STA 542: Introduction to Time Series Analysis

Lecture 1

Lasse Vuursteen

Duke University

Mon Aug 24, 2026

One workbook organizes each topic class

Before (and/or after) class
read the designated workbook; run relevant R chunks

→

In class
develop the central argument and selected examples

→

After class
review; attempt assigned and optional exercises

Workbooks and slides are the primary text.
Plan three to four hours outside class per workbook, including homework.

Canvas holds the current files

- Canvas is the source of record for materials and deadlines.
- Workbooks are Quarto `.qmd` files. Set up R, RStudio, and Quarto from the official Get Started guide.
- A local LLM may diagnose the smallest relevant error; verify the fix, then ask the TAs if the problem remains.
- Public overview and syllabus: `lassev.github.io/teaching/sta542/`.

Practice can raise the course grade; it cannot lower it

Let F, M, H, and Q be the percentage scores for the final, midterm, homework, and pop quizzes. Relative weights, when included, are

$$F : M : H : Q = 5 : 3 : 2 : 1.$$

- The most favorable scheme is applied automatically.
- The course percentage is therefore never below the final percentage.

The recorded percentage is the maximum of four schemes

$$\max\left\{F,\ \frac{5F+3M}{8},\ \frac{5F+3M+2H+Q}{11},\ \frac{5F+2H+Q}{8}\right\}$$

- Final only
- Final and midterm
- All four components
- Final, homework, and quizzes

Homework rewards serious attempts

- Nine short sheets; the best seven count; about two to three hours each.
- Two points: substantive attempt, partial attempt, or no meaningful submission.
- Handwritten hard copy at the start of Wednesday's class; tablet work must be printed.
- None in the first week; first deadline September 9. None due during the midterm and fall-break period.

Many examination questions are adapted from, or closely related to, homework problems.

A named contradiction can earn both homework points

Example.

> *I asked the LLM to expand* $\mathrm{Var}(\bar{X}_T)$ *for an AR(1) with* $\phi = 0.9$*. It returned* σ^2/T *and said the cross terms vanish. The workbook keeps* $\gamma(h)$ *for* $h \neq 0$*. Doesn't that contradict the vanishing? I do not see which assumption set those covariances to zero.*

The attempt itself must be handwritten, also the LLM output.

Pop quizzes reward recall

- Eight unannounced quizzes; the best six count.
- About five minutes, on material from previous meetings.
- Graded for correctness; no makeup for an ordinary absence.

AI and collaboration may support homework learning

Workbooks and homework

- Discussion is encouraged. AI tools may be used as tutors, critics, and coding assistants.
- Each submission is independently handwritten in the student's own words; collaborators are listed when applicable.
- A failed approach, with a precise account of where it became stuck, can earn full effort credit.

Pop quizzes and examinations

- Independent work; no collaboration or AI tools.

TA hours are the first stop for workbook questions

Arunsoumya	Tue, 10:30–11:30 a.m.
Semu	Thu, 2:00–3:00 p.m.
Lasse	Wed, 3:00–4:00 p.m.

- TA hours handle workbook and homework questions, including extensions.
- Locations are posted on Canvas. Additional meetings may be arranged by email.

Two in-class examinations

Date	Course calendar
Sep 7	Labor Day; no class
Oct 7	Midterm examination in class
Oct 12	Fall break; no class
Nov 23	Last class
Dec 9	Final examination, 9:00 a.m.–12:00 p.m.

Each exam allows one notes sheet and a calculator

- Midterm: 75 minutes, Workbooks 1–10.
- Final: cumulative, emphasizing Workbooks 11–21.
- One student-prepared, double-sided letter-size notes sheet and a nonprogrammable calculator.
- No live programming; no AI tools.

Dependence changes what one path can reveal

Time series data arrive in order along one observed path. The course asks what that path can reveal and how it can support prediction.

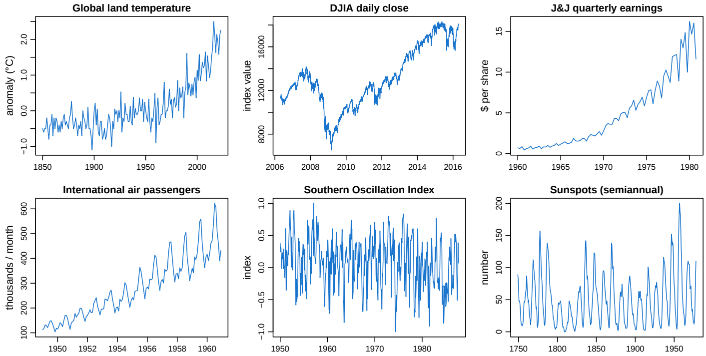
Specify a process coherent joint laws → **Learn from one path** moments and fitted models → **Forecast** points and distributions → **Criticize and adapt** vectors, states, and breaks

50% of the course is mathematics; 50% is simulations / real data in R

- Mathematical arguments identify the target and the conditions under which a method can work.
- Simulation and real-data analysis in R show how it behaves.
- Frequentist and Bayesian reasoning enter where they clarify the inferential question.

Data with a ‘temporal’ component

Each observation carries a time stamp, and the record is read in time order.

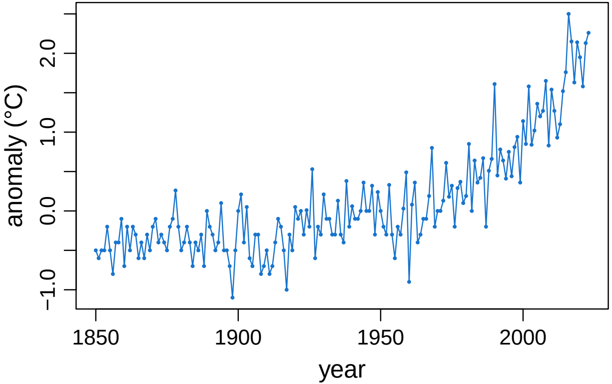


Real data — shown on a white background throughout; simulated data will appear on gray. Sources: R package `astsa` (`gtemp_land`, `djia`, `jj`, `soi`, `sunspotz`) and R `datasets` (`AirPassengers`).

One number per time moment

t (year)	x_t (°C)
1850	-0.50
1851	-0.60
1852	-0.50
1853	-0.50
1854	-0.20
$\vdots$	$\vdots$
2023	2.26

Real data: `astsa::gtemp_land`.



The raw object is an array: one number x_t per time moment t. Plotting x_t against t gives the standard picture of a time series.

The datum need not be a number

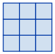
$x_t \in \mathbb{R}$

a number:
temperature, price

$x_t \in \mathbb{R}^d$

a vector:
several series jointly

$x_t \in \mathbb{R}^{m \times m}$



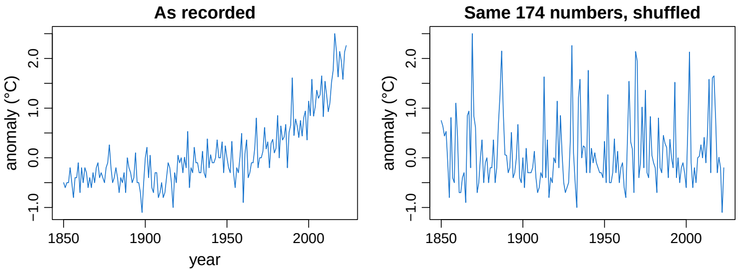
an image:
satellite frame, scan

x_t a curve

a function:
heartbeat,
EEG epoch

- One object per time moment; the time structure — order, past, future — is the same in every case.

The order carries the information



- Same numbers, same histogram — plainly not the same dataset.
- Under an i.i.d. (or any exchangeable) model, both panels are equally probable datasets.
- The trend and the clustering of warm years live in the order: past values influence future values.

Real data: `astsa::gtemp_land`; right panel deliberately shuffled.

Ordered data reach beyond time

“the words of a sensible paragraph are certainly not i.i.d.”

“paragraph the certainly words not sensible of are i.i.d. a”

- The same shuffle test, applied to text: the marginal — the word frequencies — survives; the meaning does not.
- Time-series ideas apply to any ordered data: text, DNA, event logs, queues.
- A language model is a conditional distribution for the next word given the words so far.

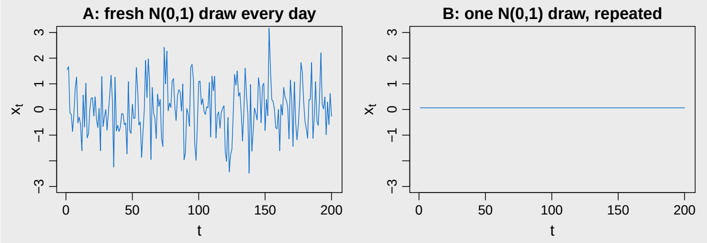
Modeling i.i.d. data takes one distribution

To model time series data, the benchmark is the i.i.d. recipe of the prerequisite courses:

$$\mathcal{P} = \{\text{candidate laws } P \text{ for } \textit{one} \text{ observation}\}\,, \qquad X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{P}.$$

- The joint law of the sample is the n-fold product: one function of one variable determines every probability.
- Fitting: pick the element that best explains the data (maximum likelihood, Bayes).
- Prediction: a future observation is just another draw — simulate from the fitted law.

The marginal does not determine the process



- In both panels every observation is exactly $N(0,1)$: the marginals coincide.
- In A, tomorrow is a fresh draw; in B, tomorrow equals today, forever: the joint families differ completely.
- A time series model must specify how past and future cohere — a collection of laws for one data point cannot say it.

Simulated data (seed 542, `figures.R`).

A process is its finite-dimensional distributions

Definition (time series)

A *time series* (or *stochastic process*) is a family $(X_t)_{t\in\mathbb{Z}}$ of $\mathbb{R}^d$-valued random vectors, specified through its *finite-dimensional distributions*: for every k and all time points $t_1 < \cdots < t_k$, the joint distribution

$$F_{t_1,\ldots,t_k}(c_1,\ldots,c_k) := \mathbb{P}\left(X_{t_1} \le c_1,\ldots,X_{t_k} \le c_k\right), \qquad c_1,\ldots,c_k \in \mathbb{R}^d,$$

the inequalities read componentwise when $d > 1$.

- To specify a process is to specify the joint law of every finite window.
- The family must be *consistent*: integrating out the last coordinate of $F_{t_1,\ldots,t_k}$ returns $F_{t_1,\ldots,t_{k-1}}$.
- General d covers the vector examples met earlier; the theory runs at $d = 1$.

A model is a collection of time series

Definition (time series model)

A *time series model* is a collection $\mathcal{P}$ of time series in the sense of the formal definition earlier — equivalently, a collection of consistent families of finite-dimensional distributions — each member regarded as a candidate mechanism for the observed series.

- The i.i.d. recipe, one level up: a member is no longer a law for one observation but a whole time series, every window specified jointly.
- The i.i.d. model sits inside as the factorizing case $F_{t_1,\dots,t_k}(c_1,\dots,c_k) = F(c_1)\cdots F(c_k)$, one distribution F determining every window.
- The modeling assumption: the observed window is one realization of one member of $\mathcal{P}$.

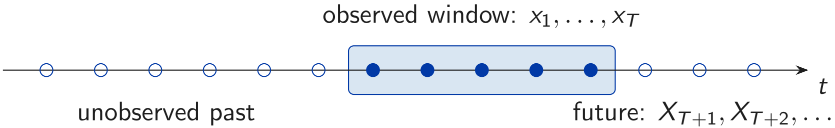
Structure enters through equations or conditional laws

Writing a consistent family down directly is not feasible. Two standing devices generate one:

$$\underbrace{X_t = g_t\left(Z_t, X_{t-1}, X_{t-2}, \ldots\right)}_{(Z_t) \text{ i.i.d. — the `noise'}} \qquad \text{or} \qquad \underbrace{\mathbb{P}\left(X_t \leq c \mid X_{t-1}, X_{t-2}, \ldots\right)}_{\text{a conditional law, given the past}}$$

- The specification is imposed at *every* t, not only over the observed stretch.
- The model must cover the future values to be predicted, and the past values that were never observed.

Finite data, infinite model



- In practice: a finite window of one realization.
- For modeling and theory: the process is defined for all time. This buys asymptotics ($T \to \infty$) and makes the organizing concepts of the course — stationarity, ergodicity, stability — expressible.

Without assumptions, one realization teaches nothing

Suppose the X_t may be completely arbitrary random variables. Then

$(X_1, \ldots, X_T)$ is a *single observation* from a completely unknown distribution on $\mathbb{R}^T$.

- Nothing can be concluded about that distribution — let alone about the future values $X_{T+1}, X_{T+2}, \ldots$
- The subject exists because regularity across time — the same mechanism operating throughout — can be assumed.

A generating equation defines a model implicitly

$$X_t = g\left(Z_t, X_{t-1}, X_{t-2}, \dots\right), \qquad g \in \mathcal{G}, \quad (Z_t) \text{ i.i.d. with a specified law.}$$

- The model: all time series satisfying the equation for some g in the class $\mathcal{G}$ and the given noise distribution.
- No distribution is ever listed — membership of $\mathcal{P}$ is decided by *solving* the equation.
- If no solution exists, $\mathcal{P}$ is empty: the consistency requirement has become an existence question.

A solution is a process, not a sequence of numbers

Solving the equation does not mean producing numbers. A solution is a whole process $(X_t)_{t\in\mathbb{Z}}$ — a consistent family of joint distributions — such that

$$X_t = g\,(Z_t, X_{t-1}, X_{t-2}, \dots) \quad \text{as an equality of random variables, at every } t.$$

- So the equation is a constraint on distributions: it may pin down one process, several, or none.
- Existence and uniqueness are genuine questions, taken up in the workbook on stability.

White noise: the raw material

Definition (white noise)

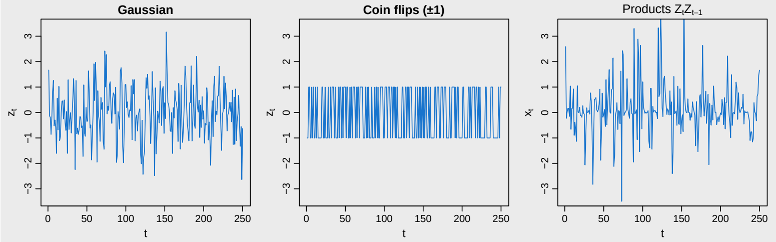
A process $(Z_t)_{t\in\mathbb{Z}}$ is *white noise* with variance σ^2, written $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$, if

$$\mathbb{E}[Z_t] = 0, \qquad \mathrm{Var}(Z_t) = \sigma^2, \qquad \mathrm{Cov}(Z_s, Z_t) = 0 \text{ for } s \neq t.$$

If the Z_t are moreover independent and identically distributed we speak of *i.i.d. noise*, and of *Gaussian white noise* if additionally $Z_t \sim N(0, \sigma^2)$.

- The definition constrains two moments and nothing else: many processes qualify, and $\mathrm{WN}(0, \sigma^2)$ denotes the whole collection — our first model.
- Not interesting on its own: it is the ingredient every construction below consumes.

Many different processes are white noise



- All three satisfy the definition with $\sigma^2 = 1$: Gaussian white noise; i.i.d. coin flips; the products $Z_t Z_{t-1}$ along an i.i.d. $N(0, 1)$ sequence.
- The third is uncorrelated but not independent — its squares are correlated: white noise that is not i.i.d. noise. Daily stock returns behave much like it.

Simulated data.

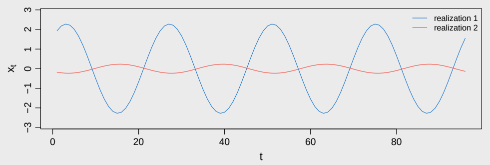
A random sinusoid is a time series

Example (a random sinusoid)

For random amplitudes A, B and a fixed frequency λ,

$$X_t = A\cos(\lambda t) + B\sin(\lambda t), \qquad t \in \mathbb{Z}.$$

- Every realization is a pure sinusoid: 'deterministic' to the eye, random only through amplitude and phase — a short stretch reveals the entire future.
- The class itself is not developed further; its idea — structure organized by frequency — returns in the frequency-domain unit.



The moving average: dependence from an explicit formula

Definition (moving average of order q)

Given white noise $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ and coefficients $\theta_1, \ldots, \theta_q$, the *moving average of order q*, MA(q), is

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad t \in \mathbb{Z}.$$

- An explicit formula: for every choice of coefficients the display exhibits a process — existence is free.
- The link to the model setup: the MA(q) time series, one for every choice of coefficients and noise process, form a model $\mathcal{P}$.
- $q = 0$ recovers white noise: the classes nest.

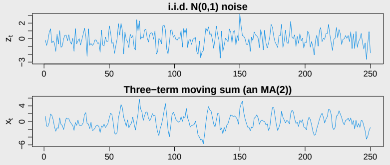
Overlapping innovations manufacture dependence

Example (a three-term moving sum)

With (Z_t) i.i.d. $N(0,1)$, an MA(2) with $\theta_1 = \theta_2 = 1$ is

$$X_t = Z_t + Z_{t-1} + Z_{t-2},$$

- Here "moving average" means a finite linear combination of current and past innovations; its coefficients need not sum to one.
- Neighboring values share noise terms and are therefore dependent; values three or more steps apart share none, so the dependence has a finite horizon.



The autoregression: an equation, not a formula

Definition (AR(p) mechanism)

Given white noise $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ and coefficients $\phi_1, \ldots, \phi_p$, the *autoregressive mechanism of order p*, AR(p), is the equation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \qquad \text{required at every } t \in \mathbb{Z}.$$

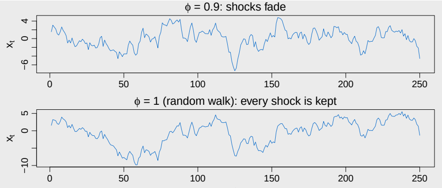
- The right-hand side involves X itself, so nothing is exhibited: the equation *constrains* a process. This is the implicit device — the AR(p) model is the set of time series satisfying the equation.
- Existence and uniqueness of a well-behaved solution now genuinely depend on the coefficients.

Feedback: today is built from yesterday

Example (AR(1), two coefficients)

$X_t = \phi X_{t-1} + Z_t$ with (Z_t) i.i.d. $N(0,1)$, run on the same noise draws with $\phi = 0.9$ and $\phi = 1$.

- At $\phi = 0.9$, shocks dissipate: the path keeps returning to the middle.
- At $\phi = 1$ — the random walk — every shock is kept forever and the path never settles. Which ϕ are safe, and in which sense, is the business of the workbook on stability.



GARCH: feedback in the variance

Definition (GARCH(1,1) mechanism)

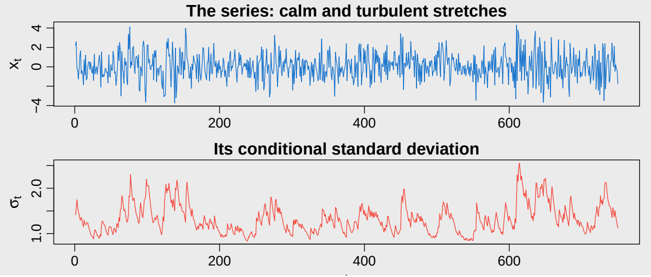
Given i.i.d. noise (Z_t) with mean 0 and variance 1, and parameters $\omega > 0$, $\alpha, \beta \geq 0$, the *GARCH(1,1) mechanism* is the pair of equations

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2.$$

- Feedback again, now in the size: yesterday's magnitude feeds today's variance. An implicit specification, like the autoregression — and whether a well-behaved solution exists depends on (α, β).
- $\alpha = \beta = 0$ recovers i.i.d. noise: the classes nest here too.

GARCH: white noise whose size has memory

- The solution is white noise — levels uncorrelated — but not i.i.d. noise: the magnitudes are dependent, and streaks form — calm days follow calm days, turbulent follow turbulent.
- The products example writ large, and the texture of daily asset returns: white noise that is not i.i.d. noise.



Stochastic volatility: the variance is its own process

Definition (stochastic volatility mechanism)

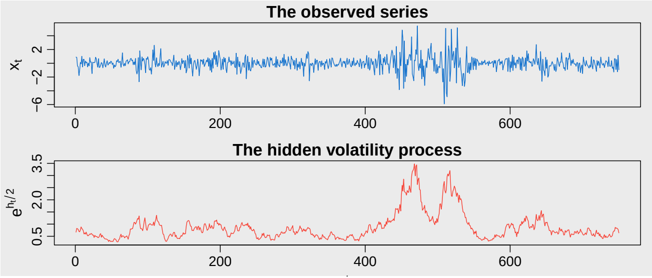
Given i.i.d. noise sequences (Z_t) and (η_t), independent of each other, the *stochastic volatility mechanism* is the pair

$$X_t = e^{h_t/2} Z_t, \qquad h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t.$$

- The log-variance h_t follows its own autoregression, driven by its own noise (η_t), and is never observed: a first latent-variable mechanism.
- Same phenomenon as GARCH, different mechanism: there the variance is fed by the observed past, here by hidden shocks. Existence reduces to the equation for h — an AR(1), just met.

Two noise sequences, one observed series

- Top: the observed series — calm and turbulent stretches again.
- Bottom: the volatility path that produced them. In data it is never seen; recovering it from the observed series alone is the estimation problem this model class carries.



What do we want from a time series?

Facing a series like those of the opening frame, we typically want some of the following.

- **Description.** Summarize the dependence: how strongly is today related to yesterday, to last month? Which regularities — trend, seasonality, cycles — are present?
- **Modeling.** Propose a mechanism that could plausibly have generated the realization; estimate its parameters; check its fit.
- **Forecasting.** Say something about $X_{T+1}, X_{T+2}, \ldots$ given the observed window, with uncertainty attached.
- **Inference under dependence.** Answer "is the warming trend real?" with standard errors that remain valid under correlation.
- **Validation.** Judge models on data they have not seen — without the future leaking into the training past.

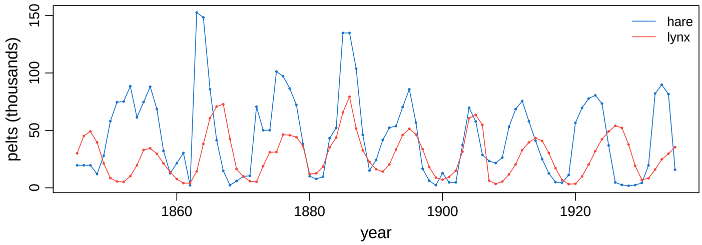
Estimating a parameter: the definitions carry over

Two of the five — inference and forecasting — can be made precise with the framework already on the table. Estimation first:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \qquad \hat{\theta}_T = g(X_1, \ldots, X_T), \qquad P_\theta(\theta \in C_T) \geq 1 - \alpha \text{ for all } \theta.$$

- A parametric model: one candidate time series P_θ per parameter value. The estimand is θ, or a feature of it; estimator and confidence interval are functions of the window.
- These are the definitions of the i.i.d. course, unchanged. The novelty: P_θ specifies the whole process jointly, so the coverage probability is computed under dependence.

The records suggest coupled predator–prey dynamics



- Hudson's Bay Company trading records, 1845–1935: pelts of snowshoe hare and Canada lynx, in thousands. The records are proxies for the underlying animal populations.
- Predator–prey feedback suggests two directions: lynx may reduce later hare abundance, while hare may support later lynx abundance.

Real data: `astsa::Hare`, `astsa::Lynx`.

The estimand is the hare-to-lynx coefficient

Let $X_{H,t}, X_{L,t}$ model the pelt records and set $Y_{H,t} = \log X_{H,t}$ and $Y_{L,t} = \log X_{L,t}$. Bundle the log histories as $\mathcal{H}_t = (\dots, Y_{H,t-1}, Y_{L,t-1}, Y_{H,t}, Y_{L,t})$. With $Z_{t+1} = (Z_{H,t+1}, Z_{L,t+1})^\top$, consider process laws satisfying

$$\begin{aligned} Y_{H,t+1} &= a_H + b_{HH} Y_{H,t} + b_{HL} Y_{L,t} + Z_{H,t+1}, & \mathbb{E}[Z_{t+1} \mid \mathcal{H}_t] &= 0, \\ Y_{L,t+1} &= a_L + b_{LH} Y_{H,t} + b_{LL} Y_{L,t} + Z_{L,t+1}, & \operatorname{Cov}(Z_{t+1} \mid \mathcal{H}_t) &= \Sigma. \end{aligned}$$

- X is the original record in thousands of pelts; Y is its log. The labels H and L identify the species throughout.
- The first subscript on b names the response and the second the predictor: b_{LH} multiplies $Y_{H,t}$ in the equation for $Y_{L,t+1}$.
- $b_{LH} = 0$ means current hare adds no linear one-step predictive information on the log scale once current lynx is included.

What does b_{LH} measure?

Holding current lynx fixed, b_{LH} describes how the current hare record changes the conditional mean of next year's log lynx record. Formally,

$$\begin{aligned}\{a_L + b_{LH}\log(cx_H) + b_{LL}\log x_L\} - \{a_L + b_{LH}\log x_H + b_{LL}\log x_L\} \\ = b_{LH}\log c.\end{aligned}$$

- $\widehat{b}_{LH} = 0.206$; nominal ordinary 95% interval $[0.113,\ 0.298]$.
- For $c = 2$, the fitted ratio of exponentiated conditional log means is 1.153 (15.3% larger).

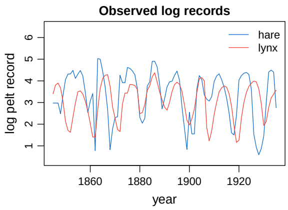
The exponentiated conditional log mean need not equal $\mathbb{E}[X_{L,t+1} \mid \mathcal{H}_t]$. The comparison is predictive, not causal.

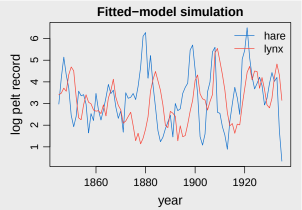
One lag leaves predictable lynx variation

Lag-one residual correlations: hare equation 0.03; lynx equation 0.52.

- Under the conditional-mean restriction, each innovation is uncorrelated with earlier innovations.
- The lynx residual correlation is evidence that one year of joint history leaves predictable structure unmodeled.
- The ordinary regression interval therefore remains nominal; its 95% coverage has not been justified for this fitted equation.

Simulation from the fit is not a model check





- To draw a path, we additionally take the yearly innovation vectors to be i.i.d. centered bivariate Gaussian with the fitted residual covariance.
- The simulated path contains linked fluctuations. Its resemblance to the data does not repair the residual failure in the lynx equation.

Predicting the next value: a different target

A one-step predictor is again a window function, aimed now at a random quantity. Under squared-error loss, its quality is measured by the mean squared prediction error:

$$\hat{X}_{T+1|T} = g\left(X_1, \ldots, X_T\right), \qquad \mathbb{E}\left[\left(X_{T+1} - \hat{X}_{T+1|T}\right)^2\right].$$

A level $1 - \alpha$ prediction interval is a window-computed set C_T with $\mathbb{P}\left(X_{T+1} \in C_T\right) \geq 1 - \alpha$.

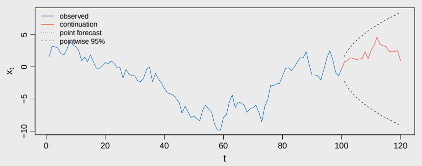
- A confidence interval concerns a fixed unknown number and may shrink as information accumulates.
- A prediction interval concerns a future random variable. It need not collapse even when the model is known, but it can collapse when the observed past determines the future.

Forecasting the random walk: a pointwise fan

Example (forecasting a random walk)

For the random walk with Gaussian white noise, the conditional distribution of X_{T+h} given the window is $N(x_T, h\sigma^2)$: for each fixed h, the point forecast is x_T and $x_T \pm 1.96\sqrt{h}\,\sigma$ is an exact 95% prediction interval.

- Across horizons, these intervals form a pointwise 95% fan, not a simultaneous 95% region for the entire future path.
- The forecast is flat and the fan widens like $\sqrt{h}$: every shock is kept, so uncertainty accumulates; mean-reverting processes differ.



How much is a century of correlated data worth?

- Every task on the list leans on the assumed regularity across time. The workbook on stationarity makes the assumption precise — *stationarity*, *ergodicity* — and asks what it buys.
- The first payoff there: the variance of a sample mean under dependence — in effect, how many independent observations a century of correlated ones is worth.