle \cdot \widetilde{X}_i = \langle \widetilde{X}_1, Y \rangle \cdot \widetilde{X}_1$$
$$= \langle 1, Y \rangle \cdot 1 = \mathsf{E}(1 \cdot Y) \cdot 1 = \mathsf{E} Y = \mu_Y$$

Of course you knew this already:
The projection of Y onto a constant is the expected value of Y

What if we use a more sophisticated space for the projection?

Let $X_1 = 1$ and $X_2 \in L_2$ and project on $\text{sp}(1, X_2)$

Let's find an orthonormal basis of $\text{sp}(1, X_2)$

We need to find a version of X_2 that has length 1

This is easy: $\widetilde{X}_2 := (X_2 - \mu_2)/\sigma_2$ achieves this:
$\|\widetilde{X}_2\| = \sqrt{\mathsf{E}\left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2} = 1$

Turns out that $\{1, \widetilde{X}_2\}$ form an orthonormal basis of $\text{sp}(1, X_2)$ (confirm this!)

The example of $\widetilde{X}_2$ offers you some intuition about orthonormal bases: it is the *standardized version* of X_2

With $\widetilde{X}_1 := 1$ and $\widetilde{X}_2 = (X_2 - \mu_2)/\sigma_2$, it follows,

$$\begin{aligned}
\widehat{Y} = \mathbb{P}_{\mathsf{sp}(1,X_2)} Y &= \sum_{i=1}^{2} \mathsf{E}(\widetilde{X}_i \cdot Y) \cdot \widetilde{X}_i \\
&= \mathsf{E}(1 \cdot Y) \cdot 1 + \mathsf{E}(\widetilde{X}_2 Y) \cdot \widetilde{X}_2 \\
&= \mathsf{E}Y + \mathsf{E}(((X_2 - \mu_2)/\sigma_2)Y)(X_2 - \mu_2)/\sigma_2 \\
&= \mathsf{E}Y + \frac{\mathsf{E}(X_2 Y) - \mu_2 \mathsf{E}Y}{\sigma_2^2}(X_2 - \mu_2) \\
&= \mathsf{E}Y + \frac{\mathsf{Cov}(X_2, Y)}{\sigma_2^2}(X_2 - \mu_2) \\
&= \beta_1^* + \beta_2^* X_2,
\end{aligned}$$

where

$$\begin{aligned}
\beta_2^* &:= \sigma_{2Y}/\sigma_2^2 \\
\beta_1^* &:= \mathsf{E}Y - \beta_2^* \mathsf{E}X_2
\end{aligned}$$

Let's generalize this once more

What if we're projecting on $\mathrm{sp}(1, X_2, X_3)$?

One might think that $\{1, \widetilde{X}_2, \widetilde{X}_3\}$ form an orthonormal basis, where $\widetilde{X}_2 := (X_2 - \mu_2)/\sigma_2$ and $\widetilde{X}_3 := (X_3 - \mu_3)/\sigma_3$

Problem: most generally $\mathrm{E}(\widetilde{X}_2 \widetilde{X}_3) \neq 0$

Intuitively, the problem is that X_2 and X_3 could be correlated

How do we construct orthonormal bases out of two random variables that are correlated?

Answer: Gram-Schmidt orthogonalization!

Gram-Schmidt orthogonalization applied to the current context, involves these simple steps:

1. Set $\widetilde{X}_1 \ := 1$
2. create $\ddot{X}_2 \ := X_2 - \mathsf{E}(X_2\widetilde{X}_1) \cdot \widetilde{X}_1$
   normalize by its length: $\widetilde{X}_2 \ := \ddot{X}_2/\|\ddot{X}_2\|$
   (notice that because we include a constant term,
   $\mathsf{Var}\ \ddot{X}_2 = \mathsf{E}\ddot{X}_2^2$ and therefore $\|\ddot{X}_2\| = \sqrt{\mathsf{Var}\ \ddot{X}_2}$)
3. create $\ddot{X}_3 \ := X_3 - \mathsf{E}(X_3\widetilde{X}_1) \cdot \widetilde{X}_1 - \mathsf{E}(X_3\widetilde{X}_2) \cdot \widetilde{X}_2$
   normalize by its length: $\widetilde{X}_3 \ := \ddot{X}_3/\|\ddot{X}_3\|$
   (notice again that $\|\ddot{X}_3\| = \sqrt{\mathsf{Var}\ \ddot{X}_3}$)

You can view Gram-Schmidt orthogonalization as an iterative algorithm to construct orthonormal bases

By the way, the order in which you are doing this does not matter

Letting $\sigma_{23} := \text{Cov}(X_2, X_3)$, we obtain (try it!):

$$\ddot{X}_2 = X_2 - \mu_2$$
$$\ddot{X}_3 = X_3 - \frac{\sigma_{23}}{\sigma_2^2}(X_2 - \mu_2) - \mu_3$$

What's going on here?

$\ddot{X}_2$ is a particular version of the original variable X_2:
take X_2 and subtract off the projection of X_2 on X_1 (the constant)

Similarly for $\ddot{X}_3$: take X_3 and subtract off the projection of X_3 on X_2 and also the projection of X_3 on X_1

In this sense we're creating orthogonalized versions of all explanatory variables

By construction, $\ddot{X}_1$, $\ddot{X}_2$, and $\ddot{X}_3$ will be uncorrelated

Their transformations $\widetilde{X}_1$, $\widetilde{X}_2$, and $\widetilde{X}_3$ are normalized to have length 1

Aside: if $\sigma_{23} = 0$ then $\widetilde{X}_3 = (X_3 - \mu_3)/\sigma_3$

With the orthonormal basis it's easy to construct the projection:

$$\mathbb{P}_{\mathsf{sp}(1,X_2,X_3)}Y = \sum_{i=1}^{3} \mathsf{E}(\widetilde{X}_i \cdot Y) \cdot \widetilde{X}_i$$

It is tedious but not difficult to show that

$$\begin{aligned}\mathbb{P}_{\mathsf{sp}(1,X_2,X_3)}Y &= \mathsf{E}Y + \beta_2^*(X_2 - \mu_2) + \beta_3^*(X_3 - \mu_3) \\ &= \beta_1^* + \beta_2^* X_2 + \beta_3^* X_3\end{aligned}$$

where

$$\begin{aligned}\beta_2^* &:= \frac{\sigma_{2Y}\sigma_3^2 - \sigma_{3Y}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \\ \beta_3^* &:= \frac{\sigma_{3Y}\sigma_2^2 - \sigma_{2Y}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} \\ \beta_1^* &:= \mathsf{E}Y - \beta_2^*\mathsf{E}X_2 - \beta_3^*\mathsf{E}X_3\end{aligned}$$

Looks awkward but it is an important result to internalize!

Look what happens when Cov$(X_2, X_3) = 0$:

$$\beta_2^* := \frac{\sigma_{2Y}\sigma_3^2 - \sigma_{3Y}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} = \frac{\sigma_{2Y}}{\sigma_2^2}$$

$$\beta_3^* := \frac{\sigma_{3Y}\sigma_2^2 - \sigma_{2Y}\sigma_{23}}{\sigma_2^2\sigma_3^2 - \sigma_{23}^2} = \frac{\sigma_{3Y}}{\sigma_3^2}$$

How would you construct an orthonormal basis for $\mathrm{sp}(X_1, X_2, \dots, X_K)$ with $X_1 = 1$ and $X_k \in L_2$ for $k = 2, \dots, K$?

Again, use Gram-Schmidt orthogonalization with the inductive definitions:

1. $\widetilde{X}_1 := 1$
2. $\ddot{X}_2 := X_2 - \mathrm{E}(X_2 \widetilde{X}_1)\widetilde{X}_1$
   $\widetilde{X}_2 := \ddot{X}_2 / \|\ddot{X}_2\| = \ddot{X}_2 / \sqrt{\mathrm{Var}\, \ddot{X}_2}$
3. $\ddot{X}_3 := X_3 - \mathrm{E}(X_3 \widetilde{X}_1)\widetilde{X}_1 - \mathrm{E}(X_3 \widetilde{X}_2)\widetilde{X}_2$
   $\widetilde{X}_3 := \ddot{X}_3 / \|\ddot{X}_3\| = \ddot{X}_3 / \sqrt{\mathrm{Var}\, \ddot{X}_3}$
4. $\ddot{X}_4 := X_4 - \mathrm{E}(X_4 \widetilde{X}_1)\widetilde{X}_1 - \mathrm{E}(X_4 \widetilde{X}_2)\widetilde{X}_2 - \mathrm{E}(X_4 \widetilde{X}_3)\widetilde{X}_3$
   $\widetilde{X}_4 := \ddot{X}_4 / \|\ddot{X}_4\| = \ddot{X}_4 / \sqrt{\mathrm{Var}\, \ddot{X}_4}$
5. and so forth

The resulting projection will have the form

$$\widehat{Y} = \mathbb{P}_X Y = \sum_{i=1}^{K} \mathsf{E}(\widetilde{X}_i \cdot Y) \cdot \widetilde{X}_i,$$

where, for simplicity, we write $\mathbb{P}_X$ for $\mathbb{P}_{\mathrm{sp}(X_1,\ldots,X_K)}$

The projection can be summarized neatly in matrix notation:

Theorem

Let $X := (X_1, X_2, \ldots, X_K)'$ *be a* $K \times 1$ *vector. Then*

$$\mathbb{P}_X Y = X'\beta^*,$$

where $\beta^* := (E(XX'))^{-1}\, E(XY)$*.*

For simplicity we will write $\mathsf{E}(XX')^{-1}$ for $(\mathsf{E}(XX'))^{-1}$