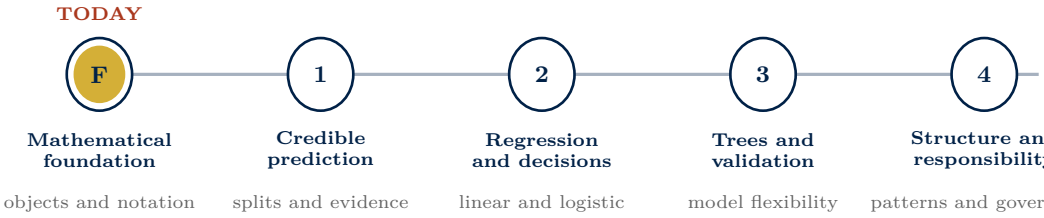
Mathematical Foundations for Machine Learning

Fatih Kansoy

Oxford

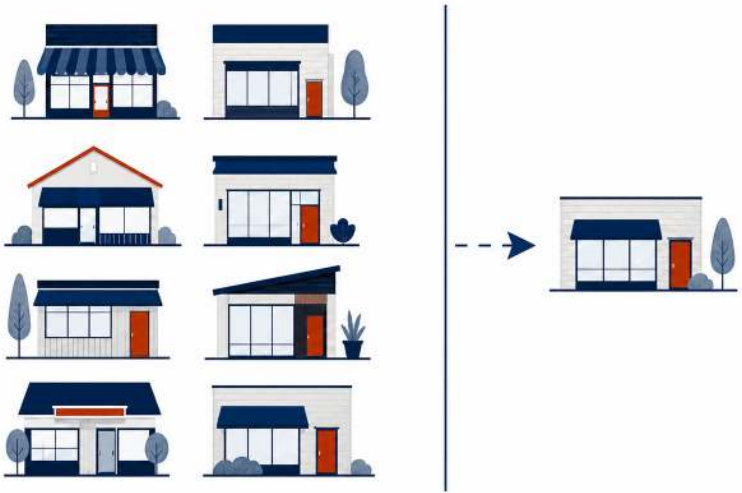
2026

A foundation workshop, then four course days



A15 · Notation reference → **A14 · Sources →**

Worked example: prediction for an unseen store



Building a model, and judging what it claims

Build: define the learning problem

- rows become vectors and matrices
- a rule maps features to predictions
- a loss defines costly errors
- optimisation picks the fitted rule

Judge: control what may be claimed

- sample is not population
- training loss is not future risk
- one estimate carries uncertainty
- association is not intervention

Every method in Days 1–4 is built on the left and defended on the right.

What this foundation workshop should leave behind

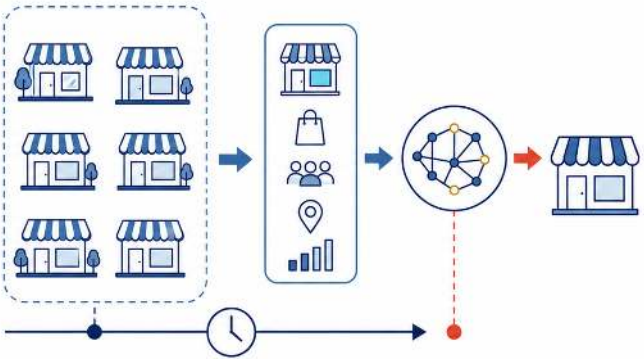
By the end of the workshop, you should be able to:

1. specify the row, prediction time, target, population, output, and loss;
2. turn a data table into vectors and a design matrix;
3. read a mean, variance, covariance, correlation, and conditional mean;
4. find the minimum of a loss curve by reading its derivative as a slope;
5. read e^x, $\ln x$, the sigmoid, and a 95% confidence interval;
6. explain empirical-risk minimisation and the least-squares objective;
7. distinguish sample fit, deployment performance, uncertainty, and causality.

Part 1

Learning problem and data structure

How can observed store-weeks support one prediction for an unseen store-week?



The forecast cutoff decides what the rule may use

Available at t_0: Q, S, I, D, K **Observed later: Y**

A **feature** is a variable supplied to the rule; the **target** is the outcome to predict. Q: previous-week demand; S: store size; I: local income; D: scheduled promotion; K: weekly capacity; Y: next-week demand. Q, K, Y use weekly demand units, S uses 100 m^2, I uses £000, and $D = 0/1$ means no/yes promotion.

For this first rule, $X = (Q, S, I, D)^\top$; K is known but deliberately reserved.

The observed sample contains eight feature–outcome records

	features used in X				reserved	target
Store	Q	S	I	D	K	Y
A	89	13	40	0	95	94
B	93	19	52	1	101	102
C	95	9	36	0	103	102
D	99	25	64	1	105	106
E	101	15	28	0	111	114
F	105	31	60	1	119	118
G	107	21	72	0	119	118
H	111	27	48	1	127	126

A **sample** is the finite collection of cases actually observed. These are synthetic data. Store A is highlighted because it becomes the first feature–outcome pair on the next slide.

One case is a pair; a sample stacks the pairs

	Features supplied to the rule	Outcome to be predicted
One case i	$\mathbf{x}_i \in \mathbb{R}^p$	$y_i \in \mathbb{R}$
All n cases	$\mathbf{X} = [\,\mathbf{x}_1^\top; \ldots; \mathbf{x}_n^\top\,] \in \mathbb{R}^{n \times p}$	$\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$

$$\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n.$$

Store A: $\mathbf{x}_A = (89, 13, 40, 0)^\top, \quad y_A = 94; \qquad n = 8,\ p = 4.$

i indexes cases, n is the number of cases, and p is the number of features. Plain X denotes a random feature vector; bold $\mathbf{X}$ is the realised design matrix. This is **supervised learning** because every fitting case pairs features with an observed target.

Sample, source population, and deployment population are different concepts

Sample
The finite cases actually observed and available for fitting.

Source population
The wider collection of cases the sample is intended to represent.

Deployment population
The future cases on which the fitted rule will be used.

Store notation: the sample is $\mathcal{D}_8$; P denotes the source distribution and P_* the deployment distribution. A distribution specifies which cases occur and how frequently.

If source and deployment differ, transfer requires evidence or an explicit assumption.

Fitting maps a sample to a fitted rule

There are two different mappings:

$$\underbrace{\mathcal{D}_n \longrightarrow \widehat{f}}_{\text{fitting}} \qquad \underbrace{\mathbf{x} \longrightarrow \widehat{f}(\mathbf{x})}_{\text{prediction}}.$$

Fitting	Input: the paired sample $\mathcal{D}_n$. | Output: one fitted rule $\widehat{f}$.
Prediction	Input: one feature vector $\mathbf{x}$. | Output: one predicted outcome $\widehat{f}(\mathbf{x})$.

$\mathcal{D}_n$ is the sample of paired feature vectors and outcomes from the previous slide; $\mathbf{x}$ is one feature vector. For an observed case, $\widehat{f}(\mathbf{x}_i) = \widehat{y}_i$ is a fitted value; for a new case it is a prediction.

The sample creates the rule; the rule then acts on one case at a time.

The store example is now complete enough to choose a method

Case and cutoff
One store-week at t_0, after D and K are known.

Features and target
Use $X = (Q, S, I, D)^\top$ to predict next-week demand Y, including unmet requests.

Fitting evidence
Eight observed feature–outcome pairs, $\mathcal{D}_8$, from the source population.

Evaluation and deployment
Separate future-like cases representing the deployment population P_*.

Output and criterion
One demand forecast; compare mean squared error with a mean-only baseline.

Intended use and claim
A planning benchmark; predictive, not a claim about intervention effects.

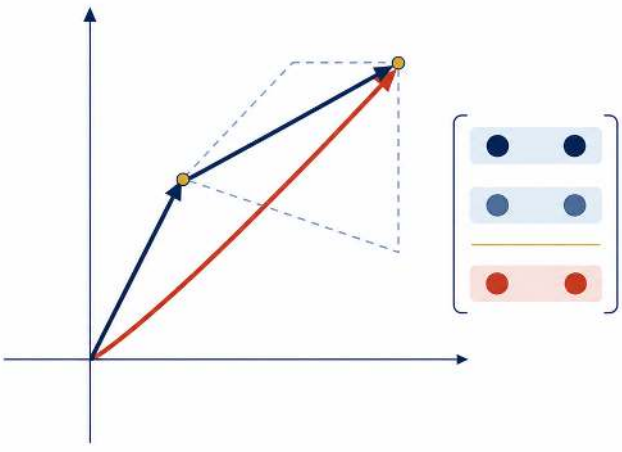
A **learning brief** fixes the unit, clock, information, target, evidence, evaluation, output, use, and permissible claim before an algorithm is selected.

The problem is now defined; next we need a mathematical representation of its cases and data.

Part 2

Vectors, geometry, and matrices

Vectors preserve coordinate meaning; dot products produce scores; matrices apply one rule; distance depends on scale.



A vector is an ordered object with fixed meanings and units

Object	General form	Meaning
Scalar	$y \in \mathbb{R}$	one real number
Ordered tuple	$(x_1, \ldots, x_p)$	values written in fixed positions
Vector	$\mathbf{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$	ordered coordinates on which vector operations are defined

Store A: $\mathbf{x}_A = (89, 13, 40, 0)^\top$ means (Q, S, I, D) in that order.

$\mathbb{R}$ is the set of real numbers; p is the number of coordinates; $^\top$ turns a row into a column. Reordering coordinates or changing their units changes what the vector means.

Three common operations act coordinate by coordinate

One colour follows one coordinate: $\mathbf{a} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 7 \\ 3 \end{bmatrix}$, $c = 6$.

SAME ROW PLUS SAME ROW

Addition $\mathbf{a} + \mathbf{b}$

$$\begin{bmatrix} 2 \\ 5 \end{bmatrix} + \begin{bmatrix} 7 \\ 3 \end{bmatrix} = \begin{bmatrix} 2+7 \\ 5+3 \end{bmatrix} = \begin{bmatrix} 9 \\ 8 \end{bmatrix}.$$

SCALAR TIMES EACH ROW

Rescaling $c\mathbf{a}$

$$6 \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 6 \times 2 \\ 6 \times 5 \end{bmatrix} = \begin{bmatrix} 12 \\ 30 \end{bmatrix}.$$

SAME ROW TIMES SAME ROW

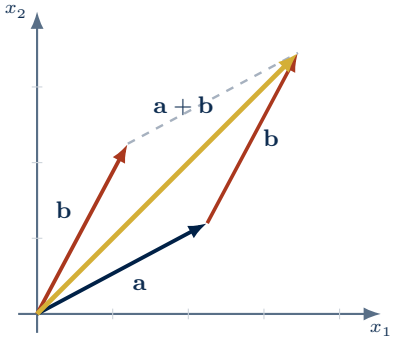
Pairwise product $\mathbf{a} \odot \mathbf{b}$

$$\begin{bmatrix} 2 \\ 5 \end{bmatrix} \odot \begin{bmatrix} 7 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \times 7 \\ 5 \times 3 \end{bmatrix} = \begin{bmatrix} 14 \\ 15 \end{bmatrix}.$$

The dot product next pairwise-multiplies, then adds: $\mathbf{a}^\top \mathbf{b} = 2 \times 7 + 5 \times 3 = 29$.

The same row-by-row rules extend to p coordinates. The symbol $\odot$ means coordinate-wise product; $\times$ normally means the three-dimensional cross product.

The same vector addition has a geometric reading



Head-to-tail rule

Move $\mathbf{b}$ without rotating or stretching it, so its tail starts at the head of $\mathbf{a}$.

Result

The arrow from the origin to the new endpoint is $\mathbf{a}+\mathbf{b}$.

The algebraic and geometric descriptions are the same operation. The diagram uses $\mathbf{a}=(2,1)^\top$ and $\mathbf{b}=(1,2)^\top$, so the endpoint is $(3,3)^\top$. Geometry is useful when vectors represent directions or differences.

A dot product multiplies matching coordinates and adds

$$\mathbf{x}^\top \boldsymbol{\beta} = x_1\beta_1 + \cdots + x_p\beta_p = \sum_{j=1}^{p} x_j\beta_j \in \mathbb{R}.$$

$$\underbrace{\mathbf{x}^\top}_{1\times p}\underbrace{\boldsymbol{\beta}}_{p\times 1} = \underbrace{\mathbf{x}^\top\boldsymbol{\beta}}_{1\times 1}.$$

Two vectors go in; one scalar comes out.

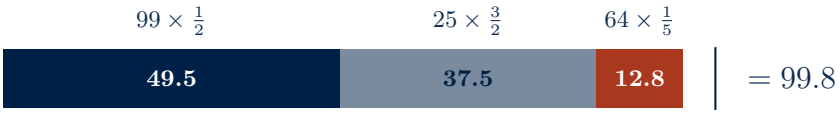
x_j and β_j occupy matching position j; p is the number of coordinates. The dot product is defined when both vectors have the same length. It measures a weighted combination here; later it also expresses length, distance, and matrix multiplication.

A linear score adds feature contributions

$$s(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j.$$

β_0 is the intercept: the part of the score not attached to one displayed feature. Each product $x_j \beta_j$ is that feature's contribution under the chosen weights.

Store D example with $\mathbf{w} = (Q, S, I)^\top$: $\mathbf{w}_D = (99, 25, 64)^\top$, $\boldsymbol{\beta} = (\frac{1}{2}, \frac{3}{2}, \frac{1}{5})^\top$, $\beta_0 = 0$.



The weights are illustrative and chosen for transparent arithmetic; nothing has been estimated on this slide.

A design matrix stacks one case per row

For n cases and p features,

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

Row i is case i; column j is feature j; x_{ij} is feature j observed for case i. A row is one feature vector written horizontally.

$$\underbrace{\begin{bmatrix} 89 & 13 & 40 & 0 \\ 93 & 19 & 52 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 111 & 27 & 48 & 1 \end{bmatrix}}_{\text{store sample } \mathbf{X} \in \mathbb{R}^{8 \times 4}} \begin{matrix} \\ \leftarrow \text{ Store A} \\ \leftarrow \text{ Store B} \\ \\ \leftarrow \text{ Store H} \end{matrix}$$

Here $n = 8$, $p = 4$, and the columns retain the order (Q, S, I, D).

Matrix multiplication applies the same score to every row

One rule uses the same intercept and weights for every case:

$$\mathbf{s} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \mathbf{x}_1^\top \boldsymbol{\beta} \\ \beta_0 + \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \beta_0 + \mathbf{x}_n^\top \boldsymbol{\beta} \end{bmatrix} \in \mathbb{R}^n.$$

$$\underbrace{\mathbf{X}}_{n\times p}\,\underbrace{\boldsymbol{\beta}}_{p\times 1} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{n\times 1}.$$

$\mathbf{s} = (s(\mathbf{x}_1), \ldots, s(\mathbf{x}_n))^\top$ is the vector of scores. $\mathbf{1}_n$ is an n-vector of ones, so $\beta_0 \mathbf{1}_n$ adds the intercept to every case. Matching inner dimensions p make the product defined; output entry i is the score for case i.

For the stores: $(8 \times 4)(4 \times 1) = (8 \times 1)$: one matrix product replaces eight dot products.

The general rule pairs each row with each column

A weight vector is one column. With several columns on the right, the same recipe runs once per column:

$$\underbrace{\begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}}_{2\times 2} \underbrace{\begin{bmatrix} 5 & 0 & 2 \\ 1 & 6 & 3 \end{bmatrix}}_{2\times 3} = \underbrace{\begin{bmatrix} 9 & 24 & 14 \\ 13 & 18 & 13 \end{bmatrix}}_{2\times 3}$$

Entry $(1, 2)$ of the result is row 1 of the left matrix against column 2 of the right one:

$$24 = 1 \times 0 + 4 \times 6.$$

Inner dimensions must match; the outer ones give the result: $(2 \times \mathbf{2})(\mathbf{2} \times 3) = (2 \times 3)$.

Each entry is a dot product, so the pairwise product followed by a sum is the operation underneath. Order matters: reversing these two matrices gives $(2 \times 3)(2 \times 2)$, whose inner dimensions 3 and 2 differ, so that product is not defined.

Distance turns a vector difference into one number

For $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$,

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}, \qquad d(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|_2.$$

$\|\mathbf{x}\|_2$ is the Euclidean norm, or length, of $\mathbf{x}$; $d(\mathbf{x}, \mathbf{z})$ is the Euclidean distance. It is non-negative and equals zero only when the two vectors match.

Store pair	**Difference on $\mathbf{w} = (Q, S, I)^\top$**	**Distance**
B versus A	$(4, 6, 12)^\top$	14
B versus H	$(-18, -8, 4)^\top$	$\sqrt{404} \approx 20.1$

In the displayed table units, B is nearer A than H.

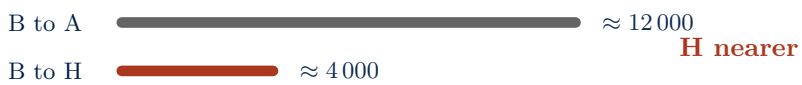
For example, $d(\mathbf{w}_B, \mathbf{w}_A) = \sqrt{4^2 + 6^2 + 12^2} = 14$. This conclusion depends on the chosen coordinate units.

Changing one unit can reverse the nearest neighbour

Keep Q and S exactly as before, and record the same incomes in pounds rather than £000, so every income coordinate is multiplied by 1000. Nothing else changes.

Income recorded in £000

The same income recorded in pounds



The stores did not change; the raw Euclidean geometry did.

A **nearest neighbour** is the observed case with the smallest chosen distance. Writing income in pounds multiplies its coordinate differences by 1000 before those differences are squared.

Standardisation measures each feature in its own standard deviations

For case i and feature j,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

$\bar{x}_j$ is the training-sample mean of feature j; s_j is its training-sample standard deviation; z_{ij} is unitless and records how many standard deviations case i lies from that mean.

$$\underbrace{\frac{52 - 50}{14}}_{\text{income in £000}} = \frac{1}{7} \quad \text{and} \quad \underbrace{\frac{52\,000 - 50\,000}{14\,000}}_{\text{the same income in pounds}} = \frac{1}{7}.$$

Estimate $\bar{x}_j$ and s_j on the training sample only; reuse them unchanged for evaluation and deployment.

For the store comparison, $d_{\text{std}}(B, A) \approx 1.34$ and $d_{\text{std}}(B, H) \approx 2.83$. The ranking no longer changes when income is converted from £000 to pounds.

Checkpoint: read the object before calculating

Given

$$\mathbf{x}_i \in \mathbb{R}^4, \qquad \mathbf{X} \in \mathbb{R}^{8\times 4}, \qquad \boldsymbol{\beta} \in \mathbb{R}^4,$$

work from meaning and dimensions:

1. distinguish the feature vector, design matrix, and coefficient vector;
2. decide whether $\mathbf{x}_i^\top \boldsymbol{\beta}$, $\mathbf{X}\boldsymbol{\beta}$, and $\boldsymbol{\beta}^\top \mathbf{X}$ are defined, and state each output dimension;
3. explain why changing income from £000 to pounds can reverse a raw-distance ranking;
4. explain why standardisation parameters must be learned on training data only.

A dimension check establishes whether an operation is defined; it does not by itself establish that the variables, units, or prediction task are scientifically appropriate.

Representation determines which operation answers the question

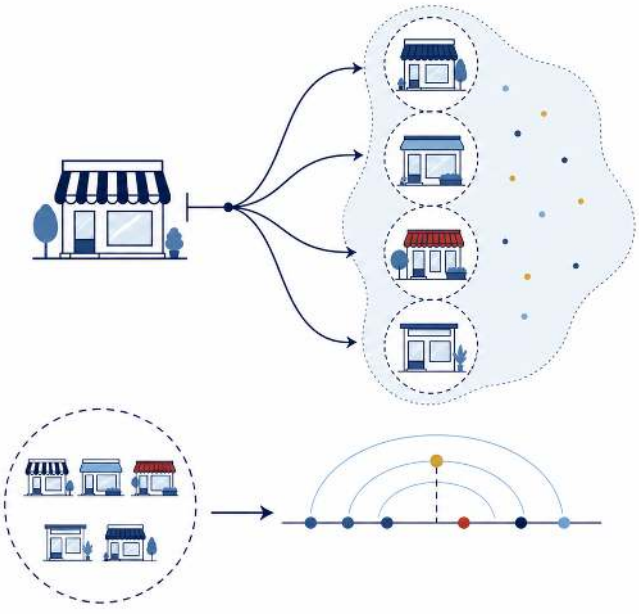
Question	Objects	Operation	Output
Score one case	$\mathbf{x}, \boldsymbol{\beta}$	$\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$	one scalar
Score all cases	$\mathbf{X}, \boldsymbol{\beta}$	$\beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}$	n-vector
Compare two cases	$\mathbf{x}, \mathbf{z}$	$\lVert \mathbf{x} - \mathbf{z} \rVert_2$	one distance
Remove unit dependence	$x_{ij}, \bar{x}_j, s_j$	$(x_{ij} - \bar{x}_j)/s_j$	unitless coordinate

Next: even at a fixed feature vector, the outcome remains uncertain; probability describes that uncertainty.

Part 3

Probability and sample statistics

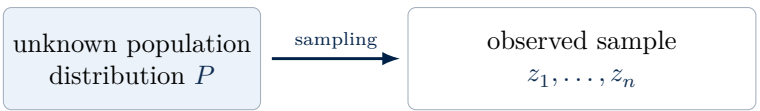
Probability names unknown population quantities; the observed sample supplies evidence about them.



A probability distribution describes possible future cases

Before a case is observed, its variables are uncertain. Write

$$Z \sim P, \qquad z_i \text{ for the value observed in case } i.$$



Use the sample to learn about future cases governed by P.

A random variable is written with a capital letter; its observed value uses the matching lower-case letter. For the store example, $Z = (X, K, Y)$, $X = (Q, S, I, D)^\top$, and $z_i = (\mathbf{x}_i, k_i, y_i)$. We treat $Z_1, \ldots, Z_n$ as independent draws from P; this is an assumption about how the sample arose.

The distribution is not observed directly; only its realised sample is.

A distribution has a centre and a spread

For any numerical random variable U,

$$\underbrace{\mathbb{E}[U]}_{\text{population centre}}, \qquad \underbrace{\mathrm{Var}(U) = \mathbb{E}[(U - \mathbb{E}[U])^2]}_{\text{population spread}}, \qquad \mathrm{sd}(U) = \sqrt{\mathrm{Var}(U)}.$$

Applying the corresponding formulas to the eight observed demands gives

$$\underbrace{\bar{y} = 110}_{\text{sample centre}}, \qquad \underbrace{v_8(\mathbf{y}) = 100}_{\text{empirical variance}}, \qquad \underbrace{s_8(\mathbf{y}) = 10}_{\text{empirical standard deviation}}.$$

Expectation and variance are properties of the unknown population distribution. The barred and subscripted quantities are descriptions of this particular sample. Variance has squared units; standard deviation has the same units as the variable. Here $v_8(\mathbf{y}) = 8^{-1} \sum_i (y_i - \bar{y})^2$.

The sample statistic is evidence about the population quantity; it is not the quantity itself.

A1 · Exact sample moments →

Covariance and correlation describe joint movement

For numerical variables U and V,

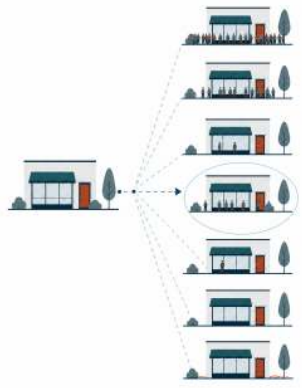
$$\mathrm{Cov}(U,V) = \mathbb{E}[(U - \mathbb{E}[U])(V - \mathbb{E}[V])], \qquad \rho_{UV} = \frac{\mathrm{Cov}(U,V)}{\mathrm{sd}(U)\,\mathrm{sd}(V)}.$$

Store feature	sample covariance with Y	sample correlation with Y
Previous demand Q	69	0.99
Store size S	47	0.67
Income I	42	0.30

Positive covariance means that above-centre values tend to occur together. Covariance carries the product of the two units. Correlation divides out both standard deviations, so it is unit-free and lies between -1 and 1. The displayed numbers describe the sample; the population values remain unknown.

Joint movement is association, not evidence that changing one variable changes the other.

Prediction is conditional: hold the features fixed, then average



For a specified feature profile $\mathbf{x}$, define

$$m(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}].$$

The bar means "given." Conceptually, keep the profile fixed and average the remaining variation in Y.

Different cases with the same profile can still have different outcomes; $m(\mathbf{x})$ is the centre of that conditional distribution.

This is the population prediction target used later under squared loss.

With continuous features, the exact profile $\mathbf{x}$ may not repeat in a small sample. A model therefore learns a smooth or structured rule across profiles. Prediction outside the population's observed support is extrapolation and needs additional assumptions.

Conditioning changes the population being averaged

$$\underbrace{\mathbb{E}[Y]}_{\text{all cases}} \qquad \underbrace{\mathbb{E}[Y \mid X = \mathbf{x}]}_{\text{cases with one profile}} \qquad \underbrace{\mathbb{E}[Y \mid D = 1]}_{\text{cases in one group}} .$$

In the store sample, conditioning only on promotion status gives

$$\widehat{m}_1 = \frac{1}{4} \sum_{i:D_i=1} y_i = 113, \qquad \widehat{m}_0 = \frac{1}{4} \sum_{i:D_i=0} y_i = 107.$$

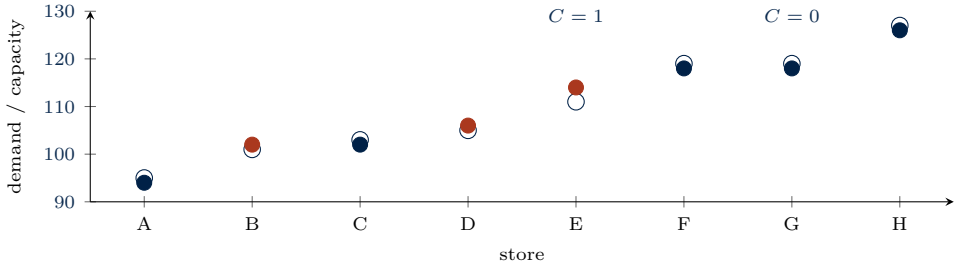
$D = 1$ means a scheduled promotion and $D = 0$ no promotion. The hats mark sample-based group means. Each mean uses four stores, and the six-unit difference is a comparison between two observed groups.

A conditional comparison need not be a causal effect: the groups may differ in other ways.

Part 6 asks how much the six-unit sample gap may wobble across samples; Part 7 asks what assumptions would permit an intervention claim.

For a binary outcome, a conditional mean is a probability

Set $C = 1$ when demand exceeds capacity, $Y > K$, and $C = 0$ otherwise.



Open circles are capacity K, filled circles demand Y; three of the eight stores exceed it.

For any $0/1$ variable, its average is the proportion of ones:

$$\mathbb{E}[C \mid X = \mathbf{x}] = P(C = 1 \mid X = \mathbf{x}).$$

Open circles show capacity K; filled circles show demand Y. Three of the eight stores have $C = 1$, so the unconditional sample proportion is $3/8$. The conditional identity is what allows a regression rule to predict a probability, as logistic regression does on Day 2.

Checkpoint: population quantity or sample statistic?

Classify each object and explain what it describes:

$$\mathbb{E}[Y \mid X = \mathbf{x}], \qquad \widehat{m}_1 = 113, \qquad \rho_{QY}, \qquad s_8(\mathbf{y}) = 10.$$

Then explain why writing income in pounds changes its covariance with demand but does not change their correlation.

Every symbol needs a population or sample referent.

Population quantities are targets; sample statistics are evidence

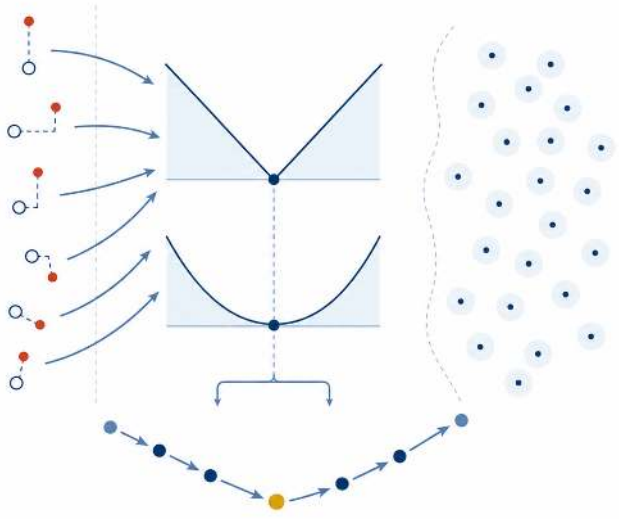
Population quantity	Question it answers	Sample evidence
$\mathbb{E}[Y]$	Where is demand centred?	sample mean $\bar{y}$
$\mathrm{Var}(Y)$	How much does demand vary?	empirical variance $v_n(\mathbf{y})$
$\mathrm{Cov}(X_j, Y)$, ρ_{jY}	How do a feature and demand move together?	sample covariance and correlation
$\mathbb{E}[Y \mid X = \mathbf{x}]$	What is mean demand at profile $\mathbf{x}$?	a fitted prediction rule, developed next

Next question: how should the sample choose one fitted rule rather than another?

Part 4

Loss, risk, and optimisation

Loss defines what counts as a good prediction; ERM and optimisation turn that definition into a fitted rule.



A loss turns one prediction error into one cost

For one case, the rule produces a prediction $a = f(\mathbf{x})$. After the outcome y is observed,

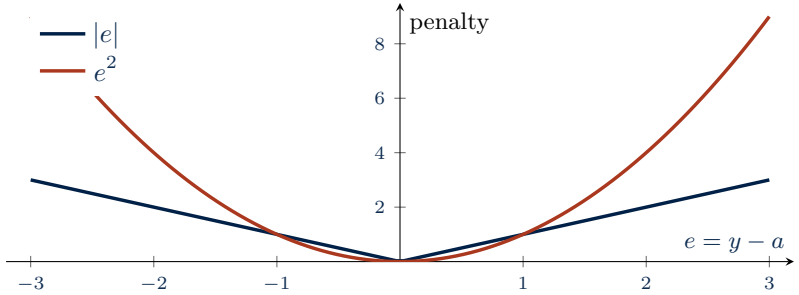
$$\underbrace{e = y - a}_{\text{signed error}} \quad \longrightarrow \quad \underbrace{\ell(y, a)}_{\text{non-negative cost}} \quad .$$

Loss	Formula	How it treats a miss
Absolute loss	$\ell_1(y, a) = \lvert y - a \rvert$	cost rises in direct proportion to the size of the miss
Squared loss	$\ell_2(y, a) = (y - a)^2$	cost grows much faster for large misses

Example: $a = 110$, $y = 94$ gives $e = -16$, absolute loss 16, and squared loss 256.

A loss is chosen before fitting and supplies the meaning of "better." Absolute loss has outcome units; squared loss has squared outcome units. RMSE takes the square root after averaging squared losses.

Different losses reward different population targets



Absolute loss rises at a constant rate; squared loss grows more quickly as the miss becomes large. This difference changes which prediction is best for an asymmetric outcome distribution.

Expected squared loss targets a conditional mean; absolute loss targets a conditional median.

The workshop uses squared loss from here on. A conditional median need not be unique.

Risk averages the loss of a whole rule

A loss scores one prediction for one case. To judge the rule f itself, take the average of those costs over many cases.

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(\mathbf{x}_i)\big) \qquad \text{empirical risk}$$

The average cost over the n cases we hold. Every term is observed, so this is a number we can work out. Take the rule that predicts 110 for every store: it misses store A by 16, a squared error of 256, and store D by 4, a squared error of 16. Averaged over the eight stores, the squared errors come to 100.

$$\mathcal{R}_P(f) = \mathbb{E}_P\big[\ell(Y, f(X))\big] \qquad \text{population risk}$$

The same average, taken over every store in the population rather than the eight we happen to have. This is the quantity we would like to make small, and it can never be worked out, because P is never observed. Written $\mathcal{R}_{P_*}(f)$ when the average is taken over the deployment population P_*.

Training risk is observable; deployment risk is the operational target.

Under squared loss, the population target is a conditional mean

Hold the profile fixed at $\mathbf{x}$ and consider any single number a proposed as the prediction there. Write $m_P(\mathbf{x}) = \mathbb{E}_P[Y \mid X = \mathbf{x}]$. Then

$$\mathbb{E}_P[(Y-a)^2 \mid X=\mathbf{x}] = \underbrace{\mathrm{Var}_P(Y \mid X=\mathbf{x})}_{\text{variation no prediction can remove}} + \underbrace{\{m_P(\mathbf{x})-a\}^2}_{\text{cost of choosing the wrong centre}}.$$

Sizes: $\mathbf{x} \in \mathbb{R}^p$ is one store's profile, a vector of p feature values, here $(Q, S, I, D)^\top$ with $p = 4$. The outcome Y, the proposal a, and the value $m_P(\mathbf{x})$ are each a single number. So m_P is a function that takes a vector and returns one number, and the display compares numbers, not vectors.

The first term does not depend on a. The second is a square, so it is never negative, and it reaches zero only at

$$a = m_P(\mathbf{x}).$$

The conditional variance is the remaining spread in outcomes among cases sharing the profile $\mathbf{x}$. It is an irreducible floor for prediction at that profile, not an error caused by the fitted model.

Squared loss makes the conditional mean the ideal population prediction.

ERM makes the training problem computable

Ideally, choose a rule with the smallest population risk. In practice, make two explicit substitutions:

1. replace the unknown population average by the observed sample average;
2. search within a stated class $\mathcal{F}$ of candidate rules.

This gives **empirical-risk minimisation (ERM)**:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}_n(f).$$

$\mathcal{F}$ is the set of rules allowed before the search; $\arg\min$ means the rule or rules at which the displayed quantity is smallest. ERM is a fitting procedure: it chooses the lowest-training-risk rule available inside $\mathcal{F}$.

Constant rules give the mean baseline; linear rules give least squares in Part 8.

The constant baseline is the simplest ERM problem

The smallest rule class ignores the profile and predicts one number c for every store:

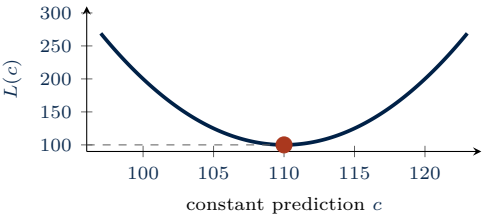
$$\mathcal{F} = \{f(\mathbf{x}) = c\}, \qquad L(c) = \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2.$$

Miss store i by $y_i - c$, square the miss, average over the eight.

The same sum, rewritten:

$$L(c) = \underbrace{v_8(\mathbf{y})}_{100} + \underbrace{(\bar{y} - c)^2}_{(110-c)^2}.$$

Insert $\bar{y}$ in the bracket and expand; the cross term cancels because deviations from the mean sum to zero (A5). The first term is the spread of the eight outcomes and holds no c; the second is how far the constant sits from the mean.



Each point is one candidate c and the empirical risk it gives.

Best training constant: 110. A richer rule must beat it on matched evaluation data.

A5 · Baseline decomposition →

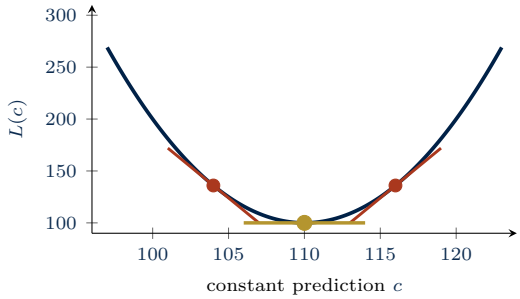
A derivative tells us which way lowers the loss

The derivative is the slope of a curve at a point:

$$L(c) = 100 + (c - 110)^2, \qquad L'(c) = 2(c - 110).$$

Read $L'(c)$ as the local change in loss when c moves a small amount to the right. Its sign tells us which direction lowers the curve.

- at $c = 104$: $L' = -12 < 0$, moving right lowers the loss;
- at $c = 116$: $L' = +12 > 0$, moving left lowers the loss;
- at $c = 110$: $L' = 0$, no direction improves.



Each short line shows the local slope. The tangent is horizontal at the minimum.

At this minimum the slope is zero. Solving $L'(c) = 0$ returns $c = 110$.

A zero derivative is necessary at a smooth interior minimum; the upward-curving parabola confirms that this stationary point is the minimum.

A gradient collects one slope for each adjustable parameter

With several parameters, the loss is a surface and each parameter has its own partial derivative. For two coefficients,

$$\frac{\partial L}{\partial \beta_0} \text{ measures the } \beta_0 \text{ direction,} \qquad \frac{\partial L}{\partial \beta_1} \text{ measures the } \beta_1 \text{ direction.}$$

Collect them in the **gradient** vector:

$$\nabla L(\boldsymbol{\beta}) = \left(\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1} \right)^{\top}.$$

At a smooth interior minimum, every directional slope is zero: $\nabla L(\boldsymbol{\beta}) = \mathbf{0}$.

A partial derivative changes one parameter while holding the others fixed. Part 8 applies this idea to five least-squares coefficients, producing five equations.

Gradient descent follows the slopes when no formula is available

When the minimum cannot be written in closed form, move against the derivative:

$$c_{\text{new}} = c_{\text{old}} - \eta\, L'(c_{\text{old}}), \qquad \eta > 0.$$

On the baseline parabola, with learning rate $\eta = 0.25$:

Step	c	$L'(c)$	**move** $-\eta L'(c)$
0	120	20	-5
1	115	10	-2.5
2	112.5	5	-1.25
3	111.25	2.5	-0.625

Each step halves the distance to 110; the sequence approaches the minimum without solving $L'(c) = 0$.

η is the learning rate. Too large can overshoot; too small moves slowly.

With several parameters, replace $L'(c)$ by $\nabla L(\boldsymbol{\beta})$. Day 2 uses numerical optimisation for logistic regression; Day 3's boosting also follows the negative gradient of a loss.

Checkpoint: trace the path from a loss to a fitted rule

1. For $a = 106$ and $y = 118$, calculate the error, absolute loss, and squared loss.
2. Explain why squared loss makes $\mathbb{E}[Y \mid X = \mathbf{x}]$ the population target.
3. State the two substitutions that turn population-risk minimisation into ERM.
4. Differentiate $L(c) = 100 + (c - 110)^2$, solve $L'(c) = 0$, and interpret the answer.
5. Starting from $c = 120$ with $\eta = 0.25$, take two gradient-descent steps.

The intended chain is: define one-case cost, average it, replace the unknown population average by a sample average, then minimise over a stated rule class.

Low training risk is not evidence of low deployment risk

Fit
Use the training sample to choose the rule and any preprocessing.

Judge
Freeze the procedure, then evaluate it on untouched future-like cases.

$$\widehat{\mathcal{R}}_n(\widehat{f}) \text{ small} \quad \not\Rightarrow \quad \mathcal{R}_{P_*}(\widehat{f}) \text{ small.}$$

ERM guarantees only that no candidate in $\mathcal{F}$ has lower risk on the same fitting sample. Generalisation needs separate evaluation and assumptions about how future cases relate to the observed data.

Part 5

Exponentials, logarithms, and the sigmoid

Classification needs probabilities. Exponentials, logarithms, and the sigmoid connect scores, probabilities, and log loss.

$$p = \frac{1}{1 + e^{-z}}$$

Exponentials and logarithms undo one another

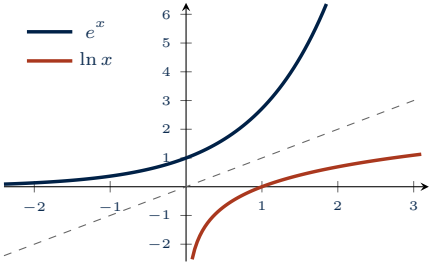
Exponential e^x: positive for every real x, increasing quickly to the right and approaching zero to the left ($e \approx 2.718$).

Natural logarithm $\ln x$: defined for $x > 0$, and the exact inverse:

$$\ln(e^x) = x, \qquad e^{\ln x} = x.$$

Anchor values:

$$e^0 = 1, \qquad \ln 1 = 0, \qquad \ln e = 1.$$



The two curves mirror one another across the dashed line $y = x$ because each function reverses the other.

As a positive number approaches zero, its logarithm falls without bound.

That boundary behaviour creates the large log-loss penalty for assigning almost zero probability to an outcome that occurs. Day 2 also uses e^{β} as an odds multiplier; Day 3 uses $e^{-1} \approx 0.368$ in the bootstrap calculation.

Logs turn products into sums and compress large values

Products become sums

$$\ln(ab) = \ln a + \ln b, \qquad \text{e.g.} \quad \ln 20 = \ln 2 + \ln 10 \approx 0.693 + 2.303 = 2.996.$$

Likelihood multiplies case-level probabilities. Taking logs turns that product into a sum that is easier to compute and optimise.

Multiplicative changes become additive changes Doubling always adds $\ln 2 \approx 0.693$, wherever the scale starts:

$$\ln 2 - \ln 1 = \ln 4 - \ln 2 = \ln 200 - \ln 100.$$

This compression can make long right-tailed quantities, such as prices or incomes, easier to model.

The first use supports likelihood and log loss; the second changes the scale on which a relationship is modelled. They are different uses of the same logarithm.

The logarithm preserves order while changing multiplication into addition.

Day 2 uses log probabilities in classification and encounters a long-right-tailed London price distribution.

Log loss penalises the probability assigned to what happened

Let q be the probability the model assigned to the outcome that actually happened. Its case-level log loss is $-\ln q$:

Claimed q for what happened	$-\ln q$	Reading
0.9	0.105	confident and right: tiny penalty
0.5	0.693	fence-sitting: moderate penalty
0.1	2.303	confident and wrong: heavy penalty
$q \to 0$	$\to \infty$	refuted certainty: unbounded

Average $-\ln q_i$ over cases to score the whole probability rule.

An extreme probability receives an extreme penalty when it is wrong.

For binary classification, $q = p$ when the observed outcome is 1, and $q = 1 - p$ when it is 0. Maximising likelihood and minimising average log loss select the same fitted rule because the logarithm turns the likelihood product into a sum.

The sigmoid turns any real score into a probability

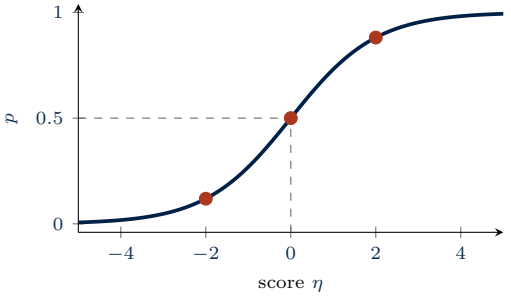
A linear score $\eta = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ can be any real number; a probability must lie between zero and one. The sigmoid provides the bridge:

$$p = \frac{1}{1 + e^{-\eta}} \in (0, 1).$$

Invert it and the score reappears as the **log-odds**:

$$\ln\left(\frac{p}{1-p}\right) = \eta.$$

If $p = 3/8$, the odds are $(3/8)/(5/8) = 0.6$, so the log-odds are $\ln 0.6 \approx -0.51$.



Scores $-2, 0, 2$ map to probabilities $0.12, 0.50, 0.88$. The curve is steepest near $p = 1/2$ and flattens near zero and one.

Logistic regression is linear in log-odds and curved in probability.

A6 · Sigmoid and log-odds identities →

Categories enter the score through 0/1 indicator columns, as they do in the design matrix of Part 8. Day 2 fits logistic regression by minimising log loss, equivalently maximising likelihood.

Checkpoint: move between score, probability, and loss

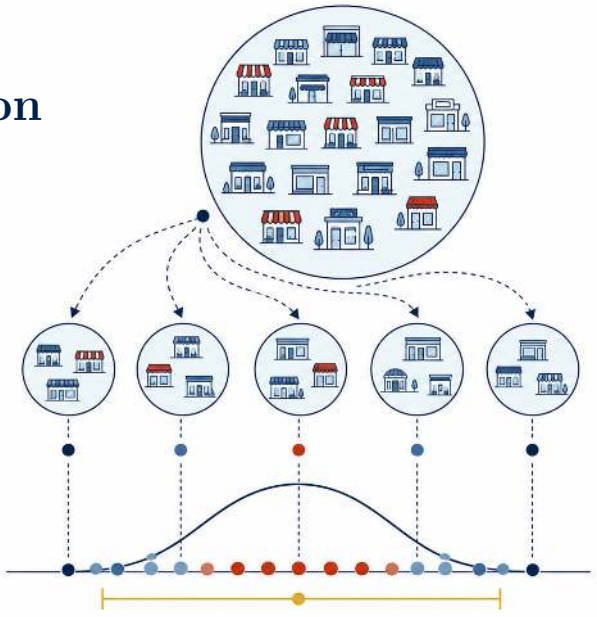
1. Using $\ln 2 \approx 0.693$ and $\ln 10 \approx 2.303$, evaluate $\ln 40$.
2. A model assigned $q = 0.25$ to what happened. Compute its case-level log loss.
3. Use the sigmoid to map $\eta = 0$ to a probability. What do positive and negative scores imply?
4. Explain why a product of case probabilities becomes a sum of case-level log losses.

Next: fitting chooses numerical scores and probabilities from a sample; how much would those fitted quantities change in another sample?

Part 6

Statistical estimation and uncertainty

A sample returns one estimate. Repeated-sample reasoning describes how that estimate can vary.



Estimand, estimator, and estimate are different objects

Estimand: the population quantity the analysis is trying to learn, fixed but unknown:

$$\theta_0 = T(P), \qquad \text{for example } \mu_Y = \mathbb{E}_P[Y].$$

Estimator: the sample-based rule used to learn that quantity; before the sample is observed, its output is random:

$$\widehat{\theta} = \widehat{\theta}(Z_1, \ldots, Z_n), \qquad \text{for example } \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Estimate: the numerical value returned by the realised sample:

$$\widehat{\theta}_{\text{obs}} = \widehat{\theta}(z_1, \ldots, z_n), \qquad \text{here } \bar{y} = 110.$$

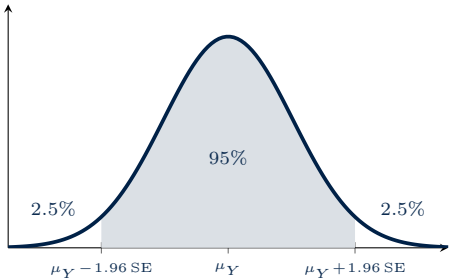
The target is fixed; the estimator varies across samples; the estimate is the value we observed.

The estimand states the question. The estimator states the procedure. The estimate is one realised answer from that procedure.

Repeated samples create a sampling distribution

Imagine drawing a new sample of n stores repeatedly and recording $\bar{Y}_n$ each time.

- the resulting estimates form a distribution;
- its centre is μ_Y under unbiased iid sampling;
- its spread is the standard error;
- for sample means, its shape becomes approximately normal as n grows.



The standardised normal curve has 95% of its area between -1.96 and $+1.96$. With a small sample and an estimated standard error, the next slide uses the wider Student t distribution.

Area under this curve describes repeated-sample probability.

The central limit theorem supplies the large-sample normal approximation under suitable sampling conditions. Normality is not a visual claim about the eight observed outcomes themselves.

A standard error measures the spread of an estimator

Under iid sampling with population standard deviation σ_Y,

$$\mathrm{SE}(\bar{Y}_n) = \sqrt{\mathrm{Var}_P(\bar{Y}_n)} = \frac{\sigma_Y}{\sqrt{n}}.$$

The population spread is unknown, so estimate it from the sample. For the eight stores, $s_Y \approx 10.69$, giving

$$\widehat{\mathrm{SE}}(\bar{Y}_8) = \frac{10.69}{\sqrt{8}} \approx 3.78.$$

Individual-store spread: $s_Y \approx 10.69$ **Estimator spread:** $\widehat{\mathrm{SE}} \approx 3.78$.

The standard deviation describes variation across individual outcomes. The standard error describes how the estimator would vary across samples of the same size. A standard error is not uncertainty about one future store.

A7 · Standard-error calculation →

A confidence interval combines an estimate with its uncertainty

Report the estimate together with how far the truth could plausibly sit from it. Reach out on both sides by a multiple of the standard error:

$$\bar{y} \pm t_{0.975,7}\,\widehat{\mathrm{SE}} = 110 \pm 2.365 \times 3.78 = [101.1,\ 118.9].$$

The multiplier 2.365 is the Student t value for 7 degrees of freedom. It is wider than the normal-curve 1.96 because the standard error is itself estimated, from only eight stores.

How to read it: intervals built by this recipe contain μ_Y in 95% of repeated samples. The 95% describes the recipe, not this one interval: once the endpoints are computed, μ_Y is either inside them or it is not.

This interval concerns the population mean, not the demand of one future store.

Day 2 applies the same structure to regression coefficients, with software selecting the critical value.

A test compares a sample gap with its standard error

Part 3 found a sample group-mean gap of $\widehat{m}_1 - \widehat{m}_0 = 113 - 107 = 6$. Compare it with the estimated repeated-sample spread of such a gap:

$$\widehat{\mathrm{SE}}(\text{gap}) \approx 7.79, \qquad t = \frac{6}{7.79} \approx 0.77.$$

The statistic t measures the observed gap in standard-error units. If the population group-mean gap were zero, the two-sided p-value is the probability of obtaining a statistic at least this far from zero. Here $p \approx 0.47$.

The p-value is not the probability that the null hypothesis is true.

Each group contains four stores and has conventional sample variance $364/3$, so $\widehat{\mathrm{SE}} = \sqrt{(364/3)/4 + (364/3)/4} \approx 7.79$. With six degrees of freedom, the two-sided value is about 0.47. This is a descriptive group comparison, not an estimate of the promotion's causal effect.

More data reduces some uncertainty, not all

Uncertainty about an estimate

$$\widehat{\theta} - \theta_0$$

contains sampling variation and possible bias.

For many fixed procedures under iid sampling, more data reduces sampling variance.

More data alone does not guarantee that systematic bias disappears.

Uncertainty about a new outcome

$$Y^* - \widehat{f}(X^*)$$

contains model-estimation error and irreducible outcome variation.

More data can reduce the estimation component; it cannot remove $\mathrm{Var}(Y^* \mid X^*)$.

Prediction uncertainty can shrink towards a floor; it does not generally shrink to zero.

A8 · Bias, variance, and the prediction floor →

A confidence interval for a population mean and a prediction interval for one future store answer different questions. The latter is wider because it includes individual-outcome variation.

Checkpoint: match each uncertainty statement to its question

1. Classify μ_Y, $\bar{Y}_8$, and 110 as estimand, estimator, or estimate.
2. Explain why $s_Y \approx 10.69$ and $\widehat{\mathrm{SE}}(\bar{Y}_8) \approx 3.78$ describe different variation.
3. Reconstruct the small-sample interval $[101.1, 118.9]$ using 2.365 and 3.78.
4. If the same six-unit gap appeared with far more stores per group, what would usually happen to its standard error and $|t|$?
5. Name one source of prediction uncertainty that more data cannot remove.

Different uncertainty statements answer different questions

Object	Question	What it describes
Standard deviation	How much do individual outcomes vary?	spread across cases
Standard error	How much would the estimator vary?	spread across repeated samples
Confidence interval	Which population values are compatible with the estimate?	estimate plus sampling uncertainty
Prediction interval	Where might one future outcome fall?	estimation uncertainty plus outcome variation

Next: uncertainty quantifies a stated target; it cannot change a predictive comparison into a causal one.

Part 7

Prediction and causal interpretation

Prediction, description, and intervention are different questions and require different evidence.

Prediction, description, and intervention ask different questions

Goal	Question	Target	Evidence required
Predict	What outcome should we expect for a future case with features $\mathbf{x}$?	$m_{P_*}(\mathbf{x})$	performance on unseen, deployment-like cases
Describe	What is true of a stated population?	for example $\mathbb{E}_P[Y]$	sampling framework and uncertainty
Intervene	What would change if promotion status were changed?	average treatment effect	design or identification assumptions

The same fitted equation can be reported for all three goals, but fitting does not supply all three kinds of evidence.

P_* is the deployment population from Part 1. A predictive target concerns future outcomes under the observed operating process. A causal target concerns outcomes under alternative actions.

A causal effect compares two potential outcomes for the same case

Let $D = 1$ denote promotion and $D = 0$ no promotion. For store i, define

$$Y_i(1) = \text{demand if promoted}, \qquad Y_i(0) = \text{demand if not promoted}.$$

Store	**Observed** D_i	**Observed** Y_i	$Y_i(0)$	$Y_i(1)$
A	0	94	94	unobserved
B	1	102	unobserved	102

Only one potential outcome is observed for each store. The individual causal effect $Y_i(1) - Y_i(0)$ is therefore never observed directly. The **average treatment effect (ATE)** is

$$\text{ATE} = \mathbb{E}_P[Y(1) - Y(0)].$$

The ATE asks for the population-average change in demand caused by switching the same cases from no promotion to promotion. It is not simply the difference between two observed groups.

The observed groups are not the missing causal comparison

The promoted and unpromoted stores differ before outcomes are compared:

Sample mean	Promoted $D = 1$	Not promoted $D = 0$	Difference
Demand Y	113	107	6
Previous demand Q	102	98	4
Store size S	25.5	14.5	11
Income I	56	44	12

The six-unit outcome gap compares different stores facing different feature profiles. It mixes any promotion effect with pre-existing group differences.

Observed groups are not automatically substitutes for the same cases under two actions.

The p-value in Part 6 quantified sampling uncertainty in the observed group comparison. It did not remove these systematic differences or convert the comparison into an ATE.

An effect claim needs assumptions that fit cannot supply

Compare promoted and unpromoted stores within the same pre-promotion profile $W = (Q, S, I)^\top$. The observed-data comparison identifies an effect only if:

- **consistency:** $Y = Y(D)$, the observed outcome matches the treatment actually received;
- **conditional exchangeability:** $(Y(0), Y(1)) \perp D \mid W$, where within W, assignment carries no further information about the potential outcomes;
- **overlap:** $0 < P(D = 1 \mid W) < 1$, so both treatment states occur at the profiles compared.

Model fit cannot establish consistency or exchangeability; observed overlap must be assessed separately.

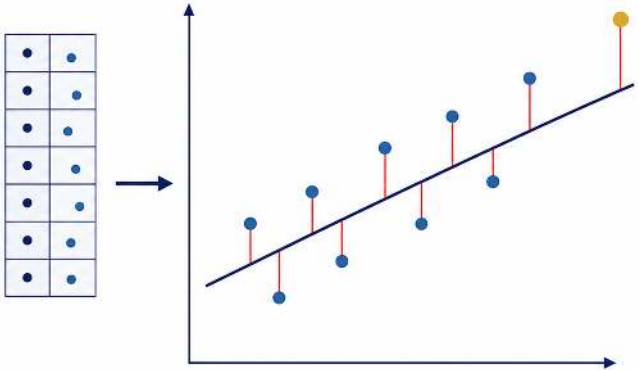
A9 · Optional identification formula →

W must be measured before promotion is set. A lack of overlap can be visible in the observed data; adequate overlap does not prove exchangeability. Variables measured after treatment can themselves be consequences of treatment. Least squares can estimate a conditional association, but its optimisation cannot supply these assumptions.

Part 8

Linear least squares

Vectors represent the cases; a linear rule predicts; squared-loss ERM selects its coefficients; calculus solves the optimisation.



The constant baseline cannot adapt to different stores

Part 4 found the best constant: predict $c = 110$ for every store.

Store	Profile (Q, S, I, D)	Demand y	Prediction	Residual $y - 110$
A	$(89, 13, 40, 0)$	94	110	-16
H	$(111, 27, 48, 1)$	126	110	$+16$

The baseline is a useful benchmark, but it deliberately ignores the feature profile.

A **residual** is observed outcome minus fitted prediction. A negative residual means the rule predicted too high; a positive residual means it predicted too low.

Can one common rule use the features to adapt its prediction from store to store?

Why start with a linear rule? It is the simplest feature-dependent class that connects directly to the dot product and design matrix; later methods relax this restriction.

A linear rule adds feature contributions to an intercept

For store i, collect the four available features:

$$\mathbf{x}_i = (Q_i, S_i, I_i, D_i)^\top.$$

Each coefficient is a weight: its feature contributes one term to the prediction.

$$f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \beta_0 + \beta_1 Q_i + \beta_2 S_i + \beta_3 I_i + \beta_4 D_i.$$

Add a leading 1 so the same rule becomes one dot product:

$$\widetilde{\mathbf{x}}_i = (1, Q_i, S_i, I_i, D_i)^\top, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top, \qquad f_{\boldsymbol{\beta}}(\mathbf{x}_i) = \widetilde{\mathbf{x}}_i^\top \boldsymbol{\beta}.$$

The same five coefficients are used for every store.

β_0 is the all-zero-feature intercept, which need not describe a realistic store. The vector $\boldsymbol{\beta}$ is still a candidate; only fitting the sample produces $\widehat{\boldsymbol{\beta}}$. Categorical features enter the same rule as 0/1 indicator columns with one level left out as the reference; Day 2 does this for room types.

One design matrix applies the rule to all eight stores

Stack one augmented feature row per store:

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \widetilde{\mathbf{x}}_1^\top \\ \vdots \\ \widetilde{\mathbf{x}}_8^\top \end{bmatrix} \in \mathbb{R}^{8\times 5}.$$

Then one matrix product gives all predictions:

$$\underbrace{\widetilde{\mathbf{X}}}_{8\times 5}\underbrace{\boldsymbol{\beta}}_{5\times 1} = \underbrace{\begin{bmatrix} f_{\boldsymbol{\beta}}(\mathbf{x}_1) \\ \vdots \\ f_{\boldsymbol{\beta}}(\mathbf{x}_8) \end{bmatrix}}_{8\times 1}.$$

Object	Meaning	Size
$\widetilde{\mathbf{X}}$	features and intercept	8×5
$\boldsymbol{\beta}$	candidate coefficients	5×1
$\widetilde{\mathbf{X}}\boldsymbol{\beta}$	eight predictions	8×1
$\mathbf{y}$	eight outcomes	8×1

The residual vector compares what happened with what the rule predicts:

$$\mathbf{r}(\boldsymbol{\beta}) = \mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta} \in \mathbb{R}^8.$$

The remaining problem: which coefficient vector makes these residuals smallest?

Least squares minimises the average of squared residuals

The name is literal: compare all allowed linear rules,

$$\mathcal{F}_{\text{lin}} = \left\{ f_{\boldsymbol{\beta}}(\mathbf{x}) = \widetilde{\mathbf{x}}^{\top}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^5 \right\}.$$

For each candidate, square its eight residuals and average them:

$$\widehat{\mathcal{R}}_8(f_{\boldsymbol{\beta}}) = \frac{1}{8}\sum_{i=1}^{8}\left(y_i - \widetilde{\mathbf{x}}_i^{\top}\boldsymbol{\beta}\right)^2 = \frac{1}{8}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2.$$

Choose the coefficient vector with the smallest score:

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^5} \widehat{\mathcal{R}}_8(f_{\boldsymbol{\beta}}).$$

Read it as a recipe: predict, compare, square, average, choose the lowest score.

Least squares is empirical-risk minimisation over linear rules under squared loss.

A10–A13 · Least-squares derivation →

$\|\mathbf{r}\|_2^2 = \sum_i r_i^2$. The factor $1/8$ makes the score an average but does not change which $\boldsymbol{\beta}$ minimises it. Squared loss alone does not make the conditional mean linear.

Calculus finds the coefficients: the normal equations

Part 4's move at full scale. Differentiate the objective with respect to the coefficient vector and set every slope to zero at once:

$$\nabla \widehat{\mathcal{R}}_8(f_{\boldsymbol{\beta}}) = \frac{2}{8}\,\widetilde{\mathbf{X}}^{\top}\big(\widetilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}\big) = \mathbf{0} \quad \Longleftrightarrow \quad \widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}\,\widehat{\boldsymbol{\beta}} = \widetilde{\mathbf{X}}^{\top}\mathbf{y}.$$

Five coefficients, five slopes, five equations: the *normal equations*. The objective is a convex quadratic, so a zero gradient marks a global minimum.

For the eight stores the five columns of $\widetilde{\mathbf{X}}$ are linearly independent, so the solution is unique:

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\top}\mathbf{y}.$$

Write the loss, differentiate, set to zero: the Part-4 routine, now producing a full model.

A10–A13 · Expansion, rank, and projection →

Software solves this system by QR or SVD rather than forming the inverse; the practical computes $\widehat{\boldsymbol{\beta}}$ for these eight stores.

The fitted rule removes 98.6% of the training-sample baseline error

$$\mathrm{SSE}_{\mathrm{base}} = \sum_{i=1}^{8} (y_i - \bar{y})^2 = 800,$$

$$\mathrm{SSE}_{\mathrm{line}} = \sum_{i=1}^{8} \left(y_i - \widetilde{\mathbf{x}}_i^{\top} \widehat{\boldsymbol{\beta}}\right)^2 \approx 10.87,$$

$$R^2 = 1 - \frac{\mathrm{SSE}_{\mathrm{line}}}{\mathrm{SSE}_{\mathrm{base}}} = 1 - \frac{10.87}{800} \approx 0.986.$$

R^2 is the fraction of the mean-only baseline SSE removed on the same data. Here the training RMSE falls from 10 to about 1.17. With eight cases and five fitted coefficients, this impressive training fit is not reliable evidence about future stores.

On new data R^2 can be negative: a fitted rule can predict worse than the evaluation-set mean.

With an intercept, least squares cannot fit its training data worse than the mean-only rule. That guarantee does not extend to evaluation data.

A fitted rule does not by itself justify a claim

Once $\widehat{\boldsymbol{\beta}}$ is computed:

- **Computation** gives a numerical prediction, $\widehat{f}(\mathbf{x}_{\text{new}}) = \widetilde{\mathbf{x}}_{\text{new}}^{\top}\widehat{\boldsymbol{\beta}}$.
- **Prediction accuracy** needs unseen data matched to future use (Parts 1 and 4).
- A **population claim** needs a target, sampling reasoning, and uncertainty (Part 6).
- A **causal effect** needs the identification assumptions of Part 7.

Optimisation supplies the fitted rule; evidence supplies the claim.

A slope describes the fitted linear surface while the other listed features are held fixed; that algebra alone is not a causal comparison.

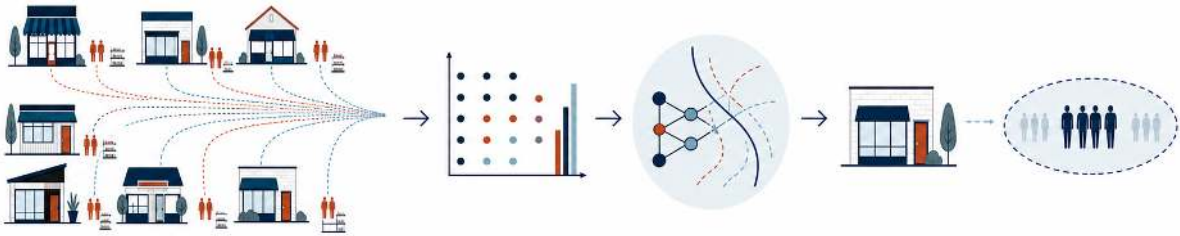
Checkpoint: read the least-squares objective as a recipe

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^5}{\arg\min}\ \frac{1}{8}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2$$

1. For store i, narrate $r_i(\boldsymbol{\beta})$ in words.
2. Where do the linear rule class, squared loss, and empirical average appear?
3. What changes if $1/8$ is removed, and name one claim minimisation alone cannot justify.

If you can narrate every symbol, you can read the training problem.

The foundation built the fit and bounded the claim



Parts 1–5 define the ingredients; Part 8 fits and scores the linear rule; Parts 6–7 bound its claim.

Numbered technical appendix

Verification, derivations, and optional extensions linked from the live lecture.

Reference only: not part of the three-hour live route.

Technical appendix map

Part 3	**A1 empirical moments →** **A2 covariance and correlation →**
Part 4	**A3–A4 conditional-mean proof →** **A5 baseline decomposition →**
Part 5	**A6 sigmoid, log-odds, and binary log loss →**
Part 6	**A7 standard-error calculation →** **A8 bias, variance, and prediction floor →**
Part 7	**A9 optional identification formula →**
Part 8	**A10–A13 objective, normal equations, rank, and projection →**
Workshop	**A14 sources →** **A15 notation reference →**

Main-slide links open the relevant detail; every numbered appendix page returns to its teaching slide.

← **Opening · Foundation and four course days**

Appendix A1: empirical moments of the eight-store sample

Supports: Part 3 · population moments and their sample counterparts

$$\bar{r} = \frac{1}{8}\sum_{i=1}^{8} r_i, \qquad v_8(\mathbf{r}) = \frac{1}{8}\sum_{i=1}^{8}(r_i - \bar{r})^2, \qquad s_8(\mathbf{r}) = \sqrt{v_8(\mathbf{r})}.$$

Variable	Mean	Variance	Standard deviation
Previous demand Q	100	49	7
Store size S	20	49	7
Income I	50	196	14
Promotion D	$1/2$	$1/4$	$1/2$
Demand Y	110	100	10
Capacity K	110	104	$\sqrt{104}$
Event C	$3/8$	$15/64$	$\sqrt{15}/8$

Exact empirical descriptions; no population standard errors or causal estimates.

All entries use divisor 8. The 0/1 variables D and C therefore have sample proportions as their means.

← **Part 3 · Centre and spread**

Appendix A2: empirical covariance and correlation

Supports: Part 3 · covariance, correlation, and the numerical ledger

$$W = (Q, S, I)^\top, \qquad \widehat{\Sigma}_W = \begin{bmatrix} 49 & 67/2 & 39 \\ 67/2 & 49 & 65 \\ 39 & 65 & 196 \end{bmatrix}.$$

$$\widehat{\mathrm{Cov}}_8(W, Y) = \begin{bmatrix} 69 \\ 47 \\ 42 \end{bmatrix}, \qquad \widehat{\boldsymbol{\rho}}_{W,Y} = \begin{bmatrix} 69/70 \\ 47/70 \\ 3/10 \end{bmatrix}.$$

Each entry is a marginal sample association, computed one feature at a time.

All entries use the eight-row empirical distribution with divisor 8. The three elements correspond to Q, S, I, in that order.

← **Part 3 · Covariance and correlation**

Appendix A3: conditional-mean proof (1/2)

Supports: Part 4 · why squared loss targets the conditional mean

Set-up. Under P, fix $X = \mathbf{x}$, write $m_P(\mathbf{x}) = \mathbb{E}_P[Y \mid X = \mathbf{x}]$, and choose one number a to minimise $\mathbb{E}_P[(Y-a)^2 \mid X = \mathbf{x}]$.

Step 1: add and subtract $m_P(\mathbf{x})$, which changes nothing:

$$Y - a = \big(Y - m_P(\mathbf{x})\big) + \big(m_P(\mathbf{x}) - a\big).$$

Step 2: square both sides.

$$(Y-a)^2 = \big(Y - m_P(\mathbf{x})\big)^2 + 2\big(Y - m_P(\mathbf{x})\big)\big(m_P(\mathbf{x}) - a\big) + \big(m_P(\mathbf{x}) - a\big)^2.$$

Step 3: average at $X = \mathbf{x}$**,** term by term:

$$\underbrace{\mathbb{E}_P\big[(Y - m_P(\mathbf{x}))^2 \mid X = \mathbf{x}\big]}_{= \mathrm{Var}_P(Y \mid X = \mathbf{x})} + \underbrace{2\big(m_P(\mathbf{x}) - a\big)\mathbb{E}_P\big[Y - m_P(\mathbf{x}) \mid X = \mathbf{x}\big]}_{= 0} + \underbrace{\big(m_P(\mathbf{x}) - a\big)^2}_{\text{a fixed number}}$$

Given $X = \mathbf{x}$, both $m_P(\mathbf{x})$ and a are constants. The middle bracket vanishes by the definition of the conditional mean; the first term is the conditional variance under P.

← Part 4 · Squared-loss target A4 · Read the result →

Appendix A4: conditional-mean proof (2/2)

Supports: Part 4 · interpreting the decomposition and the constant baseline

The identity from the previous slide:

$$\mathbb{E}_P\big[(Y-a)^2 \mid X=\mathbf{x}\big] = \mathrm{Var}_P(Y \mid X=\mathbf{x}) + \big\{m_P(\mathbf{x})-a\big\}^2.$$

Minimising. The first term does not contain a; the second is never negative and is zero only at $a = m_P(\mathbf{x})$. Hence

$$\arg\min_{a\in\mathbb{R}} \mathbb{E}_P\big[(Y-a)^2 \mid X=\mathbf{x}\big] = \{m_P(\mathbf{x})\}, \qquad \text{with minimum } \mathrm{Var}_P(Y \mid X=\mathbf{x}).$$

Two special cases of the same algebra:

- drop the conditioning: the best constant under P is μ_Y, leaving $\mathrm{Var}(Y)$;
- replace $\mathbb{E}_P$ by the sample average: $L(c) = v_8(\mathbf{y}) + (\bar{y}-c)^2$, the baseline slide of Part 4.

← **Part 4 · Squared-loss target**

The derivation needs $\mathbb{E}[Y^2 \mid X=\mathbf{x}] < \infty$ so that the conditional variance exists. Nothing else is assumed: no linearity, and no model.

Appendix A5: the constant-baseline decomposition

Supports: Part 4 · turning $\frac{1}{n}\sum(y_i - c)^2$ into a spread plus a distance

Step 1: insert the sample mean, which changes nothing because it is added and subtracted at once:

$$L(c) = \frac{1}{n}\sum_{i=1}^{n}(y_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}\big[(y_i - \bar{y}) + (\bar{y} - c)\big]^2.$$

Step 2: expand the square inside the sum, term by term:

$$L(c) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{v_8(\mathbf{y}),\ \text{no } c} + \underbrace{\frac{2(\bar{y} - c)}{n}\sum_{i=1}^{n}(y_i - \bar{y})}_{=0} + \underbrace{(\bar{y} - c)^2}_{\text{same for every } i}.$$

Step 3: the middle term vanishes, because deviations from their own mean always cancel: $\sum_i (y_i - \bar{y}) = 0$. What remains is

$$L(c) = v_8(\mathbf{y}) + (\bar{y} - c)^2, \qquad \arg\min_{c \in \mathbb{R}} L(c) = \bar{y}, \qquad \min_{c \in \mathbb{R}} L(c) = v_8(\mathbf{y}).$$

For the eight stores $v_8(\mathbf{y}) = 100$ and $\bar{y} = 110$, so $L(c) = 100 + (110 - c)^2$: at $c = 110$ it is 100, at $c = 100$ it is 200. This is the sample version of the population identity in A3–A4, with the average over P replaced by the average over the observed rows.

← **Part 4 · Constant baseline**

Appendix A6: sigmoid, log-odds, and binary log loss

Supports: Part 5 · moving between a score, a probability, and its loss

Start with the sigmoid probability

$$p = \frac{1}{1 + e^{-\eta}}.$$

Then

$$1 - p = \frac{e^{-\eta}}{1 + e^{-\eta}}, \qquad \frac{p}{1 - p} = e^{\eta}, \qquad \ln\left(\frac{p}{1 - p}\right) = \eta.$$

For an observed binary outcome $y \in \{0, 1\}$, the probability assigned to what happened is

$$q = p^{y}(1 - p)^{1 - y}.$$

Therefore its case-level log loss is

$$-\ln q = -\big[y \ln p + (1 - y) \ln(1 - p)\big].$$

The sigmoid maps score to probability; the logit reverses it; log loss scores the observed outcome.

← Part 5 · Sigmoid ← Part 5 · Log loss

Appendix A7: estimating the standard error

Supports: Part 6 · standard errors and repeated-sample variation

For iid observations with population variance σ_Y^2,

$$\mathrm{SE}(\bar{Y}_n) = \frac{\sigma_Y}{\sqrt{n}}.$$

Estimate σ_Y^2 with the conventional divisor-$(n-1)$ sample variance:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{800}{7} \approx 114.3, \qquad s_Y \approx 10.69.$$

Hence

$$\widehat{\mathrm{SE}}(\bar{Y}_8) = \frac{s_Y}{\sqrt{8}} \approx 3.78.$$

This estimates repeated-sample variation of the mean, not variation across individual stores.

The core slide's $v_8(\mathbf{y}) = 100$ uses divisor 8 because it describes the eight observed rows exactly. The divisor-$(n-1)$ quantity here serves the different task of estimating a population variance.

← **Part 6 · Standard error**

Appendix A8: bias, variance, and the prediction floor

Supports: Part 6 · what more data can and cannot reduce

For an estimator $\widehat{\theta}$ of θ_0,

$$\mathbb{E}_P[(\widehat{\theta} - \theta_0)^2] = \mathrm{Var}_P(\widehat{\theta}) + \mathrm{Bias}_P(\widehat{\theta})^2.$$

- For many fixed procedures under iid sampling, more data reduces sampling variance.
- Bias can depend on n; more data does not automatically remove systematic bias.

For a fixed rule f under squared loss, let $m_{P_*}(\mathbf{x}) = \mathbb{E}_{P_*}[Y^* \mid X^* = \mathbf{x}]$. Then

$$\mathcal{R}_{P_*}(f) = \mathbb{E}_{P_*}\left[\{m_{P_*}(X^*) - f(X^*)\}^2\right] + \mathbb{E}_{P_*}\left[\mathrm{Var}_{P_*}(Y^* \mid X^*)\right].$$

The first term is error relative to the best squared-loss point prediction; the second is the irreducible outcome-variation floor.

More data can lower prediction error towards a floor, but not generally to zero.

← **Part 6 · Uncertainty and prediction floor**

Appendix A9: optional causal identification formula

Supports: Part 7 · what the causal assumptions identify; optional and not assessed

Under consistency, conditional exchangeability given W, and overlap,

$$\mathbb{E}[Y(d)] = \mathbb{E}_W[\mathbb{E}[Y \mid D = d, W]], \qquad d \in \{0, 1\}.$$

Therefore,

$$\text{ATE} = \mathbb{E}_W[\mathbb{E}[Y \mid D = 1, W] - \mathbb{E}[Y \mid D = 0, W]].$$

This is an identification result, and it holds only under the three stated assumptions.

$\mathbb{E}_W$ averages over the population distribution of the pre-promotion covariates W, so each profile is weighted by how common it is. Under the assumptions, the effect is identified from the observed-data distribution. Estimating it well still requires adequate support, sample size, and a suitable estimator; fit alone cannot establish exchangeability.

← **Part 7 · Identification assumptions**

Appendix A10: least-squares derivation (1/4)

Supports: Part 8 · expanding the empirical-risk objective

Start from the same empirical-risk objective as the core slide:

$$\begin{aligned} J(\boldsymbol{\beta}) &= \frac{1}{n}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2 \\ &= \frac{1}{n}(\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^\top(\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) \\ &= \frac{1}{n}\left[\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\widetilde{\mathbf{X}}^\top\mathbf{y} + \boldsymbol{\beta}^\top\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\boldsymbol{\beta}\right]. \end{aligned}$$

The loss is a quadratic function of the candidate coefficients β.

Here $n = 8$, $\widetilde{\mathbf{X}}$ is 8×5, and $\boldsymbol{\beta} \in \mathbb{R}^5$. The scalar factor $1/n$ is retained to match empirical risk.

← **Part 8 · Least-squares objective** **A11 · Normal equations →**

Appendix A11: least-squares derivation (2/4)

Supports: Part 8 · differentiating the objective

Differentiate the quadratic with respect to $\boldsymbol{\beta}$:

$$\nabla J(\boldsymbol{\beta}) = \frac{2}{n}\widetilde{\mathbf{X}}^{\top}(\widetilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}).$$

At any least-squares minimiser, the gradient is zero:

$$\nabla J(\widehat{\boldsymbol{\beta}}) = \mathbf{0} \quad \Longleftrightarrow \quad \widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} = \widetilde{\mathbf{X}}^{\top}\mathbf{y}.$$

These are the normal equations.

$\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}}$ is positive semidefinite, so J is convex and every solution of the normal equations is a global minimiser. The derivative uses $\nabla_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{\top}M\boldsymbol{\beta}) = 2M\boldsymbol{\beta}$ for symmetric M.

← **Part 8 · Least-squares objective** **A12 · Rank and uniqueness** →

Appendix A12: least-squares derivation (3/4)

Supports: Part 8 · when the coefficient solution is unique

For this eight-store design, the five columns are linearly independent:

$$\operatorname{rank}(\widetilde{\mathbf{X}}) = 5 \quad \Longrightarrow \quad \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \text{ is invertible.}$$

The normal equations therefore have the unique coefficient solution

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \mathbf{y}.$$

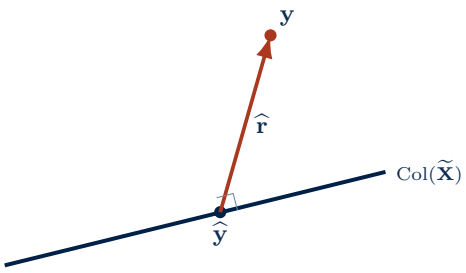
Without full column rank, coefficients may be non-unique; the in-sample fitted vector remains unique.

Rank-deficient new-point predictions need not be unique. Numerical software normally uses QR or SVD rather than explicitly forming the inverse displayed above.

← **Part 8 · Least-squares objective** **A13 · Projection geometry** →

Appendix A13: least-squares derivation (4/4)

Supports: Part 8 · the geometric meaning of the fitted values and residuals



The normal equations are the orthogonality statement

$$\widehat{\mathbf{r}} = \mathbf{y} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}, \qquad \widetilde{\mathbf{X}}^{\top}\widehat{\mathbf{r}} = \mathbf{0}.$$

Thus $\widehat{\mathbf{y}} = \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}$ is the point in $\mathrm{Col}(\widetilde{\mathbf{X}})$ nearest to $\mathbf{y}$. Under full column rank,

$$\mathbf{P}_{\widetilde{X}} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^{\top}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\top}.$$

$$\widehat{\mathbf{y}} = \mathbf{P}_{\widetilde{X}}\mathbf{y}, \quad \widehat{\mathbf{r}} = (\mathbf{I}_8 - \mathbf{P}_{\widetilde{X}})\mathbf{y}.$$

← **Part 8 · Least-squares objective**

This perpendicular geometry is in the eight-dimensional response space, not the two-dimensional feature plot. It is an in-sample algebraic fact; it does not establish future predictive validity or causality.

Appendix A14: sources

Supports: Whole workshop · mathematical, statistical, and causal references

- Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*.
- James, Witten, Hastie, Tibshirani, and Taylor, *An Introduction to Statistical Learning*.
- Murphy, *Probabilistic Machine Learning: An Introduction*.
- Wooldridge, *Introductory Econometrics: A Modern Approach*.
- Imbens and Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*.

← **Opening · Foundation and four course days**

Appendix A15: notation reference for Days 1–4

Supports: Whole course · symbols the four days use without pausing

Symbol	Read it as
$\sum_{i \in G}$	sum over the cases belonging to group G
$\arg\min_x g(x)$, $\arg\max$	the x where g is smallest / largest, not the value of g there
$F_m = F_{m-1} + \eta\, h_m$	a rule built in rounds: last round's rule plus a small correction (Day 3, boosting)
$\widehat{\theta}$, $\bar{y}$, $\mathcal{F}$, $\mathbf{X}$	hat: computed from the sample; bar: an average; calligraphic: a set of rules or settings; bold: a vector or matrix
β, γ	adjustable coefficients η: a linear score or a learning rate, as stated μ: mean, or a cluster centre (Day 4)
σ, σ^2, ρ, τ	standard deviation, variance, correlation, decision threshold
λ	eigenvalue: the variance carried by a principal component (Day 4)
Δ	a change, e.g. a daily yield change (Day 4)

Two collisions to expect: p is the number of features in Parts 1–2 but a probability in classification; η is a linear score in logistic regression but a learning rate in gradient descent and boosting. Context and subscripts distinguish them. Uppercase K is capacity in the store example; lowercase k counts clusters on Day 4.

Print this page as a bookmark for the four course days.

← **Opening · Foundation and four course days**