Conditional Expectation

Stats 220A

Motivation: “best guess given information”

We have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-field $\mathcal{G} \subseteq \mathcal{F}$.

Think of $\mathcal{G}$ as “information revealed” (e.g., what we have observed).

Goal: define $\mathbb{E}[X \mid \mathcal{G}]$ for $X \in L^1(\mathcal{F})$ so that

- it is $\mathcal{G}$-measurable (uses only information in $\mathcal{G}$),
- for every event $A \in \mathcal{G}$, it has the same average as X on A.

Definition (measure-theoretic)

Definition

Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$. A random variable Y is a version of $\mathbb{E}[X \mid \mathcal{G}]$ if

1. Y is $\mathcal{G}$-measurable, and
2. for all $A \in \mathcal{G}$,

$$\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

Notes:

- Defined up to a.s. equality.
- This is an *integral identity*, not a pointwise formula.

Two Fundamental Questions

We have defined $\mathbb{E}[X \mid \mathcal{G}]$ through an integral identity.

Natural questions:

- Does such a random variable Y actually exist?
- If it exists, is it unique?

We now show:

- Existence follows from the Radon–Nikodym theorem.
- Uniqueness holds up to almost sure equality.

Absolute Continuity and σ-Finiteness

Definition (Absolute Continuity)

Let ν and μ be measures on $(\Omega, \mathcal{F})$.
We say

$$\nu \ll \mu$$

if

$$\mu(A) = 0 \;\Rightarrow\; \nu(A) = 0 \quad \text{for all } A \in \mathcal{F}.$$

Definition (σ-finite measure)

A measure μ is called σ-finite if

$$\Omega = \bigcup_{n=1}^{\infty} A_n \quad \text{with} \quad \mu(A_n) < \infty.$$

Remark. Since $\mathbb{P}(\Omega) = 1 < \infty$, the probability measure $\mathbb{P}$ is automatically σ-finite.

Radon–Nikodym Theorem

Radon–Nikodym Theorem.
Let ν and μ be σ-finite measures on $(\Omega, \mathcal{G})$. If $\nu \ll \mu$, then there exists a $\mathcal{G}$-measurable function f such that

$$\nu(A) = \int_A f \, d\mu \quad \forall A \in \mathcal{G}.$$

The function f is called the Radon–Nikodym derivative.

Existence: Case $X \geq 0$

Assume $X \geq 0$ and $X \in L^1$.
Define a set function ν on $(\Omega, \mathcal{G})$ by

$$\nu(A) = \int_A X \, d\mathbb{P}, \qquad A \in \mathcal{G}.$$

Step 1: ν is a measure on $(\Omega, \mathcal{G})$.
If (A_n) are disjoint in $\mathcal{G}$, then by monotone convergence,

$$\nu\left(\bigcup_n A_n\right) = \sum_n \nu(A_n).$$

Step 2: $\nu \ll \mathbb{P}|_{\mathcal{G}}$.
If $\mathbb{P}(A) = 0$, then $\nu(A) = 0$.
Notation. $\mathbb{P}|_{\mathcal{G}}$ denotes the restriction of $\mathbb{P}$ to the σ-field $\mathcal{G}$, i.e.

$$\mathbb{P}|_{\mathcal{G}}(A) = \mathbb{P}(A), \quad A \in \mathcal{G}.$$

Apply Radon–Nikodym

Apply Radon–Nikodym with

$$\mu = \mathbb{P}|_{\mathcal{G}}, \qquad \nu(A) = \int_A X \, d\mathbb{P}.$$

There exists a $\mathcal{G}$-measurable function $Y \geq 0$ such that

$$\nu(A) = \int_A Y \, d\mathbb{P} \quad \forall A \in \mathcal{G}.$$

Thus

$$\int_A X \, d\mathbb{P} = \int_A Y \, d\mathbb{P}.$$

Taking $A = \Omega$ shows Y is integrable.
So Y is a version of $\mathbb{E}[X \mid \mathcal{G}]$ for $X \geq 0$.

Existence: General Case

Let $X \in L^1$.
Write
$$X = X^+ - X^-,$$
where $X^+, X^- \geq 0$ and integrable.
Define
$$Y_1 = \mathbb{E}[X^+ \mid \mathcal{G}], \qquad Y_2 = \mathbb{E}[X^- \mid \mathcal{G}].$$
Set
$$Y = Y_1 - Y_2.$$
Then for all $A \in \mathcal{G}$,
$$\int_A X \, d\mathbb{P} = \int_A Y \, d\mathbb{P}.$$
Thus $\mathbb{E}[X \mid \mathcal{G}]$ exists for all $X \in L^1$.

Uniqueness

Proposition

If Y_1, Y_2 satisfy the defining identity, then $Y_1 = Y_2$ a.s.

Proof.

Let $D = Y_1 - Y_2$.
For all $A \in \mathcal{G}$,

$$\int_A D \, d\mathbb{P} = 0.$$

Let $A_\varepsilon = \{D \geq \varepsilon\} \in \mathcal{G}$.
Then

$$0 = \int_{A_\varepsilon} D \, d\mathbb{P} \geq \varepsilon \mathbb{P}(A_\varepsilon).$$

Thus $\mathbb{P}(D > 0) = 0$ and similarly $\mathbb{P}(D < 0) = 0$. $\square$

First immediate consequence: law of total expectation

Proposition

For $X \in L^1(\mathcal{F})$,

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}]\big] = \mathbb{E}[X].$$

Proof.

Take $A = \Omega$ in the defining identity:

$$\int_\Omega \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P}.$$

$\square$

Interpretation: conditioning does not change the overall mean; it redistributes it across $\mathcal{G}$-events.

Example 1: conditioning on an event

Let $B \in \mathcal{F}$ with $0 < \mathbb{P}(B) < 1$ and $\mathcal{G} = \sigma(B) = \{\varnothing, B, B^c, \Omega\}$.

Proposition

For $X \in L^1$,

$$\mathbb{E}[X \mid \sigma(B)] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}\,\mathbf{1}_B + \frac{\mathbb{E}[X\mathbf{1}_{B^c}]}{\mathbb{P}(B^c)}\,\mathbf{1}_{B^c}.$$

Proof.

Let Y be the RHS. Then Y is $\sigma(B)$-measurable (constant on B and B^c). Check the defining identity for $A = B$ and $A = B^c$:

$$\int_B Y\,d\mathbb{P} = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}\,\mathbb{P}(B) = \mathbb{E}[X\mathbf{1}_B], \qquad \int_{B^c} Y\,d\mathbb{P} = \mathbb{E}[X\mathbf{1}_{B^c}],$$

and the identity follows for all $A \in \sigma(B)$. $\square$

Example 2: finite partition (general discrete information)

Let $\{A_1, \ldots, A_n\}$ be a partition of Ω with $\mathbb{P}(A_i) > 0$ and $\mathcal{G} = \sigma(A_1, \ldots, A_n)$.

Proposition

For $X \in L^1$,

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i]\, \mathbf{1}_{A_i} = \sum_{i=1}^{n} \frac{\mathbb{E}[X \mathbf{1}_{A_i}]}{\mathbb{P}(A_i)}\, \mathbf{1}_{A_i}.$$

Proof.

Same structure as the previous proof: define Y to be the RHS, note Y is $\mathcal{G}$-measurable, and verify the defining identity on generators $A = A_i$ (then extend by additivity over unions). □

Example 3: independence makes conditioning trivial

Proposition

If $X \in L^1$ is independent of $\mathcal{G}$, then

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X] \quad a.s.$$

Proof.

Let $Y \equiv \mathbb{E}[X]$ (constant, hence $\mathcal{G}$-measurable). For any $A \in \mathcal{G}$, independence gives

$$\int_A X \, d\mathbb{P} = \mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_A] = \mathbb{E}[X]\mathbb{P}(A) = \int_A Y \, d\mathbb{P}.$$

So Y satisfies the definition; uniqueness gives the claim. $\square$

Example 4: conditioning on a random variable

Define

$$\mathbb{E}[X \mid Y] := \mathbb{E}[X \mid \sigma(Y)].$$

A key fact: $\mathbb{E}[X \mid Y]$ is a measurable function of Y.

Theorem (Doob–Dynkin lemma (informal statement))

If Z is $\sigma(Y)$-measurable, then there exists a Borel measurable g such that $Z = g(Y)$ a.s.

So there exists g with

$$\mathbb{E}[X \mid Y] = g(Y) \quad \text{a.s.}$$

We will *use* this fact rather than prove it fully.

Concrete computation: independent sum

Let U, V be independent with $\mathbb{E}|V| < \infty$. Set $X = U + V$ and condition on U.

Proposition

$$\mathbb{E}[U + V \mid U] = U + \mathbb{E}[V] \quad a.s.$$

Proof.
Since U is $\sigma(U)$-measurable,

$$\mathbb{E}[U + V \mid U] = \mathbb{E}[U \mid U] + \mathbb{E}[V \mid U] = U + \mathbb{E}[V],$$

where $\mathbb{E}[V \mid U] = \mathbb{E}[V]$ because V is independent of $\sigma(U)$. $\square$

Concrete computation: bivariate normal (classic)

Let (X, Y) be jointly normal with means 0, variances 1, and correlation ρ.
Then:

$$\mathbb{E}[X \mid Y] = \rho\, Y, \qquad \mathrm{Var}(X \mid Y) = 1 - \rho^2.$$

Interpretation: conditioning on Y gives the best linear predictor, and for Gaussians it is the best predictor among all measurable functions.

Linearity and monotonicity

Proposition (Linearity)

If $X_1, X_2 \in L^1$ and $a, b \in \mathbb{R}$, then

$$\mathbb{E}[aX_1 + bX_2 \mid \mathcal{G}] = a\mathbb{E}[X_1 \mid \mathcal{G}] + b\mathbb{E}[X_2 \mid \mathcal{G}].$$

Proof.

Let $Y := a\mathbb{E}[X_1 \mid \mathcal{G}] + b\mathbb{E}[X_2 \mid \mathcal{G}]$. Then Y is $\mathcal{G}$-measurable. For any $A \in \mathcal{G}$,

$$\int_A Y \, d\mathbb{P} = a \int_A \mathbb{E}[X_1 \mid \mathcal{G}] \, d\mathbb{P} + b \int_A \mathbb{E}[X_2 \mid \mathcal{G}] \, d\mathbb{P} = a \int_A X_1 \, d\mathbb{P} + b \int_A X_2 \, d\mathbb{P} = \int_A (aX_1 + bX_2) \, d\mathbb{P}.$$

So Y satisfies the definition. $\square$

Proposition (Monotonicity)

If $X \leq Y$ a.s., then $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}]$ a.s.

Monotonicity (proof)

Proof.
Let $D := \mathbb{E}[X \mid \mathcal{G}] - \mathbb{E}[Y \mid \mathcal{G}]$ (which is $\mathcal{G}$-measurable). For $A := \{D > 0\} \in \mathcal{G}$ we have

$$\int_A D\, d\mathbb{P} = \int_A \mathbb{E}[X \mid \mathcal{G}]\, d\mathbb{P} - \int_A \mathbb{E}[Y \mid \mathcal{G}]\, d\mathbb{P} = \int_A X\, d\mathbb{P} - \int_A Y\, d\mathbb{P} \leq 0.$$

But on A, $D > 0$, hence $\int_A D\, d\mathbb{P} > 0$ unless $\mathbb{P}(A) = 0$. Therefore $\mathbb{P}(D > 0) = 0$, i.e. $D \leq 0$ a.s. $\square$

Tower property

Proposition (Tower)

If $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ and $X \in L^1$, then

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}\big] = \mathbb{E}[X \mid \mathcal{H}] \quad a.s.$$

Proof.

Let $Y := \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]$. For any $A \in \mathcal{H}$,

$$\int_A Y \, d\mathbb{P} = \int_A \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_A X \, d\mathbb{P},$$

since $A \in \mathcal{H} \subseteq \mathcal{G}$. Thus Y satisfies the defining identity for conditioning on $\mathcal{H}$. $\square$

Pull-out property: statement

We want the workhorse identity:

$$\mathbb{E}[ZX \mid \mathcal{G}] = Z\,\mathbb{E}[X \mid \mathcal{G}]$$

when Z is $\mathcal{G}$-measurable.
We prove it in steps:

1. $Z = \mathbf{1}_B$ with $B \in \mathcal{G}$
2. Z simple $\mathcal{G}$-measurable
3. $Z \geq 0$ $\mathcal{G}$-measurable (monotone limit)
4. integrable Z (positive/negative parts)

Pull-out property: proof (indicator and simple)

Proposition (Pull-out, bounded case)

If Z is bounded and $\mathcal{G}$-measurable and $X \in L^1$, then

$$\mathbb{E}[ZX \mid \mathcal{G}] = Z\,\mathbb{E}[X \mid \mathcal{G}] \quad a.s.$$

Proof.

Step 1: $Z = \mathbf{1}_B$, $B \in \mathcal{G}$. Let $A \in \mathcal{G}$. Then

$$\int_A \mathbf{1}_B \mathbb{E}[X \mid \mathcal{G}]\,d\mathbb{P} = \int_{A\cap B} \mathbb{E}[X \mid \mathcal{G}]\,d\mathbb{P} = \int_{A\cap B} X\,d\mathbb{P} = \int_A \mathbf{1}_B X\,d\mathbb{P}.$$

So $\mathbb{E}[\mathbf{1}_B X \mid \mathcal{G}] = \mathbf{1}_B \mathbb{E}[X \mid \mathcal{G}]$.

Step 2: $Z = \sum_{i=1}^m c_i \mathbf{1}_{B_i}$ simple $\mathcal{G}$-measurable. Use linearity and Step 1:

$$\mathbb{E}[ZX \mid \mathcal{G}] = \sum_i c_i \mathbb{E}[\mathbf{1}_{B_i} X \mid \mathcal{G}] = \sum_i c_i \mathbf{1}_{B_i} \mathbb{E}[X \mid \mathcal{G}] = Z\,\mathbb{E}[X \mid \mathcal{G}].$$

Pull-out property: proof (bounded via approximation)

Proof (continued).

Step 3: bounded $\mathcal{G}$-measurable Z. Approximate Z by simple functions Z_n with $Z_n \to Z$ a.s. and $|Z_n| \le \|Z\|_\infty$.
Then $Z_n X \to ZX$ a.s. and $|Z_n X| \le \|Z\|_\infty |X|$ which is integrable.
By dominated convergence,

$$\int_A Z_n X \, d\mathbb{P} \to \int_A ZX \, d\mathbb{P} \quad \text{for all } A \in \mathcal{G}.$$

Also $Z_n \mathbb{E}[X \mid \mathcal{G}] \to Z\mathbb{E}[X \mid \mathcal{G}]$ a.s. and is dominated by $\|Z\|_\infty |\mathbb{E}[X \mid \mathcal{G}]|$ with $\mathbb{E}|\mathbb{E}[X \mid \mathcal{G}]| \le \mathbb{E}|X| < \infty$. So again dominated convergence gives

$$\int_A Z_n \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} \to \int_A Z\mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P}.$$

Since equality holds for each n, it holds in the limit. Uniqueness gives the result. $\square$

L^2 orthogonality (full proof)

Assume $X \in L^2$. Let $Y = \mathbb{E}[X \mid \mathcal{G}]$.

Proposition (Orthogonality)

For every bounded $\mathcal{G}$-measurable Z,

$$\mathbb{E}[(X - Y)Z] = 0.$$

Proof.

Since Z is $\mathcal{G}$-measurable and bounded, $(X - Y)Z \in L^1$. By the tower property and pull-out,

$$\mathbb{E}[(X - Y)Z] = \mathbb{E}\big[\,\mathbb{E}[(X - Y)Z \mid \mathcal{G}]\,\big] = \mathbb{E}\big[\,Z\,\mathbb{E}[X - Y \mid \mathcal{G}]\,\big] = \mathbb{E}\big[\,Z\,(Y - Y)\,\big] = 0.$$

$\square$

Projection theorem in L^2

Theorem (Projection property)

Let $X \in L^2$ and $Y = \mathbb{E}[X \mid \mathcal{G}]$. Then for any $Z \in L^2(\mathcal{G})$,

$$\mathbb{E}[(X - Y)(Z - Y)] = 0 \quad \textit{and} \quad \mathbb{E}[(X - Z)^2] = \mathbb{E}[(X - Y)^2] + \mathbb{E}[(Y - Z)^2].$$

In particular, Y minimizes $\mathbb{E}[(X - Z)^2]$ over $Z \in L^2(\mathcal{G})$.

Proof.

First, $(Z - Y) \in L^2(\mathcal{G})$, so approximate $(Z - Y)$ in L^2 by bounded $\mathcal{G}$-measurable functions and apply the orthogonality proposition (density argument). Then expand

$$(X - Z) = (X - Y) + (Y - Z)$$

and square:

$$(X - Z)^2 = (X - Y)^2 + (Y - Z)^2 + 2(X - Y)(Y - Z).$$

Take expectations; the cross term is zero by orthogonality.

Law of total variance

Theorem (Total variance)

If $X \in L^2$ then

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid \mathcal{G})] + \mathrm{Var}(\mathbb{E}[X \mid \mathcal{G}]).$$

Proof.

Let $Y = \mathbb{E}[X \mid \mathcal{G}]$ and note $\mathbb{E}[Y] = \mathbb{E}[X]$.
Write

$$X - \mathbb{E}[X] = (X - Y) + (Y - \mathbb{E}[Y]).$$

Square and take expectations:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - Y)^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - Y)(Y - \mathbb{E}[Y])].$$

The cross term is 0 by orthogonality with $Z = (Y - \mathbb{E}[Y])$ (which is $\mathcal{G}$-measurable and in L^2). Finally, note $\mathbb{E}[(X - Y)^2] = \mathbb{E}[\mathbb{E}[(X - Y)^2 \mid \mathcal{G}]] = \mathbb{E}[\mathrm{Var}(X \mid \mathcal{G})]$ by definition of conditional variance. $\square$

Conditional Jensen

Theorem (Conditional Jensen)

If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $X \in L^1$ with $\varphi(X) \in L^1$, then

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[\varphi(X) \mid \mathcal{G}] \quad a.s.$$

Proof idea (standard). A convex φ can be written as a supremum of affine functions:

$$\varphi(x) = \sup_{t \in T}\{a_t x + b_t\}.$$

Then for each t,

$$a_t \mathbb{E}[X \mid \mathcal{G}] + b_t = \mathbb{E}[a_t X + b_t \mid \mathcal{G}] \leq \mathbb{E}[\varphi(X) \mid \mathcal{G}].$$

Taking $\sup_t$ preserves $\leq$ and yields Jensen.

L^p contraction

Theorem (L^p contraction)

For $p \geq 1$ and $X \in L^p$,

$$\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p.$$

Proof.

Apply conditional Jensen to the convex function $\varphi(x) = |x|^p$:

$$|\mathbb{E}[X \mid \mathcal{G}]|^p \leq \mathbb{E}[|X|^p \mid \mathcal{G}] \quad \text{a.s.}$$

Take expectations and use total expectation:

$$\mathbb{E}|\mathbb{E}[X \mid \mathcal{G}]|^p \leq \mathbb{E}\,\mathbb{E}[|X|^p \mid \mathcal{G}] = \mathbb{E}|X|^p.$$

$\square$

Worked example: conditioning reduces variance

Let $\mathcal{G} = \sigma(B)$ for an event B with $0 < \mathbb{P}(B) < 1$.
Then $\mathbb{E}[X \mid \mathcal{G}]$ is piecewise constant on B, B^c:

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid B]\mathbf{1}_B + \mathbb{E}[X \mid B^c]\mathbf{1}_{B^c}.$$

Total variance gives:

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid \mathcal{G})] + \mathrm{Var}(\mathbb{E}[X \mid \mathcal{G}]).$$

So $\mathrm{Var}(\mathbb{E}[X \mid \mathcal{G}])$ measures "how much of the variance is explained by knowing B".

Worked example: best L^2 predictor from a σ-field

Let $X \in L^2$ and consider predicting X using only information $\mathcal{G}$.

$$\widehat{X} = g(\text{information in } \mathcal{G}) \quad \Longleftrightarrow \quad \widehat{X} \in L^2(\mathcal{G}).$$

Projection theorem says the unique minimizer of $\mathbb{E}[(X - \widehat{X})^2]$ is

$$\widehat{X}^\star = \mathbb{E}[X \mid \mathcal{G}].$$

This is the mathematical meaning of "best mean-square prediction given $\mathcal{G}$".

Summary: what you should remember

- $\mathbb{E}[X \mid \mathcal{G}]$ is defined by an *integral identity* on all $A \in \mathcal{G}$.
- Existence comes from Radon–Nikodym; uniqueness holds a.s.
- Tower and pull-out are the workhorses.
- In L^2, $\mathbb{E}[\cdot \mid \mathcal{G}]$ is an orthogonal projection.
- Total variance: $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid \mathcal{G})] + \mathrm{Var}(\mathbb{E}[X \mid \mathcal{G}])$.
- Conditional Jensen $\Rightarrow$ L^p contraction.