Advanced Econometric Methods EMET3011/8014

Lecture 9

John Stachurski

Semester 1, 2011

Announcements/Reminders

- Please get a fresh copy of the course notes PDF
- "Full rank" changed to "full column rank"
- Midterm solutions are on the web
- Midterm exam marking still ongoing
- Assignment 2 posted today
- Weighting of assignment 2 is 15%, not 25
- Final exam date set: June 20, 9:15–12:00

Today's Lecture

- Conditioning
- Overdetermined Systems
- Multivariate Linear Least Squares

Conditioning

We now study conditional expectations and their properties

Key idea: Use geometric intuition about $\mathbb{R}^N$ to study RVs

Steps:

1. Define L_2 to be all random variables with finite 2nd moment
2. Define inner product and norm on L_2
3. Introduce an L_2 version of the OPT
4. Use this OPT to define and study conditional expectation

The Space L_2

Euclidean geometry for vectors:

- Inner product of $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x}'\mathbf{y} := \sum_n x_n y_n$
- The norm of vector $\mathbf{x}$ is $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_n x_n^2}$

Similarly, for random variables x and y, we define

- **Inner product** of x and y is $\langle x, y \rangle := \mathbb{E}[xy]$
- **Norm** of x is $|||x||| := \sqrt{\langle x, x \rangle} = \sqrt{\mathbb{E}[x^2]}$

Technical problem: $|||x|||$ may not be defined because $\mathbb{E}\left[x^2\right] = \infty$

Here we restrict attention to RVs with finite second moment

The standard name of this set is

$$L_2 := \{\text{ all RVs } x \text{ with } \mathbb{E}\left[x^2\right] < \infty\}$$

On L_2 the norm $|||\cdot|||$ satisfies same properties $\|\cdot\|$ does on $\mathbb{R}^N$

If $\langle x, y \rangle = 0$, we say that x and y are **orthogonal**, and write $x \perp y$

Exercise: Show that if $x \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(0,1)$ and $x \perp y$, then x and y are independent

The **distance** between x and y is

$$|||x - y||| := \sqrt{\mathbb{E}\left[(x-y)^2\right]}$$

- Analogous to euclidean distance
- A monotone transform of mean squared deviation

Orthogonal Projections in L_2

OPT in $\mathbb{R}^N$ starts with a linear subspace S of $\mathbb{R}^N$

First fix S, then think about how to project onto it

What's a linear subspace of L_2?

Linear Subspaces of L_2

Set $S \subset L_2$ is called a **linear subspace of L_2** if

$$\alpha, \beta \in \mathbb{R} \text{ and } x, y \in S \text{ implies } \alpha x + \beta y \in S$$

Example

Let $T := \{x \in L_2 : \mathbb{E}[x] = 0\}$

Suppose that $\alpha, \beta \in \mathbb{R}$ and $x, y \in T$

Then $\alpha x + \beta y \in T$ because

$$\mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y] = 0$$

As we'll see, conditional expectation is characterized by orthogonal projection in L_2

But what linear subspaces do we want to project onto?

To answer this, we need the notion of "measurability"

Measurability

Let $x_1, \ldots, x_p$ be RVs and let $\mathcal{G} := \{x_1, \ldots, x_p\}$

We say that z is **$\mathcal{G}$-measurable** if z can be written as a deterministic function of the RVs in $\mathcal{G}$

Formally, there exists a function $g \colon \mathbb{R}^p \to \mathbb{R}$ such that

$$z = g(x_1, \ldots, x_p)$$

Interpretation: z is $\mathcal{G}$-measurable if z is deterministic once the variables in $\mathcal{G}$ are realized

Notation:

- In econometrics, $\mathcal{G}$ is often called the **information set**
- If $\mathcal{G} = \{x\}$ then $\mathcal{G}$-measurable $\iff$ x-measurable

Example

If $z = 2x + 3$, then z is x-measurable

To see this formally, we can write $z = g(x)$ when $g(x) = 2x + 3$

Less formally, when x is realized, the value of z can be calculated

Example

If x and y are independent (and non-constant), then y is not x-measurable

Indeed, if y was x-measurable, then we would have $y = g(x)$ for some function g

This contradicts independence of x and y

Example

If $z = x + y$, where x and y are independent, then z is not x-measurable

Intuitively, realization of z cannot be computed until we know the realized value of y

Formal reasoning given in course notes

Example

If $y = \alpha =$ constant, then y is $\mathcal{G}$-measurable for any $\mathcal{G}$

True because $y = \alpha$ is already deterministic

Formally: Take $y = g(x_1, \ldots, x_p) = \alpha + 0 \times \sum_{i=1}^{p} x_i$

The Space $L_2(\mathcal{G})$

Given $\mathcal{G} = \{x_1, \ldots, x_p\} \subset L_2$, we define

$$L_2(\mathcal{G}) := \{\text{all } \mathcal{G}\text{-measurable random variables in } L_2\}$$

Fact The set $L_2(\mathcal{G})$ is a linear subspace of L_2

Proof: Pick $y_1, y_2 \in L_2(\mathcal{G})$ and $\alpha, \beta \in \mathbb{R}$

Let $z := \alpha y_1 + \beta y_2$

By definition, $y_i = g_i(x_1, \ldots, x_p)$

$$\therefore \quad z = \alpha g_1(x_1, \ldots, x_p) + \beta g_2(x_1, \ldots, x_p)$$

$$\therefore \quad z \in L_2(\mathcal{G})$$

Adding Information

Fact If $\mathcal{G} \subset \mathcal{H}$, then $L_2(\mathcal{G}) \subset L_2(\mathcal{H})$

Equivalent statement: If z is $\mathcal{G}$-measurable and $\mathcal{G} \subset \mathcal{H}$, then z is $\mathcal{H}$-measurable

Intuition: If z known once RVs in $\mathcal{G}$ known, then known when extra information provided by $\mathcal{H}$ is available

Example

Let $z = 2x + 3$, $\mathcal{G} = \{x\}$, $\mathcal{H} = \{x, y\}$

Here z is $\mathcal{G}$-measurable and also $\mathcal{H}$-measurable

Formally, we can write $z = g(x, y)$, where $g(x, y) = 2x + 3 + 0y$

Hence z is also $\mathcal{H}$-measurable as claimed

Conditional Expectations

Let $\mathcal{G} \subset L_2$ and let y be some RV in L_2

The **conditional expectation** of y given $\mathcal{G}$ is written as $\mathbb{E}[y \,|\, \mathcal{G}]$ and defined as the closest $\mathcal{G}$-measurable random variable to y

More formally,

$$\mathbb{E}[y \,|\, \mathcal{G}] := \operatorname*{argmin}_{z \in L_2(\mathcal{G})} \|\|y - z\|\|$$

Intuitively: $\mathbb{E}[y \,|\, \mathcal{G}] =$ best predictor of y given info contained in $\mathcal{G}$

But does it exist?

What properties does it have?

The Orthogonal Projection Theorem in L_2

The OPT in L_2 is almost identical to that for $\mathbb{R}^N$

Theorem. Given linear subspace S of L_2 and y in L_2, there is a unique $\hat{y} \in S$ such that

$$\| y - \hat{y} \| \leq \| y - z \| \quad \text{for all } z \in S$$

The RV $\hat{y}$ is called the **orthogonal projection** of y onto S

As for $\mathbb{R}^N$ case, $\hat{y}$ is the orthogonal projection of y onto S iff

1. $\hat{y} \in S$
2. $y - \hat{y} \perp S$

The L_2 OPT stated a different way:

Theorem. Given a linear subspace S of L_2, the function

$$\mathbf{P}y := \operatorname*{argmin}_{z \in S} |||y - z|||$$

is a well-defined linear function from L_2 to S

Given any $y \in L_2$, we have

1. $\mathbf{P}y \in S$
2. $y - \mathbf{P}y \perp S$
3. $\mathbf{P}y = y$ if and only if $y \in S$

To repeat:

Orthogonal projection onto arbitrary S is

$$\mathbf{P}y := \operatorname*{argmin}_{z \in S} |||y - z|||$$

Conditional expectation of y given $\mathcal{G}$ is

$$\mathbb{E}\left[y \,|\, \mathcal{G}\right] := \operatorname*{argmin}_{z \in L_2(\mathcal{G})} |||y - z|||$$

Thus, $y \mapsto \mathbb{E}\left[y \,|\, \mathcal{G}\right]$ is the map $y \mapsto \mathbf{P}y$ when $S = L_2(\mathcal{G})$

Since $\mathbb{E}[y \,|\, \mathcal{G}]$ is the orthogonal projection of y onto $L_2(\mathcal{G})$

- $\mathbb{E}[y \,|\, \mathcal{G}]$ exists, unique

Moreover, $\mathbb{E}[y \,|\, \mathcal{G}]$ is the unique point in L_2 such that

- $\mathbb{E}[y \,|\, \mathcal{G}] \in L_2(\mathcal{G})$
- $y - \mathbb{E}[y \,|\, \mathcal{G}] \perp z$ for all $z \in L_2(\mathcal{G})$

Restatement leads to our second def of conditional expectation:

Def. $\mathbb{E}[y \,|\, \mathcal{G}]$ is the unique element of L_2 such that

1. $\mathbb{E}[y \,|\, \mathcal{G}]$ is $\mathcal{G}$-measurable
2. $\mathbb{E}[\mathbb{E}[y \,|\, \mathcal{G}]\, z] = \mathbb{E}[yz]$ for all $\mathcal{G}$-measurable $z \in L_2$

We will also use the common notation

$$\mathbb{E}\left[y \,|\, x_1, \ldots, x_p\right] := \mathbb{E}\left[y \,|\, \mathcal{G}\right]$$

Also, let's record the following "obvious" fact:

Fact Given $\{x_1, \ldots, x_p\}$ and y in L_2, there exists a function $g \colon \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\left[y \,|\, x_1, \ldots, x_p\right] = g(x_1, \ldots, x_p)$

Why is this true?

Example

If x, w independent and $y = x + w$, then $\mathbb{E}[y \mid x] = x + \mathbb{E}[w]$.

To check that $h(x) := x + \mathbb{E}[w]$ is $\mathbb{E}[y \mid x]$, must show that

1. $h(x)$ is x-measurable
2. $\mathbb{E}[h(x)z] = \mathbb{E}[yz]$ for any x-measurable RV z

Part 1 is obvious

Part 2 translates to the claim that

$$\mathbb{E}[(x + \mathbb{E}[w])g(x)] = \mathbb{E}[(x + w)g(x)] \quad \text{for any function } g$$

Exercise: Check that this equality holds

Example

If x and y are random variables and $p(y \mid x)$ is the conditional density of y given x, then

$$\mathbb{E}\left[y \mid x\right] = \int t p(t \mid x) dt$$

This is a solved exercise in the course notes

Fact. Let x and y be RVs in L_2, let α and β be scalars, and let $\mathcal{G}$ and $\mathcal{H}$ be subsets of L_2. The following properties hold:

1. Linearity: $\mathbb{E}\left[\alpha x + \beta y \,|\, \mathcal{G}\right] = \alpha \mathbb{E}\left[x \,|\, \mathcal{G}\right] + \beta \mathbb{E}\left[y \,|\, \mathcal{G}\right]$
2. If $\mathcal{G} \subset \mathcal{H}$, then
$$\mathbb{E}\left[\mathbb{E}\left[y \,|\, \mathcal{H}\right] \,|\, \mathcal{G}\right] = \mathbb{E}\left[y \,|\, \mathcal{G}\right] \quad \text{and} \quad \mathbb{E}\left[\mathbb{E}\left[y \,|\, \mathcal{G}\right]\right] = \mathbb{E}\left[y\right]$$
3. If y is independent of the variables in $\mathcal{G}$, then $\mathbb{E}\left[y \,|\, \mathcal{G}\right] = \mathbb{E}\left[y\right]$
4. If y is $\mathcal{G}$-measurable, then $\mathbb{E}\left[y \,|\, \mathcal{G}\right] = y$
5. If x is $\mathcal{G}$-measurable, then $\mathbb{E}\left[xy \,|\, \mathcal{G}\right] = x\mathbb{E}\left[y \,|\, \mathcal{G}\right]$

Most of these follow directly from the L_2 OPT

The Vector/Matrix Case

Given random matrices $\mathbf{X}$ and $\mathbf{Y}$, we set

$$\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}\right] := \begin{pmatrix} \mathbb{E}\left[y_{11} \,|\, \mathbf{X}\right] & \cdots & \mathbb{E}\left[y_{1K} \,|\, \mathbf{X}\right] \\ \mathbb{E}\left[y_{21} \,|\, \mathbf{X}\right] & \cdots & \mathbb{E}\left[y_{2K} \,|\, \mathbf{X}\right] \\ \vdots & & \vdots \\ \mathbb{E}\left[y_{N1} \,|\, \mathbf{X}\right] & \cdots & \mathbb{E}\left[y_{NK} \,|\, \mathbf{X}\right] \end{pmatrix}$$

where

$$\mathbb{E}\left[y_{nk} \,|\, \mathbf{X}\right] := \mathbb{E}\left[y_{nk} \,|\, x_{11}, \ldots, x_{\ell m}, \ldots, x_{LM}\right]$$

Also,

- $\mathrm{cov}[\mathbf{x}, \mathbf{y} \,|\, \mathbf{Z}] := \mathbb{E}\left[\mathbf{x}\mathbf{y}' \,|\, \mathbf{Z}\right] - \mathbb{E}\left[\mathbf{x} \,|\, \mathbf{Z}\right]\mathbb{E}\left[\mathbf{y} \,|\, \mathbf{Z}\right]'$
- $\mathrm{var}[\mathbf{x} \,|\, \mathbf{Z}] := \mathbb{E}\left[\mathbf{x}\mathbf{x}' \,|\, \mathbf{Z}\right] - \mathbb{E}\left[\mathbf{x} \,|\, \mathbf{Z}\right]\mathbb{E}\left[\mathbf{x} \,|\, \mathbf{Z}\right]'$

Working from the definition and facts for scalar case, we obtain

Fact. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be random matrices, and let $\mathbf{A}$ and $\mathbf{B}$ be constant matrices. Assuming conformability,

1. $\mathbb{E}\left[\mathbf{AX} + \mathbf{BY} \,|\, \mathbf{Z}\right] = \mathbf{A}\mathbb{E}\left[\mathbf{X} \,|\, \mathbf{Z}\right] + \mathbf{B}\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{Z}\right]$
2. $\mathbb{E}\left[\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}\right]\right] = \mathbb{E}\left[\mathbf{Y}\right]$ and $\mathbb{E}\left[\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}, \mathbf{Z}\right] \,|\, \mathbf{X}\right] = \mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}\right]$
3. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}\right] = \mathbb{E}\left[\mathbf{Y}\right]$
4. If g is a (nonrandom) function, then

$$\mathbb{E}\left[g(\mathbf{X})\,\mathbf{Y} \,|\, \mathbf{X}\right] = g(\mathbf{X})\mathbb{E}\left[\mathbf{Y} \,|\, \mathbf{X}\right] \quad \text{and} \quad \mathbb{E}\left[g(\mathbf{X}) \,|\, \mathbf{X}\right] = g(\mathbf{X})$$

Prelude to Overdetermined Systems

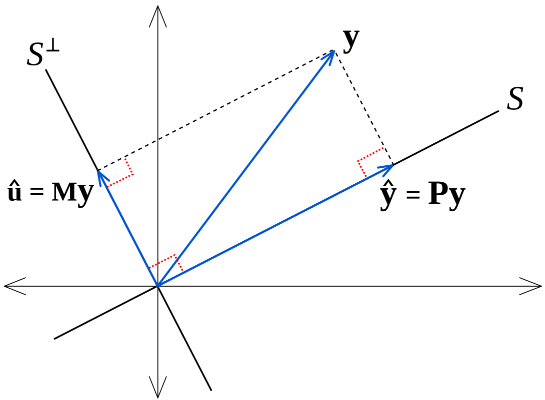
Recall the OPT Mark III

Projecting $\mathbf{y} \in \mathbb{R}^N$ onto linear subspace S

$\mathbf{Py}$ is the projection onto S

$\mathbf{My}$ is the projection onto $S^\perp$

$\mathbf{y} = \mathbf{Py} + \mathbf{My}$ and $\mathbf{Py} \perp \mathbf{My}$



$S^{\perp}$
$\mathbf{y}$
S
$\hat{\mathbf{u}} = \mathbf{My}$
$\hat{\mathbf{y}} = \mathbf{Py}$

Let the linear subspace S be given

We know that

- The orthogonal projection mapping $\mathbf{P}$ onto S is a linear function from $\mathbb{R}^N$ to $\mathbb{R}^N$
- If $f \colon \mathbb{R}^N \to \mathbb{R}^N$ is linear, then there exists an $N \times N$ matrix $\mathbf{C}$ such that $f(\mathbf{y}) = \mathbf{C}\mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^N$

Hence, exists an $N \times N$ matrix $\mathbf{C}^S$ such that $\mathbf{P}\mathbf{y} = \mathbf{C}^S\mathbf{y}$ for all $\mathbf{y}$

To put it more simply, $\mathbf{P}$ "is" a matrix

Let's calculate $\mathbf{P}$ in a specific case of interest

Let $\mathbf{X}$ be $N \times K$ with full column rank

Question: If $S = \mathrm{rng}(\mathbf{X})$, then what do $\mathbf{P}$ and $\mathbf{M}$ look like?

Answer: If $S = \mathrm{rng}(\mathbf{X})$, then

- $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- $\mathbf{M} = \mathbf{I} - \mathbf{P}$

Notation:

- $\mathbf{P}$ is called the **projection matrix** associated with $\mathbf{X}$
- $\mathbf{M}$ is called the **annihilator**

We now prove the claim that $\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Proof for $\mathbf{M}$ is an exercise

Given arbitrary $\mathbf{y} \in \mathbb{R}^N$, our claim is that

1. $\mathbf{P}\mathbf{y} \in \mathrm{rng}(\mathbf{X})$, and
2. $\mathbf{y} - \mathbf{P}\mathbf{y} \perp \mathrm{rng}(\mathbf{X})$

Here 1 is true because

$$\mathbf{P}\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{a} \quad \text{when} \quad \mathbf{a} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

On the other hand, 2 is equivalent to the statement

$$\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \perp \mathbf{X}\mathbf{b} \quad \text{for all} \quad \mathbf{b} \in \mathbb{R}^K$$

This is true: If $\mathbf{b} \in \mathbb{R}^K$, then

$$(\mathbf{X}\mathbf{b})'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \mathbf{b}'[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y}] = 0$$

Exercises: Show that

- $\mathbf{P}$ and $\mathbf{M}$ are both idempotent and symmetric
- The annihilator $\mathbf{M}$ satisfies $\mathbf{MX} = \mathbf{0}$

Overdetermined Systems of Equations

Consider system of equations $\mathbf{X}\mathbf{b} = \mathbf{y}$

- $\mathbf{X}$ is $N \times K$
- $\mathbf{b}$ is $K \times 1$
- $\mathbf{y}$ is $N \times 1$

Taking $\mathbf{X}$ and $\mathbf{y}$ as given, we seek $\mathbf{b} \in \mathbb{R}^K$ such that $\mathbf{X}\mathbf{b} = \mathbf{y}$

Assumption: $\mathbf{X}$ is full column rank

If $K = N$, then system has precisely one solution

We are going to study the case when $N > K$

Put differently:

- number of equations $>$ number of unknowns
- number of constraints $>$ degrees of freedom

In this case, system of equations said to be **overdetermined**

May not be able find a $\mathbf{b}$ that satisfies all N equations

To understand problem, recall that

$$\mathrm{rng}(\mathbf{X}) := \text{column space of } \mathbf{X} := \{\text{all } \mathbf{Xb} \text{ with } \mathbf{b} \in \mathbb{R}^K\}$$

Solution to $\mathbf{Xb} = \mathbf{y}$ exists precisely when $\mathbf{y} \in \mathrm{rng}(\mathbf{X})$

When $K < N$ this is "unlikely" because

- $\mathbf{y}$ is an arbitrary point in $\mathbb{R}^N$
- $\mathrm{rng}(\mathbf{X})$ has dimension K
- K-dim subspace has "Lebesgue measure zero" in $\mathbb{R}^N$ whenever $K < N$

If system $\mathbf{X}\mathbf{b} = \mathbf{y}$ is overdetermined, what do people do?

Answer:

1. Accept that an exact solution may not exist
2. Look instead for an approximate solution

Method: Find $\mathbf{b} \in \mathbb{R}^K$ such that $\mathbf{X}\mathbf{b}$ is as close to $\mathbf{y}$ as possible

Mathematically: Choose

$$\hat{\beta} := \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$$

Thm. The minimizer of $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|$ over $\mathbf{b} \in \mathbb{R}^K$ is

$$\hat{\beta} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Proof: Note that

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

Since $\mathbf{P}\mathbf{y}$ is the orthogonal projection onto $\operatorname{rng}(\mathbf{X})$ we have

$$\|\mathbf{y} - \mathbf{P}\mathbf{y}\| \leq \|\mathbf{y} - \mathbf{z}\| \text{ for any } \mathbf{z} \in \operatorname{rng}(\mathbf{X})$$

In other words,

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\| \leq \|\mathbf{y} - \mathbf{X}\mathbf{b}\| \text{ for any } \mathbf{b} \in \mathbb{R}^K$$

as was to be shown

Linear Least Squares Regression

We observe input $\mathbf{x} \in \mathbb{R}^K$ followed by scalar output y

Assume: The pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ are IID from some common joint distribution on $\mathbb{R}^{K+1}$

This distribution is unknown to us

Aim: Choose $f \colon \mathbb{R}^K \to \mathbb{R}$ such that $f(\mathbf{x})$ is a good predictor of y

"Goodness" measured by quadratic loss, so risk of f is

$$R(f) := \mathbb{E}\left[(y - f(\mathbf{x}))^2\right]$$

Exercise: Show that the risk minimizer is $f^*(\mathbf{x}) = \mathbb{E}\left[y \,|\, \mathbf{x}\right]$

Using principle of ERM, replace risk function with empirical risk:

$$\min_{f \in \mathcal{F}} \hat{R}(f) \quad \text{where} \quad \hat{R}(f) := \frac{1}{N} \sum_{n=1}^{N} (y_n - f(\mathbf{x}_n))^2$$

As before, $\mathcal{F}$ is called the hypothesis space

We now consider the case where $\mathcal{F}$ is the linear functions:

$$\mathcal{F} = \mathcal{L} := \{ \text{ all functions } \ell(\mathbf{x}) = \mathbf{b}'\mathbf{x} \text{ for some } \mathbf{b} \in \mathbb{R}^K \}$$

Dropping constant $1/N$, the ERM problem is then

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{b}'\mathbf{x}_n)^2$$

To solve this problem, we switch to matrix notation, with

$$\mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{x}_n := \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix} = n\text{-th obs on all regressors}$$

and

$$\mathbf{X} := \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} =: \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}$$

We assume throughout that $N > K$ and $\mathbf{X}$ is full column rank

Exercise: Verify that

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = \sum_{n=1}^{N} (y_n - \mathbf{b}'\mathbf{x}_n)^2$$

Since increasing transforms don't affect minimizers we have

$$\operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \sum_{n=1}^{N} (y_n - \mathbf{b}'\mathbf{x}_n)^2 = \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^K} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|$$

By the theory of overdetermined systems, the solution is

$$\hat{\beta} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Notation

Let $\mathbf{P}$ and $\mathbf{M}$ be the projection and annihilator associated with $\mathbf{X}$:

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{and} \quad \mathbf{M} := \mathbf{I} - \mathbf{P}$$

The **vector of fitted values** is

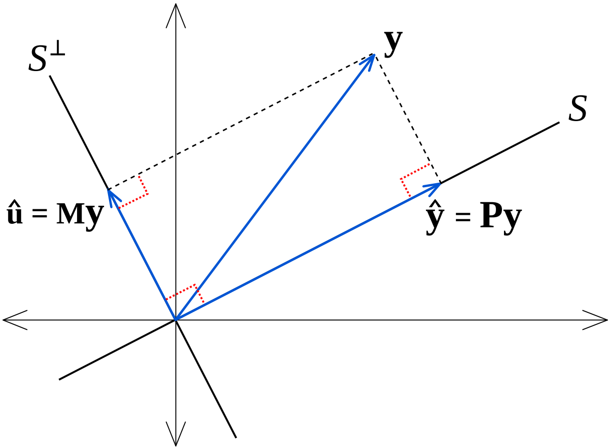
$$\hat{\mathbf{y}} := \mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$$

The **vector of residuals** is

$$\hat{\mathbf{u}} := \mathbf{M}\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}}$$

Applying the OPT we obtain

$$\mathbf{M}\mathbf{y} \perp \mathbf{P}\mathbf{y} \quad \text{and} \quad \mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$$


$S^{\perp}$
$\mathbf{y}$
S
$\hat{\mathbf{u}} = \mathbf{My}$
$\hat{\mathbf{y}} = \mathbf{Py}$

More standard definitions:

- **Total sum of squares** $=: \mathrm{TSS} := \|\mathbf{y}\|^2$.
- **Sum of squared residuals** $=: \mathrm{SSR} := \|\mathbf{M}\mathbf{y}\|^2$.
- **Explained sum of squares** $=: \mathrm{ESS} := \|\mathbf{P}\mathbf{y}\|^2$.

Exercise: Show that $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$

Transformations of the Data

How about the assumption $\mathcal{F} = \mathcal{L}$?

- The best predictor is the risk minimizer $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$
- Setting $\mathcal{F} = \mathcal{L}$ is good if $f^* \in \mathcal{L}$

We have a good chance of approximating it well with ERM over $\mathcal{L}$

Here $f^* \in \mathcal{L}$ means that

$$f^*(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \quad \text{for some } \boldsymbol{\beta} \in \mathbb{R}^K$$

But what if this is not true?

What if the system is nonlinear?

Instead of linearity, let's suppose that

- $f^*(\mathbf{x}) = \ell(\boldsymbol{\phi}(\mathbf{x})) = \gamma' \boldsymbol{\phi}(\mathbf{x})$ for some (possibly nonlinear) function $\boldsymbol{\phi} \colon \mathbb{R}^K \to \mathbb{R}^J$
- The function $\boldsymbol{\phi}$ is known, but γ is not

How to estimate γ?

We proceed as before, but regressing y on $\boldsymbol{\phi}(\mathbf{x})$

That is, we replace the data

$$(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$$

with

$$(y_1, \boldsymbol{\phi}_1), \ldots, (y_N, \boldsymbol{\phi}_N) \quad \text{where} \quad \boldsymbol{\phi}_n := \boldsymbol{\phi}(\mathbf{x}_n)$$

Example

Taking logs of the data:

$$\begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix} = \mathbf{x}_n \mapsto \boldsymbol{\phi}_n = \begin{pmatrix} \ln x_{n1} \\ \ln x_{n2} \end{pmatrix}$$

Example

Including cross products

$$\begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix} = \mathbf{x}_n \mapsto \boldsymbol{\phi}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ x_{n1}x_{n2} \end{pmatrix}$$

Example

Suppose that $K = 1$ and

$$x_n \mapsto \boldsymbol{\phi}_n = \begin{pmatrix} x_n^0 \\ x_n^1 \\ \vdots \\ x_n^{J-1} \end{pmatrix}$$

Then $\gamma' \boldsymbol{\phi}_n = \sum_{j=1}^{J} \gamma_j x_n^{j-1}$

Corresponds to univariate polynomial regression

Weierstrass: Given continuous function f, there exists a polynomial function g such that g is arbitrarily close to f

Intuition: If we take J large enough, we can approximate almost any nonlinear relationship we want

Applying linear least squares to the transformed data:

The empirical risk minimization problem is

$$\min_{\gamma \in \mathbb{R}^J} \sum_{n=1}^{N} (y_n - \gamma' \boldsymbol{\phi}_n)^2$$

Switching to matrix notation, let

$$\boldsymbol{\Phi} := \begin{pmatrix} \boldsymbol{\phi}_1' \\ \boldsymbol{\phi}_2' \\ \vdots \\ \boldsymbol{\phi}_N' \end{pmatrix} \in \mathbb{R}^{N \times J}$$

Switching into matrix form, the objective function is

$$\sum_{n=1}^{N} (y_n - \gamma' \boldsymbol{\phi}_n)^2 = \|\mathbf{y} - \boldsymbol{\Phi}\gamma\|^2$$

Since increasing functions don't affect minimizers,

$$\operatorname*{argmin}_{\gamma \in \mathbb{R}^J} \sum_{n=1}^{N} (y_n - \gamma' \boldsymbol{\phi}_n)^2 = \operatorname*{argmin}_{\gamma \in \mathbb{R}^J} \|\mathbf{y} - \boldsymbol{\Phi}\gamma\|$$

Assuming that $\boldsymbol{\Phi}$ is full column rank, the solution is

$$\hat{\gamma} := (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{y}$$

Example

To add an intercept to the regression, use transformation

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_K \end{pmatrix}$$

Exercise: In this case,

- the vector of residuals must sum to zero
- the mean of the fitted values equals the mean of $\mathbf{y}$

Work through it in the course notes

In most of what follows, we don't discuss transformations explicitly

- regress y on $\mathbf{x}$, not on $\boldsymbol{\phi}(\mathbf{x})$

No loss of generality is entailed:

We can just imagine that the data has already been transformed, and $\mathbf{x}$ is the result

Hence we use

- $\mathbf{X}$ to denote the data matrix instead of Φ
- $\hat{\boldsymbol{\beta}}$ to denote the least squares estimator $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$