Econometrics Across Disciplines

Measurement, models, and decisions

Ian Helfrich, PhD

October 2026

Table of contents

Reading this working edition **4**
Four reading levels 4
Claims and calculations 4

Economic questions, useful predictions, and interventions **6**
○ Orientation: The forecast meeting 6
○ Orientation: What would count as an answer? 6
○ Orientation: Six possible days 7
◇ Core: Averages with stated information 10
◇ Core: Why the conditional average minimizes squared error 10
◇ Core: Several mechanisms, identical observations 12
◇ Core: Assigning the alert 13
○ Orientation: A warning as information and an action 14
◆ Theory: What the records cannot identify 14
◇ Core: A design that supplies the missing comparison 16
◇ Core: A prediction and a recommendation 16
◇ Core: One declared calculation in three languages 17
✦ Workshop: Diagnose and repair a claim 17
○ Orientation: The professional handoff 19
Assignment comparison 20

Constructing a measured world **21**
○ Orientation: A reading is the end of a process 21
○ Orientation: Power, energy, and an axis label 22
◇ Core: A controlled meter and billing system 23
◇ Core: How a line summarizes co-movement 26
◆ Theory: Why classical error flattens this slope 27
◇ Core: A calibration error can behave differently 28
◆ Theory: Recovering a coefficient and predicting a bill 29
◆ Theory: What independent repeated readings can identify 29
◇ Core: A denominator can manufacture growth 30
◆ Theory: Aggregation changes the weights 31
✦ Workshop: Worked exercises and an audit 31
Professional handoff: a measurement dictionary and memo 32
Instrument comparison 34

Observation clocks, releases, vintages, and provenance **35**
○ Orientation: A quarter ends before its estimate arrives 35
◇ Core: A growth rate also has a convention 36
◇ Core: Selecting a value as of a decision time 36
◆ Theory: A future release cannot change an earlier calculation 39
◇ Core: A forecast can leak without seeing its target 40
◆ Theory: Information groups possible worlds 40
◆ Theory: More information improves the population optimum 42
◇ Core: The target has a publication history too 43
✦ Workshop: A release audit and worked answers 43
Professional handoff: A reproducible information ledger 44
Release comparison 46

Reading this working edition

Econometrics begins with a question about a measured world. This handbook develops the mathematical tools needed to state that question, investigate it, and explain the conditions under which an answer follows. Concrete examples and pictures accompany fully explained proofs and independent Python, Julia, and R calculations.

This working edition contains three foundation chapters: economic questions, measurement, and observation clocks. The fifteen-lecture time-series course is published separately at ihelfrich.github.io/time-series and keeps its own lecture numbering.

Four reading levels

○ **Orientation** begins with a concrete question and defines unfamiliar terms. Arithmetic and pictures support a route for a reader whose mathematical training is still developing.

◇ **Core** develops the calculation, its assumptions, and its economic meaning. Later chapters build the algebra, probability, calculus, and linear algebra they require.

◆ **Theory** gives precise definitions and complete proofs. Chapter 1's finite identification argument can be followed before advanced probability. Later technical passages add the spaces, convergence, and identification conditions needed for doctoral work.

✦ **Workshop** gives a bounded exercise or research investigation. Its status is explicit. Chapter 1's workshop consists of worked practice in a controlled example, not new empirical findings or open research problems.

The glyphs accompany text labels. They indicate a passage's requirements and purpose, not a reader's ability. The introductory route retains the assumptions and meaning of a result even when its more technical proof is deferred.

Claims and calculations

The first chapter supplies a question contract, exact finite calculations, a proof of the squared-error forecast rule, and proofs of nonidentification and the limits of estimation. The second

supplies a measurement dictionary, proofs of the population projection and classical attenuation, a repeated-reading correction with explicit restrictions, calibration failures, and denominator/aggregation audits. The third builds an as-of release ledger, proves prefix invariance, develops finite filtrations and adapted forecasts, proves the information-gain identity, and separates input and target vintages. Direct section links name the definition or result being used. The first two chapters use entirely controlled examples. The third adds a small attributed BEA release-history extract alongside separate controlled models; it does not fit a GDP nowcast. The empirical Texas electricity dataset is separate and does not estimate the alert effect or validate the controlled meter.

Python, Julia, and R independently reproduce the declared numerical quantities. Their agreement is a calculation check. Mathematical claims rely on the written proofs, and economic validity relies on the stated model and design. Source, input manifest, native verification results, and original figures accompany this edition.

Public practice includes complete worked answers. Live assessments, grading keys, instructor scripts, generator truth, and held-out graded continuation remain private.

Working edition, October 2026. © Ian Helfrich. All rights reserved. Authored text, code, and figures remain author-owned. Cited works are linked and not redistributed. Bundled font files retain their included licenses.

Online handbook · Chapter 1 · Three-language code

Economic questions, useful predictions, and interventions

○ Orientation: The forecast meeting

A utility sends an alert when it expects a hot day. The alert asks customers to reduce electricity use. Its records show higher peak demand on days with alerts. One analyst wants to use the alert to predict tomorrow's demand. Another wants to know whether sending more alerts would reduce demand. A manager needs to decide whether the reduction would cover the program's cost.

All three questions concern the same two columns. Each requires a different argument. An alert may carry useful information about heat and still reduce the demand that heat would otherwise produce. The demand observed after an alert does not reveal the demand that the same day would have had without it.

We will build a small economy in which every possibility can be written down. It is a controlled example, with invented numbers and known mechanisms. The observed Texas electricity case uses actual provider data for forecasting; this chapter does not estimate an alert effect from those data.

Our example has six possible kinds of day. Working with six makes the reasoning inspectable. The central difficulty will survive when the number of recorded days grows.

○ Orientation: What would count as an answer?

An **observation** is a recorded value, such as a day's peak demand. A **population** is the collection of days or units the question concerns. A **target** is the quantity we want to learn about that population. A target stated mathematically is often called an **estimand**. An **estimator** is a calculation that uses recorded observations to estimate it. An **estimate** is the number produced by that calculation.

Suppose tomorrow's alert is known before the demand forecast is made. A forecasting question asks for average demand among days with that alert status. An intervention question asks how demand would change if we assigned the alert status, keeping the day and the specified background conditions comparable. A decision question adds costs, benefits, constraints, and whose interests count.

Before calculating, write a short question contract:

Item	Forecast	Intervention
Population	Days served by this utility under the current alert rule	Those same days under an assigned alert rule
Outcome	Daily peak load, measured in megawatts (MW)	The same peak-load measure
Target	Average load given the available alert status	Average load if an alert is assigned, compared with no alert
Information	The issued alert; load has not yet occurred	An explicit model or design for assigning the alert
Horizon	Tomorrow's daily peak	The daily peak following the assigned alert
Error or value	Squared forecast error, in MW^2	A specified value of avoided peak load, less program cost
Permissible claim	A prediction under the stated information and population	A response under the stated intervention and assumptions

The outcome definition needs care. **Power**, measured in MW, is a rate of energy use. **Energy**, measured in megawatt-hours, accumulates over time. The same distinction appears between water flowing through a tap and water accumulated in a tank: the flow rate and the accumulated quantity answer different questions. A day's peak power does not give its total energy use. We will keep MW throughout the load calculations. The unit check deliberately converts MW to kilowatts and shows why squared error scales by a million.

The contract also records when information becomes available. An alert known after the peak cannot serve as an input to tomorrow's forecast. The time-series chapter's reference-time and release-time distinction explains this timing issue in detail. Here the alert is available beforehand by construction.

○ Orientation: Six possible days

Let U indicate an unrecorded heat condition: $U = 0$ for an ordinary day and $U = 1$ for a hot day. Each has probability one half. Let V be a demand shock, in MW, taking values $-1, 0$, and 1 with equal probability. We assume the heat condition and shock are **independent**: knowing one does not change the probabilities of the other. Every pair in the table therefore has probability $(1/2)(1/3) = 1/6$.

Let A indicate the alert: zero means no alert and one means an alert. The current rule is $A = U$. In this example the utility's internal heat condition is absent from the analyst's records. The analyst sees only A and peak load Y.

Day type	Heat condition U	Shock V, MW	Alert A	Observed load Y, MW	Probability
1	0	−1	0	9	1/6
2	0	0	0	10	1/6
3	0	1	0	11	1/6
4	1	−1	1	11	1/6
5	1	0	1	12	1/6
6	1	1	1	13	1/6

These rows list possibilities and their probabilities. They are not six sampled observations. A randomly recorded sequence of days could contain one type repeatedly and omit another. Enumeration gives exact population calculations for this declared model; it does not eliminate sampling uncertainty in an empirical study.

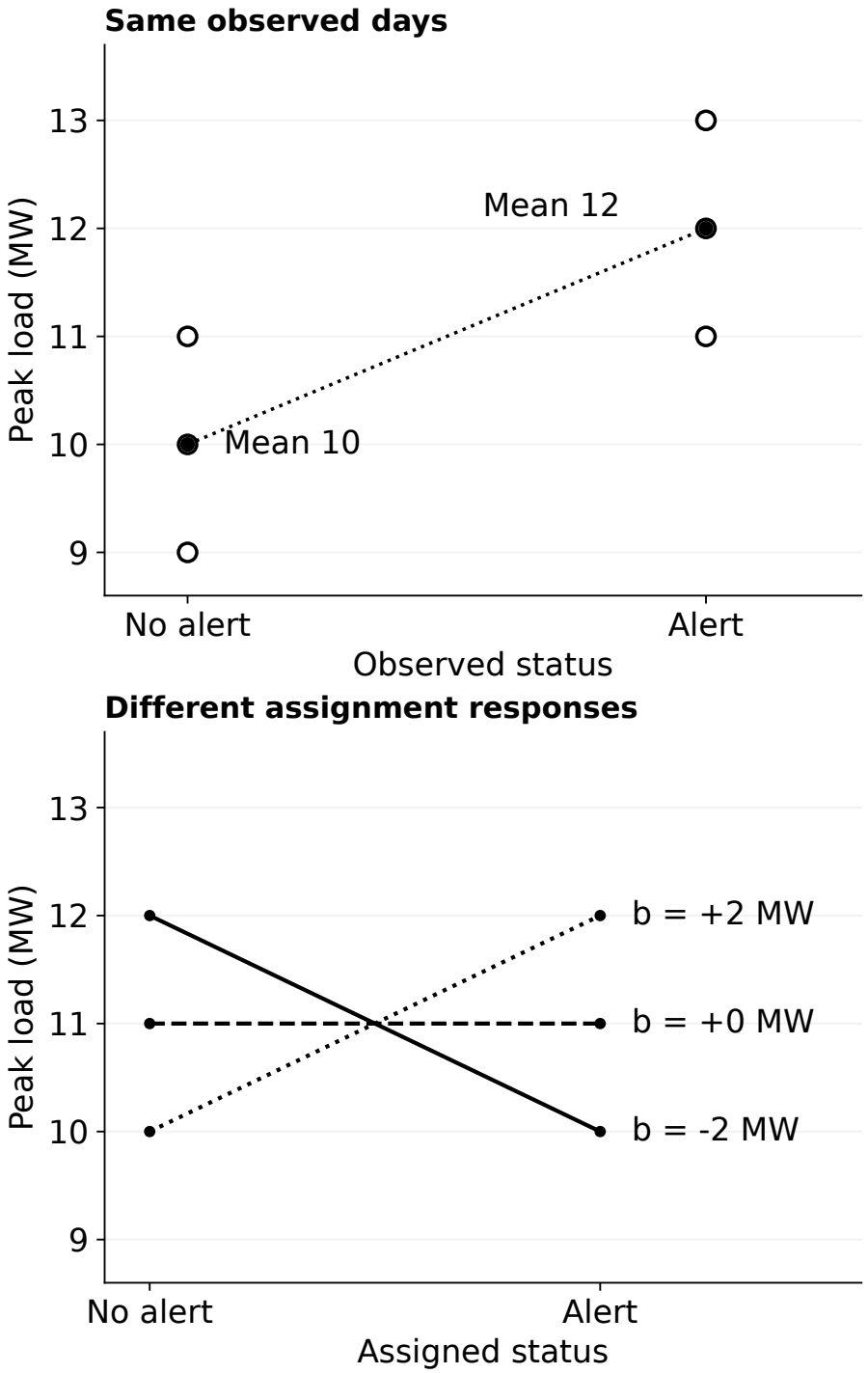


Figure 1: The first panel is common to every mechanism below. The second panel compares population averages under assigned alert status; each line uses the same distribution of heat conditions and shocks. The line labels give the intervention effect in MW. The intervention calculation explains each endpoint.

◇ Core: Averages with stated information

Write p_i for the probability of day type i and y_i for its load. The symbol $\sum$ means "add the following terms over the listed day types." The **expectation** of load is the probability-weighted average:

$$E[Y] = \sum_{i=1}^{6} p_i y_i = \frac{9 + 10 + 11 + 11 + 12 + 13}{6} = 11 \text{ MW}.$$

The square brackets identify the quantity being averaged. They do not mean that a sampled average must equal eleven.

To **condition** on an alert, retain the three alert-day possibilities and give them probabilities that add to one. Their original probabilities add to $3/6 = 1/2$. Each retained probability becomes $(1/6)/(1/2) = 1/3$. Thus

$$E[Y \mid A = 0] = \frac{9 + 10 + 11}{3} = 10, \qquad E[Y \mid A = 1] = \frac{11 + 12 + 13}{3} = 12.$$

The vertical bar means "given." The difference of these conditional averages is $12 - 10 = 2$ MW. It describes two groups of days selected by the existing alert rule. We have not yet calculated what changing that rule would do.

The rule $m(A) = 10 + 2A$ gives the correct conditional average for either alert status. The letter m names a function: insert zero to obtain ten, or one to obtain twelve. Its error is $Y - m(A) = V$, so the error takes values -1, 0, and 1 equally often in each group. The **mean squared error**, or MSE, is the average squared error:

$$E[(Y - m(A))^2] = \frac{(-1)^2 + 0^2 + 1^2}{3} = \frac{2}{3} \text{ MW}^2.$$

The model contains unpredictable variation even when its conditional average is known exactly. Prediction accuracy concerns the declared information and loss; it does not establish an intervention effect.

◇ Core: Why the conditional average minimizes squared error

Result 1. Within this six-day model, among all forecasts that use only A, the function $m(A) = E[Y \mid A]$ uniquely minimizes population mean squared error. Both alert groups have positive probability, so uniqueness applies to each of the two forecast values.

The proof has five steps and uses no calculus.

1. **Fix the information.** Choose an alert value a, either zero or one, and let q be a proposed forecast for that group. Write m_a for its conditional average: $m_0 = 10$ and $m_1 = 12$, as established in the finite-average calculation. Add and subtract m_a:

$$Y - q = (Y - m_a) + (m_a - q).$$

2. **Expand the squared error.** Multiplying $(r+s)(r+s)$ gives $(r+s)^2 = r^2 + 2rs + s^2$. Apply that identity to the two terms above, then average within the group:

$$E[(Y-q)^2 \mid A = a] = E[(Y - m_a)^2 \mid A = a] \\ + 2(m_a - q)E[Y - m_a \mid A = a] + (m_a - q)^2.$$

   Constants can be taken outside a finite weighted sum because each term contains the same factor.

3. **Remove the cross term.** By the definition of m_a, the average deviation from it is zero:

$$E[Y - m_a \mid A = a] = E[Y \mid A = a] - m_a = 0.$$

   The middle term in step 2 therefore vanishes. Using the residual average square calculated in Averages with stated information leaves

$$E[(Y-q)^2 \mid A = a] = \frac{2}{3} + (m_a - q)^2.$$

4. **Locate the minimum within the group.** The first term is the irreducible error. The second is a square, which is nonnegative and equals zero exactly when $q = m_a$. This proves the claim within either alert group.

5. **Average across groups.** For an arbitrary two-value forecast g, weight the two group errors by their probabilities of one half:

$$E[(Y - g(A))^2] = \frac{2}{3} + \frac{1}{2}(10 - g(0))^2 + \frac{1}{2}(12 - g(1))^2.$$

   Both extra terms must vanish to attain $2/3$. A constant eleven-MW forecast pays an extra $(1/2)(1)^2 + (1/2)(-1)^2 = 1$ MW², giving MSE $5/3$.

The alert is useful for this prediction. The proof has said nothing about the consequence of assigning one.

Every average here is finite and has positive conditioning probability. For an unrestricted random outcome, the corresponding squared-error argument requires a finite second moment. The existing course's conditional-forecast proof gives the general version; the calculation above supplies its finite foundation.

◇ Core: Several mechanisms, identical observations

We now specify how load responds to an alert and a heat condition. A **structural equation** describes that response together with the rule generating the alert. Consider the following family:

$$A = U, \qquad Y = 10 + bA + (2 - b)U + V, \qquad -2 \leq b \leq 2.$$

The parameter b measures the load response, in MW, to changing A from zero to one while keeping U and V fixed. The coefficient $2 - b$ measures the heat contribution in this family. The parameter range is part of the example's model class; it is not a confidence interval inferred from observations. It keeps all of the assigned-alert loads positive in our six day types.

The family keeps the following assumptions explicit:

- The current operating rule is $A = U$.
- The six background states and their probabilities are the ones in Six possible days, including independence of heat and the shock.
- The load equation applies across both assigned alert values. Assignment does not change the background distribution or add another causal pathway.
- The model class permits $-2 \leq b \leq 2$. The data do not supply that restriction.

Substitute $A = U$ into the load equation:

$$Y = 10 + bU + (2 - b)U + V = 10 + \{b + (2 - b)\}U + V = 10 + 2U + V.$$

The first equality uses the alert rule. The second collects terms multiplying the same heat indicator. The third uses $b + 2 - b = 2$. The resulting observed equation contains no b.

This cancellation holds separately for every row, not merely for an average or correlation. Changing b leaves every observed pair (A, Y) unchanged. Since each row also retains probability $1/6$, the entire observable probability distribution is unchanged. Such mechanisms are **observationally equivalent** for these records.

In a world with $b = -2$, the alert reduces load by two MW while the heat contribution is four MW. With $b = 0$, the alert has no effect and heat contributes two MW. With $b = 2$, the alert increases load by two MW and the direct heat contribution in this simple equation is zero. All three worlds produce the table in Six possible days.

The example establishes a logical possibility. It does not claim that the third mechanism describes any particular utility or that real demand has this simple form. Restrictions excluding a mechanism must have a substantive or design justification. An observed alert–load association alone cannot provide it.

◇ Core: Assigning the alert

An intervention replaces the rule generating the variable being assigned. For do$(A = a)$, replace $A = U$ by $A = a$. Keep the load equation, the six heat–shock possibilities, and their probabilities fixed. This definition specifies what remains comparable across the two assignments. It assumes that the assignment itself has no additional effect outside the stated equation.

Under that intervention the load of a day with condition u and shock v is

$$Y(a; u, v) = 10 + ba + (2 - b)u + v.$$

The notation before the semicolon gives the assigned alert; the notation after it identifies the same background day. For each day,

$$Y(1; u, v) - Y(0; u, v) = [10 + b + (2 - b)u + v] - [10 + (2 - b)u + v] = b.$$

All shared terms cancel. The difference is the same across the six types because this model assumes a constant alert response. Averaging a constant leaves it unchanged, so the population intervention effect is also b.

To calculate the two intervention levels, use $E[U] = 1/2$ and $E[V] = (-1 + 0 + 1)/3 = 0$. Linearity follows by distributing and regrouping their finite weighted sums:

$$E[Y \mid \text{do}(A = a)] = 10 + ba + (2 - b)\frac{1}{2} + 0 = 11 + b\left(a - \frac{1}{2}\right).$$

Mechanism	Assigned no alert, MW	Assigned alert, MW	Intervention difference, MW	Observed group difference, MW
$b = -2$	12	10	−2	2
$b = 0$	11	11	0	2
$b = 2$	10	12	2	2

Conditioning on $A = 1$ selects only hot days under the old rule. Assigning $A = 1$ retains both ordinary and hot days in the population. That change in which backgrounds are being compared explains the different answers. The finite-average calculation and the intervention definition together identify exactly where the two operations diverge.

The intervention convention is established causal-inference machinery. Pearl, Glymour, and Jewell's *Causal Inference in Statistics: A Primer*, §3.1, printed pp. 53–55 explains equation replacement and conditioning; our finite utility example and calculations are authored here. The chapter links to their text and does not redistribute it.

The assignment experiment lets you choose a mechanism, predict the direction of the comparison, and then inspect the two means and all six day types. Change the mechanism and explain why the observed comparison stays fixed. The table above and worked exercise 1 give the same investigation in print.

○ Orientation: A warning as information and an action

A sensor warning offers a useful connection outside economics. Suppose a machine has a high-heat state U and a display records it as $A = U$. If the display is isolated from the machine's thermal process and nobody changes the machine in response to it, changing only the display has no effect on the machine's temperature. That claim depends on the stated isolation and absence of feedback. It does not follow from an association between the display and temperature.

The sensor and utility examples share the assignment relation $A = U$. In both, observing the warning selects a background state. The utility alert also asks people to act, so the load equation allows a response to A. Replacing the display in the isolated sensor example and sending a demand-response alert therefore require different substantive restrictions, even though their observational calculations have the same form. The equation-replacement definition in Assigning the alert identifies the common mathematical operation and the mechanism that must be specified separately.

A second connection concerns dynamics. Consider a specified physical relaxation model in which a fixed fraction ϕ, with $0 < \phi < 1$, of a deviation from a fixed reference temperature remains after each equal sampling interval. Its gap satisfies $x_{k+1} = \phi x_k$. Repeated substitution gives $x_k = \phi^k x_0$: each interval multiplies the previous gap by the same fraction. These assumptions exclude changing reference temperatures and multiple interacting relaxation modes. The course's difference-equation derivation applies the same multiplication rule to an economic deviation and then accounts for additional disturbances. Economic persistence needs its own interval, disturbance model, and interpretation; the shared recursion alone does not establish a thermal explanation for an economic series.

◆ Theory: What the records cannot identify

Definition. A target is **identified** in a model class if every mechanism in that class yielding the same observable distribution gives the same target value. The phrase "in a model class" matters: observations and restrictions jointly determine identification.

Result 2. In the family $-2 \le b \le 2$ specified in Several mechanisms, the alert's population intervention effect is not identified from (A, Y). Its identified set is the entire interval $[-2, 2]$.

Proof. The row-by-row substitution in Several mechanisms proves that every allowed b gives the same observable distribution. The paired intervention calculation in Assigning the alert

proves that the target equals b. Every value in the allowed interval is therefore compatible with the records, and no value outside it is admitted by this model class. Those two facts establish the identified set. In particular, the effects -2 and 2 disagree despite identical observations. This violates the definition of point identification.

Result 3. Under independent sampling of days from this same observational distribution, no estimator using only those records can consistently estimate b for every mechanism in this family.

"Independent sampling" means each draw has the six probabilities and learning one day's type does not change another's probabilities. "Consistently" means that, as the number n of recorded days grows, the probability of an estimation error larger than any specified positive tolerance approaches zero. Let T_n denote the estimator, including any internal randomness independent of the data and having the same distribution in every mechanism.

Proof. Consider the worlds $b = -2$ and $b = 2$. They give identical distributions for one observed day. Under the stated sampling rule they give identical distributions for every n-day dataset, because the probability of a particular sequence is the product of the same per-day probabilities. Applying the same estimator with the same independent randomization preserves equality of output distributions. Call that shared distribution Q_n.

Take the tolerance to be one MW. Consistency in the first world would require

$$Q_n\{|T_n + 2| < 1\} \longrightarrow 1.$$

Consistency in the second would require

$$Q_n\{|T_n - 2| < 1\} \longrightarrow 1.$$

The braces specify events, or sets of possible estimator outputs. The first requires $-3 < T_n < -1$; the second requires $1 < T_n < 3$. These intervals are disjoint. For every n, the probabilities of disjoint events add, and their sum is at most one. They cannot both approach one, since their sum would then approach two. This contradiction proves the result.

The proof concerns a declared sampling and model class. A new experiment, additional defensible restrictions, or new observations of relevant mechanisms can change the class or the available information. More records drawn from the unchanged observational process improve knowledge of that process; they do not distinguish the two worlds used in the proof.

Merely recording U also fails to supply the missing comparison here. Within $U = 0$, the old rule always gives $A = 0$; within $U = 1$, it always gives $A = 1$. The other alert status has probability zero in each stratum. A conditional average for that absent status is not available from this design. This is a failure of **overlap**, the availability of both comparison conditions in the relevant population strata.

◇ Core: A design that supplies the missing comparison

Consider a controlled experiment in which a fair, independent coin assigns A. Keep the heat and shock probabilities and the load equation fixed. Each of the twelve triples (A, U, V) now has probability $1/12$.

For either assigned alert status, the two heat conditions remain equally likely and the three shocks retain their original probabilities. Thus

$$E[Y \mid A = a] = 10 + ba + (2 - b)E[U \mid A = a] + E[V \mid A = a] = 11 + b\left(a - \frac{1}{2}\right).$$

The first equality averages the specified load equation. The second uses independence of the randomized alert from both background variables, giving conditional averages $1/2$ and zero. Subtracting the average at $a = 0$ from the one at $a = 1$ leaves b. Both groups have positive probability. The experimental comparison therefore identifies the effect in this model.

Random assignment supplies comparable background distributions; maintaining the same load equation supplies the link to the intervention target. Real experiments must also define the treatment, handle noncompliance and missing outcomes, and consider interference between units. An alert sent to one region may affect neighboring demand. Those questions need an expanded model and are not resolved by this twelve-state demonstration.

Hernán and Robins's *Causal Inference: What If*, chapters 1–3 develops causal effects, randomized experiments, and observational studies through potential outcomes. Its framework gives another precise way to state the comparisons used here. The chapter source register records the particular sections, use, and reading scope of both references.

◇ Core: A prediction and a recommendation

Suppose a planner values an expected one-MW reduction in the daily peak at v dollars and pays c dollars to issue an alert. Under a deliberately simple linear valuation, the incremental net benefit is

$$N(b) = v\{E[Y(0)] - E[Y(1)]\} - c = -vb - c.$$

The braces contain a reduction, so its sign is the negative of the load effect b. Let $v = 1{,}000$ dollars per MW of daily peak reduction and $c = 500$ dollars per day. Then $N(-2) = 1{,}500$, $N(0) = -500$, and $N(2) = -2{,}500$ dollars per day. The same observed distribution is compatible with a profitable alert program and an unprofitable one under this valuation.

This calculation is a stated decision rule, not a complete welfare analysis. A utility's avoided capacity cost, customer inconvenience, distributional effects, and system reliability need their own quantities. Even a causal peak-load estimate would not by itself settle those choices.

For the current forecasting task, the conditional-average forecast remains valid in all three worlds. For the intervention decision, the records leave the sign unresolved. A professional memo should preserve both conclusions. It can recommend the forecast for the defined information set while explaining what evidence the alert decision still needs.

◇ Core: One declared calculation in three languages

The six-state input file contains heat, shock, and an integer probability weight. The six weights add to six; dividing each by that sum produces the probabilities used above. It contains no empirical utility observations. Its input manifest records the schema, units, generator, and file hash.

Independent Python, Julia, and R programs enumerate the same states. Each calculates the observed row values, conditional means, squared-error decomposition, assigned-alert means, randomized-design means, and decision value. The implementation also evaluates $b = -1$ and $b = 1$ between the three plotted mechanisms. Those five values are numerical checks within the continuous family; the algebraic proof covers the full interval.

For example, the Python operation for an assigned alert is `10 + b*a + (2-b)*u + v`. The Julia and R versions implement their own enumeration and weighting. None reads another language's output as its answer. The verification program runs the three programs, compares their results, and checks independent analytic targets and row invariants. The checked results identify the actual native runtimes. This is computational verification of specified finite calculations. The mathematical proof is the separate argument in Results 2 and 3.

Converting load from MW to kW multiplies each load and forecast error by $1{,}000$. Squaring gives $1{,}000^2 = 1{,}000{,}000$, so forecast MSE becomes $(2/3) \times 1{,}000{,}000$ kW². The intervention difference becomes $1{,}000b$ kW. A reported MSE that rises only a thousandfold has mixed up the unit of an error with the unit of its square. Each program checks this conversion explicitly.

The execution guide gives commands and the scope of the checks. These are native programs; the page does not execute Python, Julia, or R in the browser.

✦ Workshop: Diagnose and repair a claim

The following are worked exercises for this chapter's controlled model. They are not claims of new research results. Write your argument before reading the solutions.

1. **A mechanism between the plotted cases.** Set $b = -1$. Calculate the observed group difference and the two assigned-alert means. Does a positive observed difference rule out a load-reducing alert?

   Solution. The observed equation is still $Y = 10 + 2U + V$, so the group difference remains two MW. The formula in Assigning the alert gives $11 + (-1)(0 - 1/2) = 11.5$ MW without the assigned alert and $11 + (-1)(1 - 1/2) = 10.5$ MW with it. The effect is -1 MW. Thus the observed two-MW difference coexists with a one-MW reduction under intervention.

2. **A less useful forecast.** A forecast reports nine MW when $A = 0$ and thirteen MW when $A = 1$. Calculate its MSE without listing all six squared errors.

   Solution. Use the decomposition proved in Result 1. Its extra error is $(1/2)(10 - 9)^2 + (1/2)(12 - 13)^2 = 1$ MW². Adding the unavoidable $2/3$ gives $5/3$ MW². This equals the error of the constant eleven-MW forecast despite using the alert. Using an informative variable does not guarantee that a particular forecasting rule uses it well.

3. **A proposed statistical repair.** An analyst obtains the previously unrecorded heat indicator and proposes comparing alert and non-alert days within each heat category. Which needed cells are absent?

   Solution. Under $A = U$, the pair $(U = 0, A = 1)$ never occurs and the pair $(U = 1, A = 0)$ never occurs. The proposed comparisons require averages in these zero-probability cells. The overlap discussion explains why measuring a variable and having variation conditional on it are separate requirements. The independent assignment in A new design supplies all four heat–alert combinations.

4. **A decision threshold.** Under the linear valuation $v = 1{,}000$ and $c = 500$, which values of b give positive net benefit? Is that condition identified from the old records?

   Solution. Solve $-1{,}000b - 500 > 0$: add five hundred to both sides to obtain $-1{,}000b > 500$, then divide by a negative number, reversing the inequality, to get $b < -0.5$. The identified set $[-2, 2]$ contains values on both sides of this threshold. The sign of net benefit therefore remains unresolved under the stated model class and valuation.

5. **A stronger statement than the evidence allows.** Rewrite "alerts increase demand by two MW, so cancel them." Include the useful finding and the unanswered question.

   Solution. "Under the existing alert rule, alert days have two MW higher mean peak load. The alert improves a conditional-average forecast in the six-state model. These records do not identify the load response to assigning an alert; the defined model class admits effects from −2 to 2 MW. An alert recommendation requires an assignment design or justified restrictions, together with the cost and benefit model."

○ Orientation: The professional handoff

A short memo can state the question, the result, its scope, and the next evidence needed without burying the distinction in technical language:

> The alert is useful for predicting daily peak demand under the current operating rule. In the controlled model, mean peak load is ten MW without an alert and twelve MW with an alert, and the conditional-average forecast has mean squared error of two thirds MW^2. The same records are compatible with alerts reducing, preserving, or increasing demand. They therefore do not support a recommendation to expand or cancel the alert program. A comparable assignment design would address the intervention question. Program costs and the value of avoided load must be assessed separately.

Attach the question contract, the input manifest, and the calculation checks. Link the claim about forecast accuracy to Result 1, the identification limit to Results 2 and 3, and the proposed assignment design to its explicit conditions. This makes the memo traceable for both a manager and another analyst.

The same discipline will recur when an area-level quantity is used to make a claim about a person, a noisy score is used to measure a skill, or a forecast is used to choose a policy. First name the economic object and the comparison. Then establish what the observations and assumptions allow us to learn.

Assignment comparison

The online experiment asks for a prediction before revealing assigned-alert means. These static comparisons use the same six background states under each assignment. The observed group means remain 10 and 12 MW throughout.

Effect b, MW	Assigned no alert, MW	Assigned alert, MW	Paired effect, MW
-2	12	10	-2
-1	11.5	10.5	-1
0	11	11	0
1	10.5	11.5	1
2	10	12	2

Conditioning selects the background types under the existing rule; assignment retains the common background distribution. The intervention proof explains why those operations answer different questions.

Constructing a measured world

○ Orientation: A reading is the end of a process

A facility manager wants to understand the relation between electricity use and the daily bill. A meter supplies a number. An accounting system supplies another. Joining the two files looks straightforward until someone asks what each row means.

Does electricity use mean energy consumed over the whole day, average power during an hour, or the largest instantaneous load? Does the bill include a fixed charge? Do both records describe the same facility and interval? Was the meter calibrated before the readings were collected? A regression cannot settle these questions for us. They determine which regression would have a useful interpretation.

A **measurement map** follows a concept into a recorded field. Begin with the object from chapter 1's question contract, then specify the operation that makes it observable. For daily energy use, the map might be: define the facility boundary; accumulate energy between two meter readings; align the start and end times; convert the unit; record any missing interval; attach the facility identifier. Every operation can change the object represented by the final number.

The distinction matters beyond meters. Annual income may refer to a person or household, before or after tax, and to earnings or all receipts. A skill score represents performance on a specified set of tasks under specified conditions. A temperature field represents a sensor, a location, an averaging interval, and a unit. A network degree counts connections among the nodes retained in a particular graph. The column name alone does not define the quantity.

For each recorded field, write down:

- **Concept and operation:** what is being measured, and how is the value produced?
- **Unit and support:** what does one numerical unit mean, and which person, place, or interval does the row describe?
- **Coverage:** which members of the target population enter the numerator and any denominator?
- **Timing:** which interval is described, when was the value available, and can it be revised?
- **Validation evidence:** what comparison or calibration would support the intended interpretation?

These are substantive choices. If the question concerns access to a service, counting buildings with a connection may miss whether occupants can afford to use it. The recorded field can be reproducible and still answer a narrower question.

Reliability concerns the stability of a measurement under specified repetitions. **Validity** concerns whether the measurement supports its intended interpretation. An incorrectly calibrated instrument can repeat the same wrong value. An AI system can assign the same label every time while its labels measure writing style rather than the economic construct of interest. Repetition gives evidence about stability; a defensible interpretation also needs a measurement map and a relevant external comparison.

○ Orientation: Power, energy, and an axis label

Think of water flowing into a tank. The flow rate describes how quickly water enters. The accumulated water describes how much has arrived. Electrical power and energy have the same rate-versus-accumulation relationship: power describes energy per unit time; energy accumulates power over an interval. The connection is mathematical. A meter's engineering limitations still require their own evidence.

If power remains at 2 MW for three hours, the energy consumed is 6 MWh. More generally, piecewise-constant power readings P_j over intervals of length Δt_j give

$$E = \sum_j P_j \Delta t_j.$$

Here P_j is in MW, each Δt_j is in hours, and E is in MWh. This expression is exact for the declared piecewise-constant path; using intermittent readings to approximate an unknown path introduces an additional approximation. Unequal intervals require their actual lengths. Adding power readings without their intervals does not produce an energy total.

The NIST SI definitions, §2.3.4, Table 4 distinguish the joule as energy and the watt as joules per second. Since an hour contains 3,600 seconds,

$$1\ \text{MWh} = 10^6\ \frac{\text{J}}{\text{s}} \times 3{,}600\ \text{s} = 3.6 \times 10^9\ \text{J}.$$

The peak-load outcome in chapter 1's utility example is power in MW. This chapter's meter measures daily energy in MWh. They describe different objects. Dividing a day's energy by its duration gives average power; it does not recover the day's peak.

Changing MWh to kWh multiplies an energy number by 1,000. A bill remains the same amount of money. Its coefficient in dollars per kWh must therefore be one-thousandth of its coefficient in dollars per MWh. We will check that transformation in each implementation rather than trusting a changed axis label.

◇ Core: A controlled meter and billing system

We can see measurement error clearly in a small world where the underlying quantity is known. This is an authored teaching model, not an observed tariff, a test of actual meter accuracy, or a causal estimate from utility data.

Let X be true daily energy use at a facility, in MWh. It takes values 8 and 12 with equal probability. Let Z_1 and Z_2 independently take values $-1, 0, 1$, each with probability $1/3$. Let ε, independently of all three, take values -20 and 20 dollars with equal probability. Crossing these choices gives 36 equally likely states.

The bill Y and two recorded meter readings M_1, M_2 are

$$\begin{aligned} Y &= 30 + 50X + \varepsilon, \\ M_1 &= X + \kappa(X - 10) + hZ_1, \\ M_2 &= X + \kappa(X - 10) + hZ_2. \end{aligned}$$

The fixed charge is 30 dollars. The constructed energy charge is 50 dollars per MWh. The perturbation ε is in dollars. The noise amplitude h is in MWh; Z_1, Z_2 and the calibration parameter κ are dimensionless. The 10 inside the meter equation denotes 10 MWh. Each addition has compatible units.

When $\kappa = 0$, a reading is true use plus independent, mean-zero noise. When $\kappa = -1/2$, the reading's distance from 10 MWh is compressed by half before random noise is added. When $\kappa = -2$, that distance is reversed. The latter is a deliberately broken calibration, useful for testing a claim about the direction of bias. Both meters share the same calibration in this model, while their random noise components are independent.

At the baseline $h = 2$ and $\kappa = 0$, the first meter's six signal/noise combinations are:

True use X, MWh	Noise hZ_1, MWh	Recorded use M_1, MWh	Mean bill at this true use, dollars
8	−2	6	430
8	0	8	430
8	2	10	430
12	−2	10	630
12	0	12	630
12	2	14	630

Each row has probability $1/6$ after averaging over the second meter and billing perturbation. The actual bill in a state is 20 dollars below or above the displayed mean. These rows enumerate possibilities; they are not six sampled observations.

True use has mean 10 MWh and variance 4 MWh^2. The baseline meter noise has variance $8/3$ MWh^2. A reading of 10 MWh can come from low use with a positive error or high use with a negative error. The record does not distinguish those backgrounds.

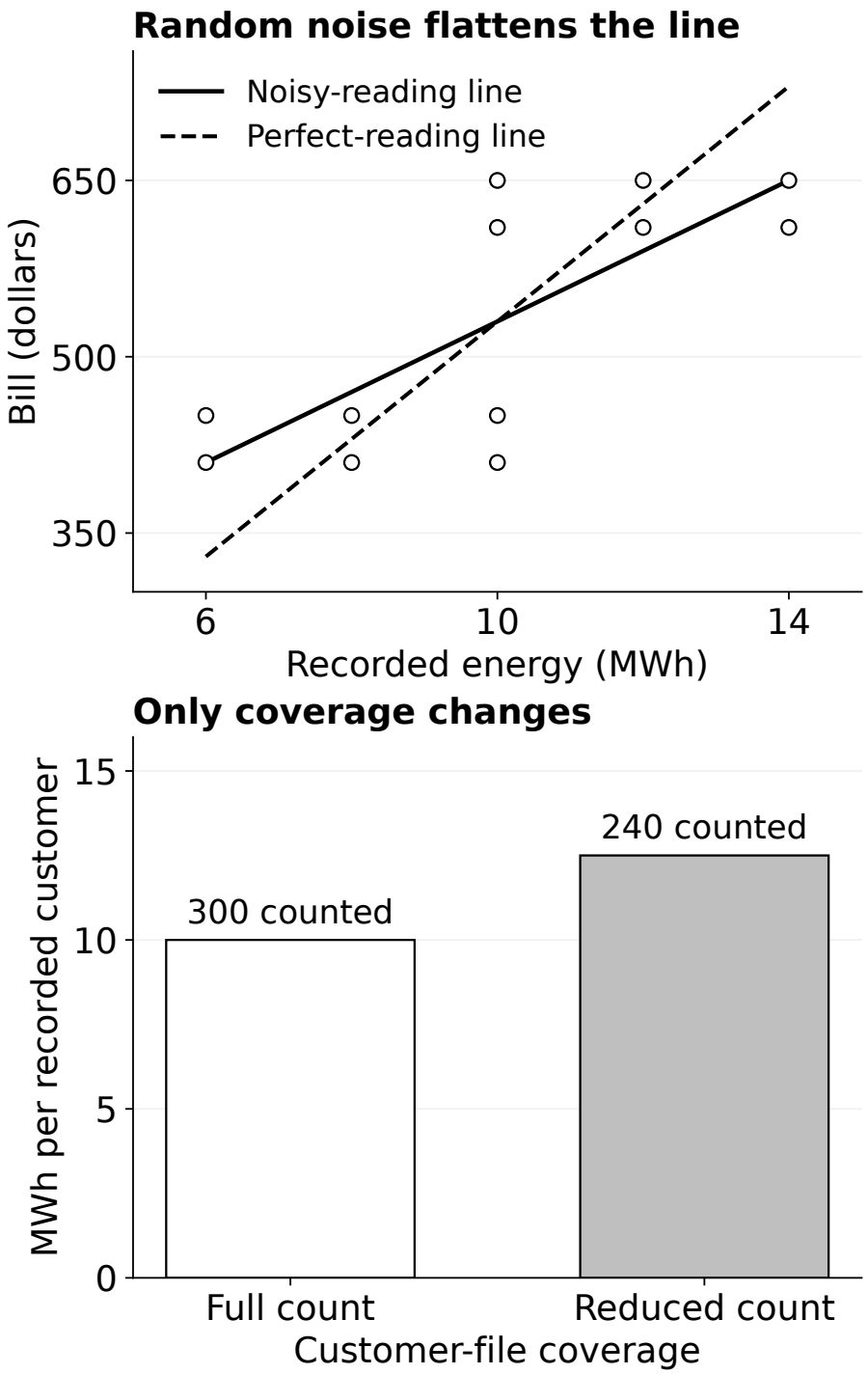


Figure 1: The first panel compares the perfect-reading relation with the population linear projection using noisy readings. Points represent the two possible bills for each of the six signal/noise combinations. The second panel holds energy and the true customer population fixed while the recorded denominator changes. Both panels describe controlled examples.

◇ Core: How a line summarizes co-movement

For a finite population with state probabilities p_i, the mean is the weighted sum $E[A] = \sum_i p_i A_i$. The probabilities sum to one. Chapter 1 develops this operation in its finite-average calculation. The **variance** of A and the **covariance** of A with B are

$$\begin{aligned} \mathsf{Var}(A) &= E[(A - E[A])^2], \\ \mathsf{Cov}(A, B) &= E[(A - E[A])(B - E[B])]. \end{aligned}$$

Variance averages squared distances from a mean. Covariance averages products of two centered distances. It is positive when above-mean values of one quantity tend to accompany above-mean values of the other. Covariance has the product of the two variables' units; it is not a unit-free correlation.

The **population linear projection** chooses an intercept a and slope b to minimize $E[(Y - a - bM)^2]$. It is the best line for that population and squared loss. When $\mathsf{Var}(M) > 0$, its coefficients are

$$b_M = \frac{\mathsf{Cov}(M, Y)}{\mathsf{Var}(M)}, \qquad a_M = E[Y] - b_M E[M].$$

Here is a direct proof without an optimization routine.

1. Define centered variables $M_c = M - E[M]$ and $Y_c = Y - E[Y]$. For a fixed slope b, write the error as $Y_c - bM_c + d$, where $d = E[Y] - a - bE[M]$. The centered part has mean zero. Its mean squared error is therefore $E[(Y_c - bM_c)^2] + d^2$. It is minimized by $d = 0$, giving $a = E[Y] - bE[M]$.
2. Expand the remaining square using the variance and covariance definitions:

$$E[(Y_c - bM_c)^2] = \mathsf{Var}(Y) - 2b\,\mathsf{Cov}(M, Y) + b^2\,\mathsf{Var}(M).$$

3. Complete the square in b. The same expression becomes

$$\mathsf{Var}(Y) - \frac{\mathsf{Cov}(M, Y)^2}{\mathsf{Var}(M)} + \mathsf{Var}(M)\left(b - \frac{\mathsf{Cov}(M, Y)}{\mathsf{Var}(M)}\right)^2.$$

   The first two terms do not depend on b. The last term is nonnegative and vanishes only at the stated slope because $\mathsf{Var}(M) > 0$. This proves the unique minimizing line and its risk.

If every reading is identical, the variance is zero. Dividing by it is undefined: a line's intercept and slope cannot be separated using a constant regressor. Only their fitted value at the observed reading matters.

This best-line result also holds for nonfinite populations with finite second moments. The algebra uses expectations of products, not the number of states. It does not assert that the true conditional mean is a line, that a finite-sample slope is unbiased, or that a causal effect has been identified.

◆ Theory: Why classical error flattens this slope

Consider the general scalar model $Y = \alpha + \beta X + \varepsilon$, observed through $M = X + W$. Assume:

- X, W, ε have finite second moments, and $\mathrm{Var}(X) > 0$.
- $E[W] = E[\varepsilon] = 0$.
- $\mathrm{Cov}(X, W) = \mathrm{Cov}(X, \varepsilon) = \mathrm{Cov}(W, \varepsilon) = 0$.

Independence of the three components is sufficient for the covariance restrictions, but the proof only needs those restrictions. These assumptions describe a particular additive error model. They are not established merely because a recorded variable looks noisy.

1. Expand the numerator in the projection slope. Covariance is linear in each argument, as follows by expanding its centered-product definition:

$$\begin{aligned}\mathrm{Cov}(M, Y) &= \mathrm{Cov}(X + W, \alpha + \beta X + \varepsilon) \\ &= \beta\,\mathrm{Var}(X) + \beta\,\mathrm{Cov}(W, X) + \mathrm{Cov}(X, \varepsilon) + \mathrm{Cov}(W, \varepsilon) \\ &= \beta\,\mathrm{Var}(X).\end{aligned}$$

   A constant has zero covariance. The three other terms vanish by the listed restrictions.

2. Expand the denominator. Using $M - E[M] = (X - E[X]) + W$ and taking the expected square gives

$$\mathrm{Var}(M) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, W) + \mathrm{Var}(W) = \mathrm{Var}(X) + \mathrm{Var}(W).$$

3. Divide these expressions and define the signal fraction λ:

$$b_M = \beta\lambda, \qquad \lambda = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \mathrm{Var}(W)}.$$

   Both variances are nonnegative and the signal variance is positive. Thus $0 < \lambda \leq 1$. Positive noise variance makes $\lambda < 1$, shrinking the slope's magnitude toward zero. For

a negative β, shrinkage makes the slope less negative; "downward bias" would be an imprecise description.

In the baseline meter world, $\mathrm{Var}(X) = 4$ and $\mathrm{Var}(W) = 8/3$. Therefore $\lambda = 3/5$, the recorded-energy slope is 30 dollars per MWh, and the intercept is $530 - 30(10) = 230$ dollars. The constructed billing slope remains 50 dollars per MWh. A larger collection of readings from this same population would not make its population projection equal 50.

The dimensions provide a useful check: the variance ratio is dimensionless, so multiplying by β preserves dollars per MWh. Zero noise gives $\lambda = 1$. Holding the signal variance fixed while noise variance increases moves λ toward zero. Converting both signal and noise to kWh multiplies both variances by a million and leaves λ unchanged.

This is the scalar classical errors-in-variables result. Hansen's January 10, 2010 manuscript, chapter 11, opening measurement-error example provides a related general regression treatment. The finite billing construction and proof here are authored for this chapter.

◇ Core: A calibration error can behave differently

The classical proof removes $\mathrm{Cov}(X, W)$ using an assumption. It cannot remove that term after a calibration error makes it nonzero.

In our meter family, $W = \kappa(X - 10) + hZ_1$. The independence and zero mean of Z_1 give $\mathrm{Cov}(X, W) = 4\kappa$. The recorded centered reading is $(1 + \kappa)(X - 10) + hZ_1$. Substituting into the centered-product definitions yields

$$\begin{aligned} \mathrm{Var}(M_1) &= 4(1+\kappa)^2 + 2h^2/3, \\ \mathrm{Cov}(M_1, Y) &= 200(1+\kappa), \\ b_{M_1} &= \frac{200(1+\kappa)}{4(1+\kappa)^2 + 2h^2/3}. \end{aligned}$$

The numerical constants carry the units inherited from the meter model. With no random noise and $\kappa = -1/2$, this slope is 100 dollars per recorded MWh. One recorded-unit change represents two true-unit changes; the coefficient doubles. With $\kappa = -2$ and no random noise, it is -50. The instrument reverses ordering. Neither result follows the classical attenuation conclusion because its covariance restriction fails.

Two instruments with the same compressed calibration can agree perfectly. Their agreement alone does not reveal their distance from true use. Calibration needs a relevant reference or design restriction, not just another copy of the same reporting process.

The interactive meter comparison lets you predict the slope, change noise or calibration, and then reveal the fitted line. Explain which term in the displayed covariance calculation changed. A correct prediction needs that reason; "measurement error causes bias" does not specify its direction.

◆ Theory: Recovering a coefficient and predicting a bill

If the aim is to predict a bill using the recorded meter, a slope of 30 is the correct slope for the baseline population's best straight-line predictor. Replacing it with 50 while retaining the noisy horizontal variable would generally worsen that line's forecast. A structural coefficient and a predictor solve different problems, as chapter 1's distinction between forecasting and intervention demonstrates.

The conditional mean can improve on the best line. For the baseline readings $M_1 = 6, 8, 10, 12, 14$, it is respectively $430, 430, 530, 630, 630$ dollars. At 10, the two true-use values are equally likely; the bill's perturbation has mean zero. At the other four readings, true use is determined in this finite example. Thus this conditional mean is not a line.

The projection risk calculation gives a line MSE of 4,400 dollars squared. The conditional-mean risk is $400 + 10{,}000/3 = 11{,}200/3$ dollars squared: billing noise contributes 400 everywhere; ambiguity between the two true-use means contributes 100^2 at the central reading, which occurs with probability $1/3$. The conditional mean's optimality follows from chapter 1's complete squared-loss proof.

These numbers depend on the finite support. Changing a noise amplitude can create or remove overlaps among recorded values. Consequently the best conditional-mean risk need not increase monotonically with the amplitude in an artificial discrete model. The classical best-line slope does shrink monotonically as its noise variance rises. Keep the claim attached to the object proved.

Neither a good forecast nor an attenuation correction proves that changing real energy use would change bills according to this equation. The equation is a declared generator here. An empirical claim would additionally need a defensible economic model and identification design.

◆ Theory: What independent repeated readings can identify

Suppose two readings of the **same** underlying quantity are $M_1 = X + W_1$ and $M_2 = X + W_2$. Retain finite second moments, positive signal variance, and the billing relation. Assume each measurement error is uncorrelated with X and ε, and $\mathrm{Cov}(W_1, W_2) = 0$. Assume also $\mathrm{Cov}(X, \varepsilon) = 0$. Under those restrictions,

$$\beta = \frac{\mathrm{Cov}(M_1, Y)}{\mathrm{Cov}(M_1, M_2)}.$$

1. Expand the readings' covariance:

$$\mathrm{Cov}(M_1, M_2) = \mathrm{Var}(X) + \mathrm{Cov}(X, W_2) + \mathrm{Cov}(W_1, X) + \mathrm{Cov}(W_1, W_2) = \mathrm{Var}(X).$$

The denominator is strictly positive by the signal-variance assumption.

2. Expand Cov(M_1, Y) exactly as in the attenuation numerator. The restrictions leave β Var(X). Dividing by the first expression cancels the positive signal variance, proving the result.
3. Interpret the cancellation. Independent noise expands each reading's own variance but does not expand the readings' shared covariance. The shared part isolates signal under the assumptions. A single proxy's variance does not by itself supply that separation.

At the classical baseline, the numerator is 200 dollar-MWh and the denominator is 4 MWh^2, giving 50 dollars per MWh. Copying the first meter into a second column instead gives denominator $20/3$ and ratio 30. It supplies no new measurement.

In our compressed-calibration world, the independently noisy meters still share the wrong scale. Their ratio is 100, because both violate the restrictions involving X. A second instrument must satisfy the proposed measurement model; the presence of two columns is insufficient evidence.

Two consecutive dates are also not automatically repeated readings of the same X. The true quantity may change, and instrument errors may persist over time. A time-series version would have to specify latent dynamics and error dependence. That leads toward the course's state-space definition of a latent state and observation equation, rather than silently treating adjacent observations as independent repeats.

Averaging genuinely independent equal-variance errors reduces their variance: Var$((W_1 + W_2)/2) = (\sigma_W^2 + \sigma_W^2)/4 = \sigma_W^2/2$. The classical baseline average-meter projection therefore has slope 37.5, closer to 50 but still attenuated. The covariance ratio and an averaged regressor perform different operations. A common error would survive the average.

◇ Core: A denominator can manufacture growth

Consider a separate controlled reporting example. A district uses 3,000 MWh during each of two days and has 300 actual customers on both days. Its true daily energy per customer is 10 MWh. The numerator covers all customers on both days. The customer file records all 300 on the first day but only 240 on the second.

Dividing the unchanged full-coverage numerator by these recorded counts produces 10 and 12.5 MWh per recorded customer: apparent growth of 25 percent. Nobody's consumption changed. The numerator and second-day denominator describe different coverage. These invented records illustrate an audit failure; they do not allege a problem with a named data provider.

For positive totals Q_t and denominators N_t, define $R_t = Q_t/N_t$. Its exact proportional growth is

$$1 + g_R = \frac{R_t}{R_{t-1}} = \frac{Q_t/Q_{t-1}}{N_t/N_{t-1}} = \frac{1 + g_Q}{1 + g_N}.$$

The first equality defines growth, the second substitutes the ratio twice, and the third substitutes numerator and denominator growth. In this example $g_Q = 0$ and $g_N = -0.2$, so $g_R = 1/0.8 - 1 = 0.25$. Subtracting percentage changes would give 20 percent, an approximation that is inaccurate here. For positive values the log identity $\Delta \log R = \Delta \log Q - \Delta \log N$ is exact, because a log of a ratio is the difference of logs. Neither identity tells us whether a changed denominator represents an actual population change or a reporting change. The measurement map must supply that interpretation.

◆ Theory: Aggregation changes the weights

Suppose disjoint zones have totals Q_j and positive counts N_j, all with aligned coverage and intervals. Their aggregate per-customer measure is

$$\frac{\sum_j Q_j}{\sum_j N_j} = \sum_j \frac{N_j}{\sum_k N_k} \frac{Q_j}{N_j}.$$

To verify the equality, multiply each zone's ratio by its count weight. The N_j cancels, leaving $Q_j / \sum_k N_k$. Summing gives the left side. The weights are nonnegative and sum to one, so the aggregate lies between the smallest and largest zone ratios.

For two zones with 1,000 MWh and 100 customers, and 3,000 MWh and 200 customers, the ratios are 10 and 15. The aggregate is $4{,}000/300 = 40/3$, about 13.33 MWh per customer. Giving the two zones equal weight instead gives 12.5. The first describes the total divided by total customers; the second describes an equally weighted average zone. Both calculations are defined, but they answer different questions.

This is a bridge to geospatial econometrics: changing boundaries changes which individuals share an aggregate and how groups are weighted. Matching units on both axes does not make differently supported aggregates equivalent. In a network, removing nodes likewise changes the domain over which degree or an average is computed. A later geographic or network analysis must preserve the population and boundary definitions needed for its claim.

✦ Workshop: Worked exercises and an audit

These exercises have complete public answers. The controlled calculations test definitions and identities; they are not additional completed empirical cases.

1. **Convert the baseline slope.** What happens when both meter readings change from MWh to kWh? The slope becomes $30/1{,}000 = 0.03$ dollars per kWh, and the intercept remains 230 dollars. Predicted bills stay unchanged because $0.03(1{,}000M) = 30M$. Reliability is unchanged because both relevant variances acquire the same million-fold factor. The native metric `h2_classical_slope_per_kwh` checks this operation.
2. **Make random noise larger.** Set $h = 4$ and $\kappa = 0$. Noise variance is $2(4^2)/3 = 32/3$, so $\lambda = 4/(4 + 32/3) = 3/11$. The slope is $150/11$, about 13.64 dollars per MWh. This uses the classical covariance restrictions, not a general claim about all imperfect proxies.
3. **Compress the instrument without random noise.** Set $h = 0$ and $\kappa = -1/2$. The calibration formula gives 100. The meter repeats without random noise but uses the wrong scale. Calling that repetition "accurate" confuses reliability with validity.
4. **Compare a second reading with a copied column.** At the classical baseline the independently noisy twin gives ratio 50; a copied reading gives 30. Follow the repeated-reading denominator proof and identify which covariance differs. Independence is a design claim that needs evidence in real work, not a property created by a new column name.
5. **Audit reported growth.** A report says energy per customer rose from 10 to 12.5 while total energy stayed at 3,000. Recover the implied denominators: $3{,}000/10 = 300$ and $3{,}000/12.5 = 240$. Check coverage before interpreting the rise as behavior. The fixed true customer population in the declared example identifies this as a reporting change.
6. **Choose the aggregate.** For the two zones in the aggregation proof, explain why 12.5 and $40/3$ differ. Write the weights explicitly: $1/2, 1/2$ versus $1/3, 2/3$. State whether the client wants an average zone or a total-per-customer quantity before selecting the result.

An AI assistant can help audit a measurement dictionary by listing ambiguities or proposing a deliberately broken example. Give it the actual schema, units, coverage, and requested claim. Require it to identify which equation and assumption justify a correction. Then independently check its answer against the definitions, fixture, and calculations. A fluent explanation is not validation evidence. Record the prompt and correction in the project log; omit invented source documentation and private records from a public portfolio.

Professional handoff: a measurement dictionary and memo

The companion measurement dictionary records the fields used in this controlled example. The fixture manifest defines all 36 states, units, probabilities, equations, and authorship. The execution guide gives the native Python, Julia, and R commands; the measurement verification report records what actually ran. A program's output is a calculation check, while the written proofs justify the population identities.

A professional memo can now say: "The controlled baseline's regression on recorded energy has slope 30 dollars per MWh, although the declared billing relation uses 50. Independent additive measurement noise explains the difference under the stated covariance restrictions. Independent repeated readings recover the declared coefficient in that classical model; copied readings and shared calibration errors do not. A separate reporting example creates a 25-percent ratio increase through denominator coverage alone."

For an empirical portfolio project, replace the controlled records with a permitted source and document its measurement operations. Deliver the dictionary, a unit/coverage audit, one reproducible comparison of competing proxies, and a memo stating which interpretation is supported. Preserve uncertain definitions as unresolved questions rather than choosing a convenient meaning. The time-series project's provenance requirements give an existing example of documented source fields and limitations. Chapter 3's three-clock definition distinguishes reference, publication and collection times; its as-of selection preserves the information available at a decision origin.

This chapter's main equation checks are explicit: dollar terms add to dollars; signal fractions are unit-free; zero noise recovers the correct classical slope; a constant regressor has no separately identified slope; kWh conversion rescales coefficients without changing bills; independent repetition removes only the noise its assumptions cover; positive denominators are required for the ratio and log calculations. These checks support the proofs but do not replace them.

Instrument comparison

The online experiment asks for a prediction before revealing a population line. These static comparisons retain all nine parameter settings for this edition. The billing coefficient is 50 dollars per true MWh throughout. A calibration error changes the interpretation of both recorded readings.

Noise h, MWh	Calibration	Recorded slope, dollars/MWh	Twin covariance ratio, dollars/MWh
0	classical	50	50
0	compressed	100	100
0	reversed	-50	-50
2	classical	30	50
2	compressed	27.273	100
2	reversed	-30	-50
4	classical	13.636	50
4	compressed	8.5714	100
4	reversed	-13.636	-50

The twin ratio recovers the declared coefficient only under the repeated-reading assumptions proved in this chapter. The archive includes each native implementation and its checked outputs.

Observation clocks, releases, vintages, and provenance

○ Orientation: A quarter ends before its estimate arrives

An analyst preparing a report on April 24, 2024 can know that the first quarter has ended. That does not give the analyst the Bureau of Economic Analysis's first estimate of that quarter's growth. The estimate has its own publication date. A file downloaded much later can put the number beside January 1, the quarter's starting date, without showing when it became available.

This is how a tidy spreadsheet can give a historical analyst knowledge of the future. The problem is in what the row represents, before any forecasting method is chosen. Begin with chapter 2's measurement map and add three clocks:

- **Reference time** identifies the interval the statistic describes. Here that interval is January through March 2024.
- **Publication time** identifies when a particular estimate was released by its provider.
- **Collection time** identifies when a particular system received and could use it. Publication and collection need not coincide.

A fourth timestamp, the date of our present-day research download, documents how we obtained the archived source. It cannot establish that our system possessed the file in 2024. These distinctions also apply to financial filings, revised geospatial boundaries, delayed survey records, and corrected network edges. The definition of the economic object and the history of access are separate pieces of evidence.

For the same quarter, the archived BEA headlines give this sequence:

Estimate and original source	Published, UTC	Real GDP growth, percent
Advance	April 25, 12:30	1.6
Second	May 30, 12:30	1.3
Third	June 27, 12:30	1.4

These are seasonally adjusted annualized percentage changes from the preceding quarter. Each original header specifies 8:30 a.m. Eastern Daylight Time; that is 12:30 UTC on these dates. All three numbers describe the same quarter. Their differences are revisions to an estimate, not

growth across three successive quarters. “Third” is the third estimate in this extract, not a claim that no subsequent revision occurred. These archived releases have been superseded.

At noon UTC on May 15, the advance estimate was already published and the second was still ahead. In our idealized zero-delay reconstruction, the available headline is 1.6. Using 1.3 or 1.4 as that origin’s input would use a later publication. The three-row ledger and source receipts preserve the observed values, dates, identifiers, source links and checked document hashes. They do not constitute a fitted GDP nowcast or a study of predictive performance.

◇ Core: A growth rate also has a convention

The percent sign does not tell us the interval or transformation. If a quarter’s growth factor is $1 + q$, an annualized growth rate repeats that factor four times:

$$1 + g = (1 + q)^4.$$

Here g and q are dimensionless proportions. The tabulated 1.6 percent means $g = 0.016$. Taking the positive fourth root gives

$$q = (1 + g)^{1/4} - 1.$$

Multiplying this proportion by 100 gives about 0.398 percent for the quarter. Dividing the displayed annualized percentage by four is an approximation; it does not reproduce compounding exactly. The positive-factor condition $1 + g > 0$ is required for this real-valued interpretation. The archived advance release’s Statistical conventions, PDF page 5 describes its annual-rate convention. Our calculation converts the reported rounded headline; it does not reconstruct BEA’s unrounded source values.

The second-minus-advance revision is $1.3 - 1.6 = -0.3$ percentage points. The third-minus-second revision is $+0.1$ percentage point, and the net third-minus-advance revision is -0.2 percentage point. A percentage-point difference is not a proportional percentage change. Preserve the convention and rounding in the provenance manifest, alongside the unit and coverage fields introduced in chapter 2’s measurement definition.

◇ Core: Selecting a value as of a decision time

Keep each publication as a separate event. For a series and reference period, an event r has a provider publication time p_r, an availability time a_r, a value v_r, and an identifier. Availability cannot precede the permitted publication time. If a system receives the event later, the receipt time determines when it can use the event.

In this chapter's reconstruction, we specify a constant collection delay $\delta \geq 0$, in seconds, and set

$$a_r = p_r + \delta.$$

Adding seconds to a timestamp produces another timestamp. We test $\delta = 0, 60, 3600$. These are declared scenarios; they are not measured BEA delivery latencies. In production, preserve actual receipt times where they exist. A provider's release header supplies the publication clock but not a particular customer's collection clock.

At an origin τ, the eligible event set is

$$\mathcal{A}_\tau = \{r : a_r \leq \tau\}.$$

Equality is eligible in this declared convention. This assumes the value is usable at the stated availability instant. A system that requires parsing, quality checks or execution time should use the later usable timestamp. For an intraday decision, knowing only a publication date is insufficient evidence of exact availability within that day.

An **as-of selection** returns, for each series/reference-period key, the newest provider publication among its eligible events. If none exists, that period is missing. It is not assigned a zero or filled with a later value. Separate series, geographic identifiers and measurement definitions belong in the key; matching a date alone is insufficient.

The implementation follows a definite order:

1. Validate the schema, finite numerical values, UTC timestamps and unique reference-period/publication keys. An ambiguous duplicate is an error, not a reason to take an arbitrary row.
2. Remove events whose availability is after the origin. A reference period ending before the origin does not make its publication eligible.
3. Within each retained key, select the greatest provider publication time. Preserve the selected event's identifier and both clocks with its value.
4. Join that origin's selected records to the declared forecast input. Retain absent periods as absent and state how the method handles them.

Filtering must precede selecting the last version. Selecting today's latest value and then attaching an old reference date loses the information needed to reconstruct the old decision. If receipts arrive out of order, the latest receipt need not be the newest provider estimate; the two clocks still perform different jobs.

On April 25 at 12:30:00 UTC, zero-delay availability includes the advance estimate. Under a 60-second delay, it is still absent at that instant and becomes eligible at 12:31:00. On May 30 at 12:30:00, the zero-delay selection switches to the second estimate, while the delayed system still selects the advance estimate. A one-second boundary test can expose a mistake hidden by a month-level join.

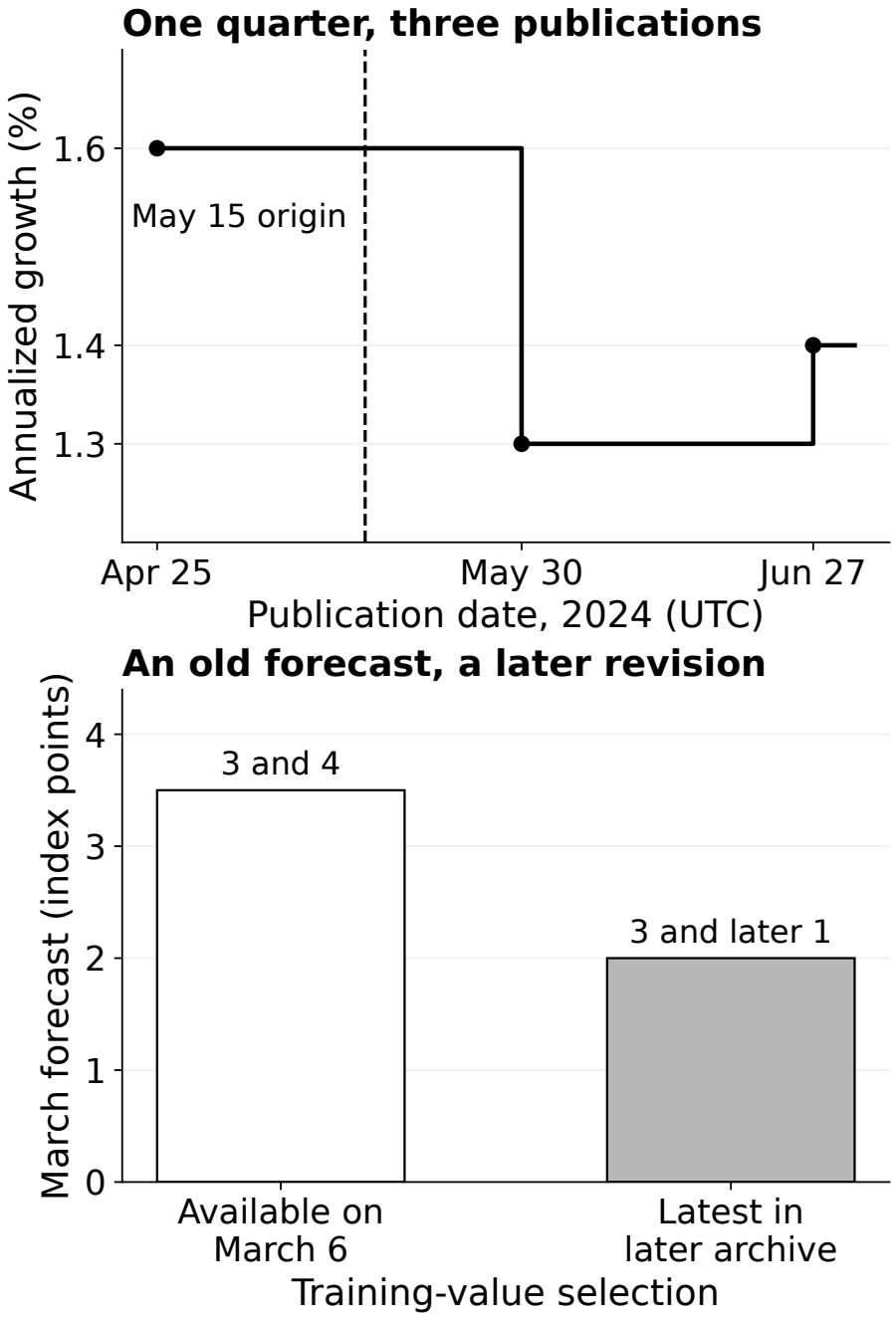


Figure 1: The first panel uses the attributed BEA extract with an idealized zero collection delay. Its horizontal axis is publication time, not the quarter being measured. The second panel uses the separate authored index example developed below. It illustrates a pipeline failure and does not estimate a GDP forecasting effect.

◆ Theory: A future release cannot change an earlier calculation

The archive can contain events from many dates. An earlier calculation must depend on the relevant history, not on everything the archive eventually contains. Write $\mathcal{H}_\tau(D)$ for the ordered history of all events available by τ in a valid ledger D. The history retains their values, identifiers and relevant metadata. An origin's forecast is

$$\hat{y}_\tau = T(\mathcal{H}_\tau(D), c_\tau),$$

where T is a deterministic calculation and c_τ is the configuration allowed at the origin. The configuration includes the model specification, transformations, fitted parameters, tuning decisions and, when applicable, a fixed random seed. Calling parameters "configuration" does not make later-fitted parameters historically available; their provenance must be established too.

Prefix invariance. Suppose two valid ledgers D and D' have identical histories through τ, and use the same permitted configuration. Then their forecasts at τ are identical.

1. Every available event in one history has the same key, clock, value and relevant metadata in the other. Any added or altered event is still strictly after the origin. Thus $\mathcal{H}_\tau(D) = \mathcal{H}_\tau(D')$.
2. Applying the as-of selection to either history gives the same eligible records and the same newest eligible event for each key. Missing keys remain missing in both.
3. The deterministic transformation, join and forecasting calculation receive identical inputs and the same configuration. A function applied to identical arguments has identical output. Therefore

$$T(\mathcal{H}_\tau(D), c_\tau) = T(\mathcal{H}_\tau(D'), c_\tau).$$

The statement requires changes to remain outside the origin's history. Backdating an altered release into the eligible set changes the premise. Changing the historical model or a random seed also changes an argument of T. For a randomized procedure, holding the origin-permitted random input fixed gives this pathwise equality; with a different random input, identical numerical outputs are not required.

This proof gives an adversarial diagnostic. Alter future values drastically or append a future release, rerun the early origin and compare its selected identifiers, transformations and forecast. The early result should be unchanged. Passing several mutations checks those paths; it does not prove that every possible leakage path has been removed. A model trained on later data could ignore the mutated field and still pass. The proof describes the required dependency; the audit must inspect the entire dependency chain.

The interactive release audit changes a future value while preserving the origin's eligible history. Predict which selected value can change before revealing the comparison. The native programs check the same operation at nine origins and three collection delays, including missing periods and exact release boundaries.

◇ Core: A forecast can leak without seeing its target

Consider an authored monthly index. These six invented records are independent of the BEA extract:

Reference month	First publication and value	Later publication and value
January	February 5, 09:00 UTC: 2	February 15, 09:00 UTC: 3
February	March 5, 09:00 UTC: 4	March 15, 09:00 UTC: 1
March	April 5, 09:00 UTC: 5	April 15, 09:00 UTC: 7

The numbers are index points. On March 6 at 09:00 UTC, suppose a transparent baseline predicts March using the mean of the available January and February records. January's revision to 3 is already available; February's first reading is 4. The locked forecast is

$$\hat{y}_{\text{Mar6}} = (3 + 4)/2 = 3.5.$$

If we later collapse each month to its last version before evaluating that old origin, January is 3 and February is 1. The reconstructed forecast becomes $(3+1)/2 = 2$. It never used the March target. It nevertheless used a February revision that was unavailable on March 6. Training-data revisions alone can create leakage.

Before February's first publication, on March 4, this baseline has only January's value and returns 3. That is a different input set, explicitly allowed by this example's available-mean rule. A model requiring a balanced two-month input should instead decline to forecast then. Missing information is part of the contract, not permission to choose whichever backfill improves a score.

Increase every release after March 6 by 1,000 index points while preserving its publication date. The locked forecast stays 3.5. The incorrect latest-version forecast becomes $(3 + 1001)/2 = 502$. The enormous change makes the dependency visible. It is a deliberately corrupted test archive, not a claim about actual index revisions. The controlled ledger and verified outputs retain both calculations.

◆ Theory: Information groups possible worlds

The deterministic prefix argument tells us how a pipeline must behave. Probability gives another way to describe what a forecaster can distinguish. Begin with four equally likely possibilities, collected in a set $\Omega = \{1, 2, 3, 4\}$. Their future outcomes are $Y = (0, 2, 4, 6)$ in abstract outcome units.

At time 0, no reading distinguishes them. At time 1, a signal says whether the state is in $C_0 = \{1, 2\}$ or $C_1 = \{3, 4\}$. At time 2, the exact state is revealed. A **partition** groups states that

have the same observable history. Its groups are called cells. A forecast using only that history must give the same answer to every state within a cell: the forecaster has no observed basis for selecting one state over another.

The events distinguishable at time 1 are $\emptyset$, C_0, C_1 and Ω. This collection is an **information sigma-algebra**, written $\mathcal{F}_1$. In a finite example it contains exactly the unions of partition cells. The name means that complements and countable unions remain in the collection. Complementing a union of cells selects the remaining cells; taking any union of such unions still selects cells. The collection also contains the whole set and empty set.

Here $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and $\mathcal{F}_2$ contains every subset of the four states. Consequently

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2.$$

A nested family of information sigma-algebras is a **filtration**. It represents accumulation of distinguishable information. Retaining the whole publication ledger allows later information to include the earlier estimate and its revision. A storage system that overwrites the old estimate without retaining it can discard history; its latest-value table alone should not be confused with the full filtration.

A forecast process is **adapted** to this filtration when its time-t forecast is measurable with respect to $\mathcal{F}_t$: events described by its output can be decided using the information available then. In this finite setting, measurability is equivalent to being constant on every information cell.

1. If a forecast is constant on each cell, the states where its value belongs to any specified set of numbers form a union of cells. That union is an event in the information sigma-algebra, so the forecast is measurable.
2. If it assigns different values to two states within one cell, take a numerical threshold strictly between those values. The event that the forecast exceeds that threshold splits the cell. Such an event is not a union of information cells and cannot belong to the sigma-algebra. The forecast is not measurable.

The time-1 forecast $(1, 1, 5, 5)$ is adapted. The apparent forecast $(0, 2, 4, 6)$ splits both cells and uses information unavailable at time 1. Its perfect score would reveal an impossible information advantage, not a successful signal-based forecast. At time 2, after exact revelation, those same four outputs are permitted. A calculation's validity depends on when its inputs are known.

This finite proof is exact and requires no density or asymptotic approximation. General probability spaces need the full measurable-space definition rather than four-state enumeration; those foundations will receive their own treatment. The finite model here establishes the meaning needed for the release audit.

◆ Theory: More information improves the population optimum

The squared-loss proof in chapter 1 applies within each positive-probability information cell. The minimizing forecast is its weighted outcome mean. With no information the forecast is 3 and its risk is

$$R_0 = (9 + 1 + 1 + 9)/4 = 5.$$

With the coarse signal, the two cell means are 1 and 5. The four squared errors are all 1, giving $R_1 = 1$. With the exact state, $R_2 = 0$. The means do not claim that intermediate outcomes occur; they minimize average squared loss over the possibilities.

The improvement has a precise proof. Let m_0 and m_1 denote optimal conditional means for nested finite information partitions, with positive probability for each retained cell. On every finer cell, $E[Y - m_1 \mid \text{cell}] = 0$. The older forecast m_0 is constant there because the finer cell lies inside an older one.

1. Decompose the older error:

$$Y - m_0 = (Y - m_1) + (m_1 - m_0).$$

2. Square and average. On a finer cell the second term is constant, so its cross-product with $Y - m_1$ has conditional mean zero. Summing over the cell probabilities keeps that cross-product zero. Thus

$$E[(Y - m_0)^2] = E[(Y - m_1)^2] + E[(m_1 - m_0)^2].$$

3. The last term is a mean of squares and is nonnegative. Therefore the finer-information optimum cannot have larger squared-error risk. Equality holds precisely when the two conditional means agree on every positive-probability cell.

In our four states the means move from 3 to 1 or 5, each two units away. The gain is $E[(m_1 - 3)^2] = 4$, matching $R_0 - R_1 = 5 - 1$. Risks have squared outcome units; the forecast shifts have outcome units before being squared. This is a population result for optimal forecasts. It does not promise that a fitted model improves whenever someone adds a feature. Estimation error, a misspecified function and changing distributions can defeat that empirical promise.

◇ Core: The target has a publication history too

What will count as a forecast error? In the controlled monthly example, March's first published value is 5 and its later value is 7. The unchanged locked forecast of 3.5 has squared error 2.25 against the first and 12.25 against the later. Scoring the same forecast against different publications changes the evaluation target.

Write $e_v = y^{(v)} - \widehat{y}$ for error against target vintage v, and $d = y^{(w)} - y^{(v)}$ for the revision to vintage w. Then $e_w = e_v + d$. Expanding the square gives the exact identity

$$e_w^2 - e_v^2 = 2e_v d + d^2.$$

In the example, $e_v = 1.5$ and $d = 2$, so the change is $6 + 4 = 10$ squared index points. A revision can increase or decrease a score: the cross term can be negative. A later estimate need not move away from a forecast. Declare whether the task forecasts the first release, a specified later publication, or a target defined by an explicit revision horizon. "Final GDP" without a dated definition is insufficient.

Three evaluation designs answer different questions:

- **A fixed-vintage historical exercise** uses one frozen extract for every origin and withholds later reference periods. It documents that extract's information advantage when older values were revised later.
- **A reconstructed real-time exercise** rebuilds each origin's inputs from documented release histories and explicit availability assumptions. This chapter's BEA audit is such a reconstruction with declared collection-delay scenarios.
- **A prospective production exercise** logs actual inputs, receipt times, model/configuration identities and locked predictions as decisions occur, then scores them against the predeclared target publication.

Keeping the target out of training is necessary in each design; it does not convert the first into the third. The existing Texas case's fixed-vintage disclosure is retained rather than relabelled as an intraday historical experiment.

✦ Workshop: A release audit and worked answers

The following exercises have public worked answers. The BEA part is a small observed release-history audit. The larger C02 nowcast, latent-state inference and revision-attribution case remains to be developed; three headlines cannot establish a forecasting advantage.

1. **Choose the eligible headline.** At May 15, 12:00 UTC with zero delay, the eligible estimate is advance, 1.6 percent. At May 30, 12:30 UTC with a 60-second delay, it is still advance; the second becomes eligible a minute later. Use the availability inequality, not the quarter's starting date.

2. **Distinguish absence from zero.** At March 31, no headline in the three-row extract is published. Report "not released in this extract." Assigning 0 would assert a growth estimate the provider did not supply. A forecast could still use other permitted evidence, but this ledger contains none of that evidence.
3. **Convert the annualized headline.** Substitute $g = 0.016$ in the compounding formula. The resulting quarter growth is about 0.398 percent. Explain why 0.4 is an approximation and retain the source headline's rounding.
4. **Find the invisible revision.** The incorrect March 6 forecast uses February's later value of 1. Replace it with the eligible first value of 4 and recover 3.5. The March target never entered either training calculation; the error arose from the training-vintage choice.
5. **Explain a perfect score.** At signal time the two information cells each contain two distinct outcomes. An output equal to the exact outcome separates states inside both cells, violating the finite measurability criterion. Perfect scoring against a withheld outcome is not evidence that the permitted signal supplied it.
6. **Reconcile two scorecards.** Keep the forecast fixed at 3.5. Its errors against targets 5 and 7 are 1.5 and 3.5. Square them, then check that their difference equals $2(1.5)(2) + 2^2$. Name the target-vintage identity in the memo instead of calling the entire score change a model failure.

Ask an AI assistant to inspect the schema and propose three ways a future value could enter an earlier origin. Require it to name the affected field, clock and operation, and to produce a valid mutation that stays after the origin. Verify those proposals yourself. Reject invented publication times, claims that shifting a reference date proves availability, and changes to an origin's model that silently invalidate the prefix-invariance premise. Record the accepted test, correction and executed result in the portfolio log.

Professional handoff: A reproducible information ledger

The manifest defines the extract's scope, units, timestamps, keys, delay scenarios, hashes and separate rights. The execution guide supplies commands for independent Python, Julia and R implementations. Their native verification report records matched calculations and boundary/failure checks. These records establish what we actually tested; the written arguments establish why the operations have their claimed properties.

Deliver a preserved extract, its hash, the manifest, selected-event identifiers for each origin, a frozen forecasting specification, declared target vintages, and a short audit memo. A hash can show that two copies contain the same bytes. It does not establish that the file existed at an earlier date or that the contents were correct; the source and receipt evidence perform those tasks. Preserve the original clock fields through unit conversions, joins and aggregation.

ALFRED's download documentation, Observations by Real-Time Period describes observation dates and the intervals during which versions were current. Its date-level vintages are useful for

archived macroeconomic work, but they should not be treated as exact intraday receipt timestamps. A release with no numerical revision may also need a separate event receipt; a list of changed-value vintages is not automatically a complete publication calendar.

A memo for this bounded example can say: "The archived 2024 Q1 GDP headlines were 1.6, 1.3 and 1.4 percent at three documented publication dates. Our reconstructed selection respects the declared collection delay and inclusive availability rule. Future-only mutations and appended releases leave earlier selections unchanged. A separate controlled index experiment shows how a later training revision can alter an old forecast even when the target is withheld. We have not estimated a GDP nowcast's predictive value."

The same logic links economic forecasting to event logs and delayed observations in physical systems: the event described, the message sent and the message received can occur at different times. This shared ordering problem does not imply that statistical revisions behave like physical noise or that a publication lag identifies an economic effect. Geographic boundary revisions and changing network membership likewise need versioned definitions. Later applications can build on this ledger without pretending that differently dated or differently supported records describe the same information set.

Release comparison

The online experiment asks for a prediction before showing an origin-specific selection. These static comparisons use zero collection delay and add 10 percentage points only to events unavailable at each origin. The mutations are authored audit tests, not BEA estimates.

Origin date, UTC	Time, UTC	Eligible stage	Eligible growth, percent	Entire test archive, percent
2024-03-31	23:59:59	Not released	Missing	11.4
2024-04-25	12:29:59	Not released	Missing	11.4
2024-04-25	12:30:00	Advance	1.6	11.4
2024-04-25	12:31:00	Advance	1.6	11.4
2024-05-15	12:00:00	Advance	1.6	11.4
2024-05-30	12:30:00	Second	1.3	11.4
2024-06-27	12:29:59	Second	1.3	11.4
2024-06-27	12:30:00	Third	1.4	1.4
2024-07-01	12:00:00	Third	1.4	1.4

At the first-publication boundary, a 60-second delay retains missingness until 12:31 UTC; a 3,600-second delay retains it until 13:30 UTC. At subsequent boundaries, the prior estimate remains selected during the delay. The native programs check all 27 origin/delay combinations and keep absence distinct from zero.