Advanced Econometrics
09 Time Series Models

Eduard Brüll

Fall 2025

Lecture 9

Advanced Econometrics

Literature: Greene Chapter 12; Wooldridge (2020) Chapter 10; Hamilton (1994) Chapters 1–3

9.1: The Generalized Regression Model

Recap: Ordinary Least Squares

Linear model:

$$y = X\beta + \varepsilon, \qquad E[\varepsilon|X] = 0, \quad Var(\varepsilon|X) = \sigma^2 I_n$$

Key properties:

- Unbiased and consistent under exogeneity.
- Efficiency proof relied on **equal and uncorrelated error variance** (Assumption A4).
- **BLUE** (Best Linear Unbiased Estimator) if errors are spherical:

$$Var(u|X) = \sigma^2 I_n$$

- Used when $Var(u_i|X) = \sigma^2 / w_i$ (known or estimable).
- WLS is a *special case* of GLS with diagonal $\Omega = W^{-1}$.

Takeaway: So far, we assumed uncorrelated errors. Next, we generalize to *correlated* errors and introduce the **GLS estimator**.

From Robust SEs to Efficient Estimation

We have seen that when errors are **not spherical**:

$$Var(\varepsilon|X) = \sigma^2\Omega, \quad \Omega \neq I_n,$$

OLS remains unbiased and consistent, but no longer efficient.

The **sandwich estimator** corrects inference:

$$\widehat{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1},$$

but leaves the point estimate unchanged.

Question: Can we improve not only inference, but also the estimator itself when Ω has known structure?

$\Rightarrow$ This leads to the **Generalized Least Squares (GLS)** estimator.

What is $\Omega^{-1/2}$?

We will transform our data with the matrix $\Omega^{-1/2}$ for **GLS**

Notation:

$$\Omega^{-1/2} \quad \text{is the matrix such that} \quad \Omega^{-1/2}\Omega(\Omega^{-1/2})' = I_n.$$

It is the **matrix square root of the inverse** of Ω:

$$(\Omega^{-1/2})'(\Omega^{-1/2}) = \Omega^{-1}.$$

Intuition:

- It rescales and rotates the data so that transformed errors become uncorrelated.
- In practice, we rarely compute it directly:
- For a diagonal Ω: divide by each σ_i (That's WLS if we use it to deal with non-homoskedastic variance).

The Generalized Regression Model

Consider:

$$y = X\beta + \varepsilon, \qquad E[\varepsilon|X] = 0, \quad Var(\varepsilon|X) = \sigma^2 I_n$$

where Ω is positive definite but not necessarily identity.

Pre-multiplying by $\Omega^{-1/2}$ yields transformed variables:

$$\tilde{y} = \Omega^{-1/2} y, \quad \tilde{X} = \Omega^{-1/2} X, \quad \tilde{u} = \Omega^{-1/2} u,$$

for which $Var(\tilde{u}|X) = \sigma^2 I_n$.

Applying OLS to these transformed data gives:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} y.$$

GLS transforms the errors to be uncorrelated with the same spread across observations

When Does GLS Help?

Case 1: Heteroskedasticity

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \quad \Rightarrow \quad w_i = 1/\sigma_i^2$$

$\Rightarrow$ Weighted Least Squares (WLS)

Case 2: Autocorrelation

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \dots \\ \rho & 1 & \rho & \dots \\ \rho^2 & \rho & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$\Rightarrow$ Time-ordered dependence of errors

GLS handles both — but *requires* knowledge (or estimation) of Ω.

OLS vs GLS in a Nutshell

OLS:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y, \quad Var(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}$$

GLS:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, \quad Var(\hat{\beta}_{GLS}|X) = \sigma^2(X'\Omega^{-1}X)^{-1}$$

- GLS is **BLUE** under non-spherical errors.
- OLS is a special case when $\Omega = I_n$.
- In practice: Ω unknown $\Rightarrow$ use *Feasible GLS (FGLS)*.

Alternative: Robust (sandwich) SEs maintain validity without changing $\hat{\beta}$.

Why Feasible GLS Is Less Popular Than Huber–Eicker–White

Feasible GLS (FGLS):

$$\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$$

requires estimating the full covariance structure $\hat{\Omega}$.

Practical issues:

- Estimating Ω reliably is hard; especially in small samples or with complex dependence.
- Model misspecification of $\hat{\Omega}$ can make FGLS **less efficient than OLS**.

Why robust SEs became dominant:

- Huber-Eicker-White ("sandwich") inference works under very general forms of heteroskedasticity and autocorrelation.
- No need to specify or estimate Ω.
- Easy to compute and valid in large samples even if the true structure is unknown.

Bottom line:

FGLS is efficient only if the covariance model is correct; robust SEs are **safer and simpler** in applied work.

From Correlated Errors to Serial Correlation

GLS reminded us: when Ω has off-diagonal elements, errors are **correlated across observations**.

In most time-series data:

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \dots \\ \rho & 1 & \rho & \dots \\ \rho^2 & \rho & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \implies Cov(u_t, u_{t-h}) = \rho^h \sigma^2$$

Interpretation:

- Correlated errors often arise from **time persistence** in omitted variables.
- We will later see a model where: $u_t = \rho u_{t-1} + \varepsilon_t$ due to persistence $\Rightarrow$ Then, even if ε_t is white noise, the composite error u_t will be serially correlated.

Next: We'll illustrate this with the **ice-cream consumption example.**

Example: Omitted Variables and Serial Correlation

The ice cream example:

- Suppose we regress daily ice cream consumption y_t on price p_t:

$$y_t = \beta_0 + \beta_1 p_t + u_t.$$

- True data-generating process includes temperature T_t:

$$y_t = \beta_0 + \beta_1 p_t + \beta_2 T_t + \varepsilon_t,$$

  but we omit T_t.

What happens?

- The omitted factor T_t varies smoothly over time: Hot days follow hot days.
- $\Rightarrow$ The composite error term

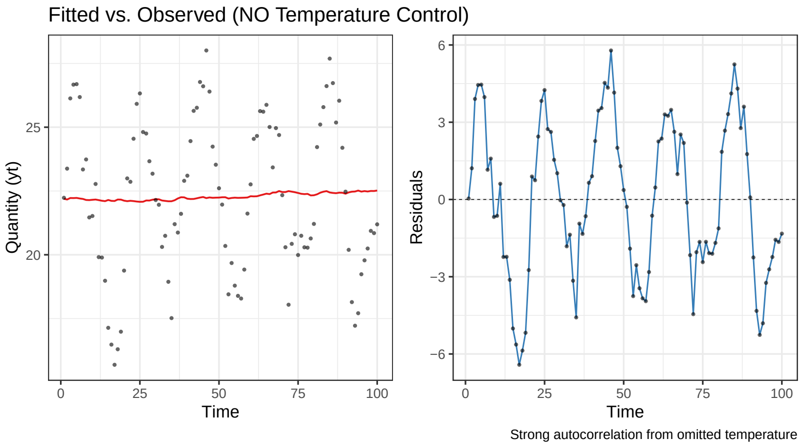
$$u_t = \beta_2 T_t + \varepsilon_t$$

  is **serially correlated**, even if ε_t itself is white noise.

Intuition:

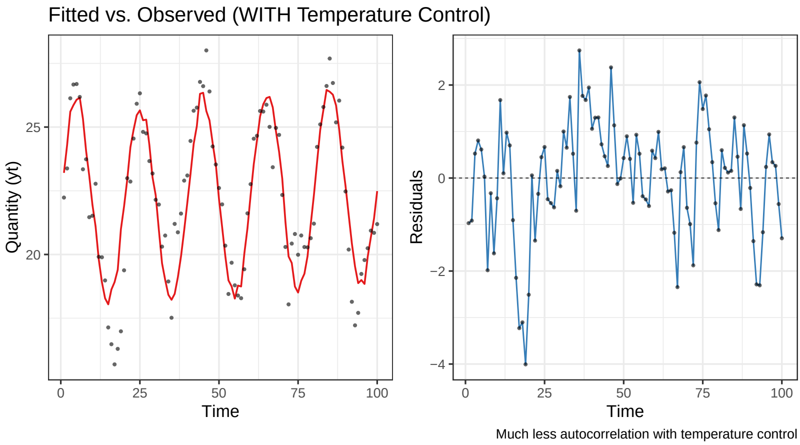
- OVB in time series can manifest as **autocorrelation of residuals**.
- The problem isn't just inefficiency: Autocorrelation hints at **missing dynamics or omitted variables**.

Illustration: The Ice-Cream Example



- **True driver:** temperature T_t changes gradually over time.
- If T_t omitted, the residual inherits its correlations structure over time.

Illustration: The Ice-Cream Example



- **True driver:** temperature T_t changes gradually over time.
- If T_t omitted, the residual inherits its correlations structure over time.

From GLS to Time-Series Models

Why discuss GLS now?

- In cross-section data, Ω captures **heteroskedasticity**.
- In time-series data, Ω captures **serial correlation** between errors.

Example: AR(1) error process

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad |\rho| < 1.$$

Implications:

- OLS remains unbiased (if exogenous) but is **no longer efficient**.
- GLS (or feasible GLS) can restore efficiency by accounting for correlation.
- Serial correlation often signals **dynamic structure** in the data:

$$y_t \text{ may depend on } y_{t-1}, \text{ not just } X_t.$$

Next step:

We move from correcting correlated errors (GLS) to explicitly **modeling dynamics** in y_t through AR and dynamic regression models.

Why Time-Series Feels Different

Applied micro, finance, and macro all study causality, but their data worlds differ

Cross-section / panel world:

- Units are independent or conditionally independent.
- Variation comes from **who** or **where**.
- We seek **causal identification**: treatment effects, policy impacts.

Time-series world:

- Observations are dependent through time: today's data predict tomorrow.
- Variation comes from **when**.
- We seek **dynamic structure** and **forecasting ability**.

Consequence: Even when we use the same estimators (OLS, ML, GMM), the questions and diagnostics are different.

A Different Kind of Identification

In applied micro, identification means isolating a **causal effect**. In time-series, it often means isolating a **dynamic mechanism**.

Applied micro:

- "What would have happened without the policy?"
- Exogeneity through design or instruments.

Time-series:

- "How does a one-time shock propagate over time?"
- Identification through **model structure** (lags, filters, restrictions).

Takeaway: Time-series econometrics is less about finding quasi-experiments and more about describing and testing the laws of motion that generate the data.

Time-Series Thinking: Common in Finance, Rarer in Applied Micro

Finance and macroeconomics: Time-series methods are part of everyday empirical work.

- Forecasting returns, volatility, and risk premia
- Market efficiency and asset-pricing tests depend on serial dependence structures
- Typical data: high-frequency, long panels, few identification concerns but strong modeling discipline

Applied micro and business fields: Usually rely on cross-sectional or panel identification.

- Emphasis on exogeneity and local variation (DiD, IV, RCT).
- Time-series often treated as nuisance (trends, persistence, serial correlation).
- Yet: dynamic responses and long-run adjustment paths are inherently *time-series questions*.

Yagan, D. (2019). *Employment Hysteresis from the Great Recession*. American Economic Review, 109(10), 3495–3529.

Research question: Did local labor markets hit hardest by the Great Recession experience persistent employment losses even after the recovery?

Challenge: Distinguishing temporary cyclical shocks from long-term structural effects in aggregate employment data.

Approach: Combines panel microdata on local labor markets with a **forecasting-based time-series counterfactual** for employment recovery paths.

Main Result:

Shows strong persistence (“hysteresis”): local shocks have long-lasting employment effects.

9.2: Dynamic Linear Models

Dynamic Linear Models

Dynamic models arise when past realizations of y_t or the errors affect current outcomes.

$$y_t = x_t'\beta + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

Examples of sources of dynamics:

- Partial adjustment or habit formation
- True state dependence
- Serial correlation of omitted factors

Substituting the AR(1) error structure:

$$y_t = \rho y_{t-1} + x_t'\beta - \rho x_{t-1}'\beta + u_t$$

$\Rightarrow$ Lagged dependent variables can arise from either genuine dynamics or serial correlation.

The First-Order Autoregression: AR(1)

A simple way to describe time dependence is to relate y_t directly to its own past:

$$y_t = \mu + \rho y_{t-1} + u_t,$$

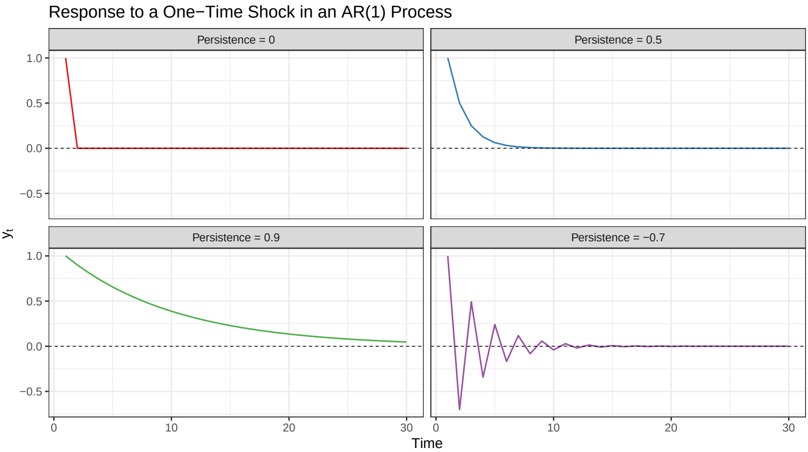
where u_t is a white noise shock (zero mean, uncorrelated over time).

Interpreting the parameter ρ:

Value of ρ	**Behavior of the process**
$\rho = 0$	No persistence. Each y_t is independent noise.
$0 < \rho < 1$	Positive persistence. Shocks decay gradually over time.
$\rho \approx 1$	Very high persistence.
$\rho < 0$	Negative persistence. The series alternates above and below its mean.

When ρ is small, y_t quickly forgets past shocks. When ρ is near one, shocks fade only slowly. When ρ is negative, the process each shock flips the direction of the next period.

AR-1 Impulse Response for different ρ



Each line shows how a one-time shock today affects future y_t values, depending on ρ. We abstract from the white-noise error u_t in every period to isolate persistence.

Why start here?

- The **AR(1) process** is the simplest way to capture persistence: how today's outcome depends on yesterday's.
- It provides a clean laboratory to see how serial correlation, stability, and long-run behavior are connected.
- Understanding AR(1) intuition helps interpret more complex dynamic models:
  - When persistence reflects genuine state dependence
  - When it reflects correlated shocks or omitted dynamics
- In the next section, we'll formalize what makes such a process *stable* or *stationary*.

GLS and Dynamic Models: Connecting the Dots

Recall from Section 9.1:

$$Var(\varepsilon|X) = \sigma^2\Omega, \quad \Omega \neq I_n.$$

GLS transforms the data to remove correlation in ε_t:

$$\tilde{y} = \Omega^{-1/2}y, \quad \tilde{X} = \Omega^{-1/2}X.$$

In time series data, Ω typically has an AR(1) structure:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

Two equivalent viewpoints:

- **GLS approach:** Treat serial correlation as a nuisance and transform the data to get spherical errors
- **Dynamic model approach:** Treat serial correlation as evidence of meaningful time dependence in y_t itself

Both perspectives address the same issue: Persistence over time

- GLS focuses on efficiency
- Dynamic models seek **economic interpretation**

9.3: Serial Correlation and Stationarity

Serial Correlation and Stability

Big question

How do we know whether a dynamic model describes a process that settles down or one that drifts endlessly?

Example: AR(1) process

$$y_t = \mu + \rho y_{t-1} + u_t$$

Two possibilities:

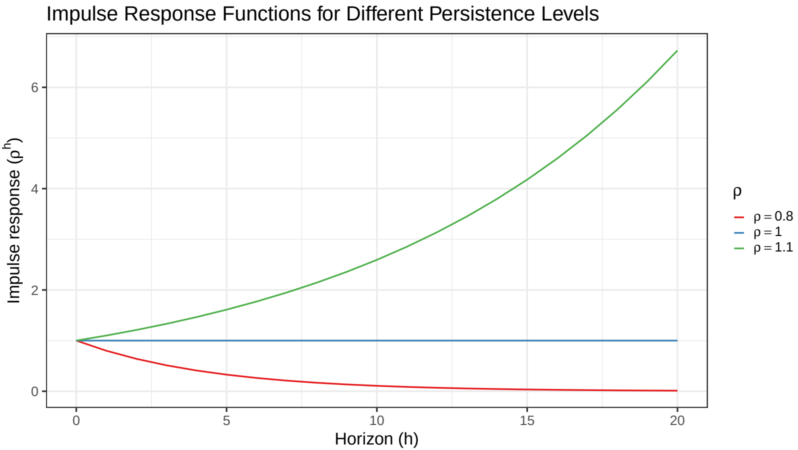
- $|\rho| < 1$: shocks die out $\Rightarrow$ process is **stable**.
- $|\rho| \geq 1$: shocks persist or explode $\Rightarrow$ process is **unstable**.

Why we care:

- Stability ensures the process has a fixed mean and variance.
- Otherwise, we can't summarize its long-run behavior. Every shock leaves a permanent mark!

Next: The formal notion of **stationarity** makes this idea precise.

Illustration: AR(1) with $\rho \geq 1$



Why Stationarity Matters: A Cautionary Example

Consider a simple AR(1) process:

$$y_t = \mu + \rho y_{t-1} + u_t, \quad u_t \sim (0, \sigma_u^2)$$

If $\rho = 1$, the process has a **unit root**:

$$y_t = y_{t-1} + u_t \quad \Rightarrow \quad y_t = y_0 + \sum_{i=1}^{t} u_i$$

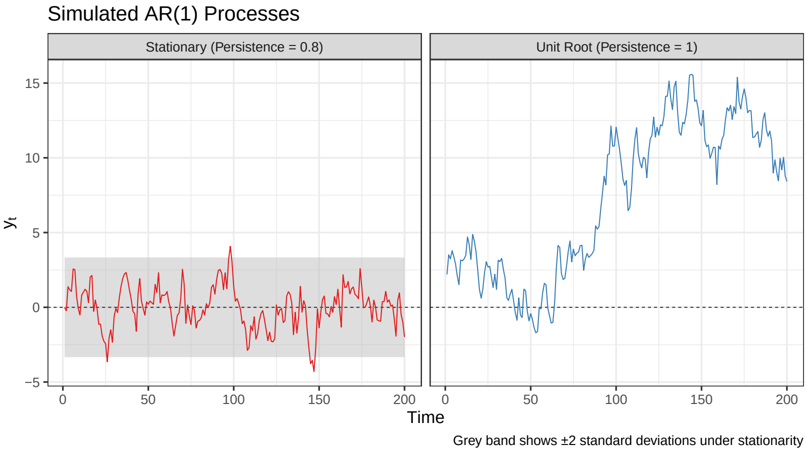
Implication:

- Variance grows over time: shocks have permanent effects.
- The process never settles around a stable mean.

Why this matters:

- When moments change over time, regression results lose meaning.
- Econometric theory (LLN, CLT) relies on **time-invariant distributions**.
- That property is what we'll call **stationarity**.

Illustration: Stationary vs. Unit Root AR(1)



Interpretation:

- **Left:** Stationary AR(1) with $\rho = 0.8$ fluctuates within a constant standard deviation band .
- **Right:** Unit root process ($\rho = 1$) shows variance increasing over time.

A process is **stationary** if its statistical properties stay the same over time.

Formally (weak stationarity):

1. $\mathbf{E}[y_t] = \mu$ (constant mean)
2. $\mathbf{var}(y_t) = \sigma_y^2$ (constant variance)
3. $\mathbf{cov}(y_t, y_{t-k}) = \gamma_k$ depends only on the lag k

Interpretation:

- The process behaves the same way today, tomorrow, or next year.
- We can use one sample path to learn about its long-run distribution.
- If these properties drift over time $\longrightarrow$ inference breaks down.

Condition for Stationarity

Iterate the process:

$$y_t = \mu(1 + \rho + \rho^2 + \cdots) + \rho^t y_0 + \sum_{i=0}^{t-1} \rho^i u_{t-i}.$$

For $|\rho| < 1$:

$$\rho^t \to 0 \Rightarrow y_t \text{ forgets its starting point.}$$

Therefore: The AR(1) is *stationary* if and only if $|\rho| < 1$.
Intuition: shocks fade out geometrically.

When Stationarity Meets a Trend

If many time series, like stock prices or GDP, keep going up, doesn't that break AR(1) stationarity?
No!

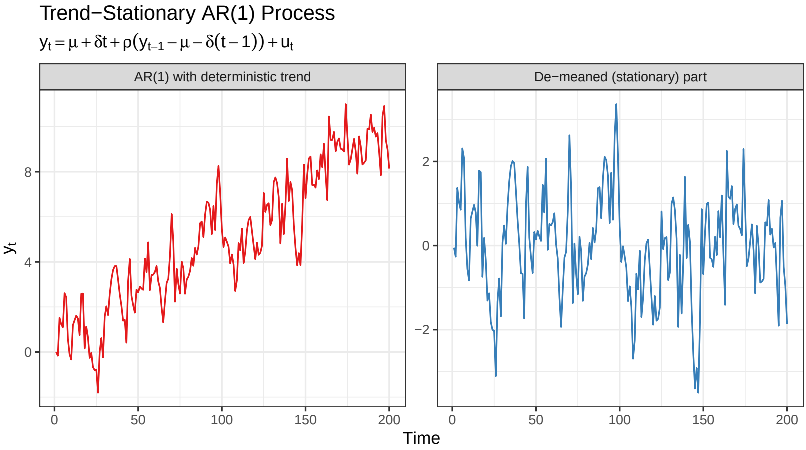
- A process can be **stationary around a trend**.
- The trend may be:
  - **Deterministic:** fixed slope δ (trend-stationary).
  - **Stochastic:** random walk component (difference-stationary).

Example:

$$y_t = \mu + \delta t + \rho(y_{t-1} - \mu - \delta(t-1)) + u_t$$

- δt: deterministic linear trend
- ρ: AR(1) persistence around that moving mean

Illustration: Stationarity with a Trend



A trend doesn't violate stationarity: We just need to remove it to analyze persistence.

Stationary vs. Ergodic: The Intuition

Stationary: The *rules of the game* don't change over time.

- Mean, variance, and covariances stay constant.
- What happens today follows the same distribution as what could happen tomorrow.
- Required for a process to have stable long-run properties.

Ergodic: The process is *not path dependent*.

- In the long run, the starting point doesn't matter.
- Past events lose influence as time passes.
- Time averages $\approx$ population averages $\Rightarrow$ allows inference from a single realization.

Example: Is the prevalence of **QWERTY** keyboards a case of path dependence (non-ergodicity)?

See: Paul David (1985), AER, and Liebowitz & Margolis (1994), JEP.

Ergodicity: The Intuition

In time series, we usually observe only **one realization** of the process $\{y_t\}_{t=1}^{T}$, i.e. one country's GDP, one firm's stock price, one macro variable.

The question: Can we learn about population properties (like $\mathbf{E}[y_t]$) from this single history through time?

Analogy:

- In cross-sections, we average over *individuals*.
- In time series, we average over *time*.

Key idea:

- **Stationarity** says the process behaves the same way at all times.
- **Ergodicity** says that one long history reveals those stable properties.

If a process "forgets" where it started, its long-run average tells us the true mean.

Ergodicity: Formal Statement

Formally, a process $\{y_t\}$ is **ergodic** if time averages converge to ensemble averages:

$$\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t \quad \xrightarrow{p} \quad \mathbf{E}[y_t].$$

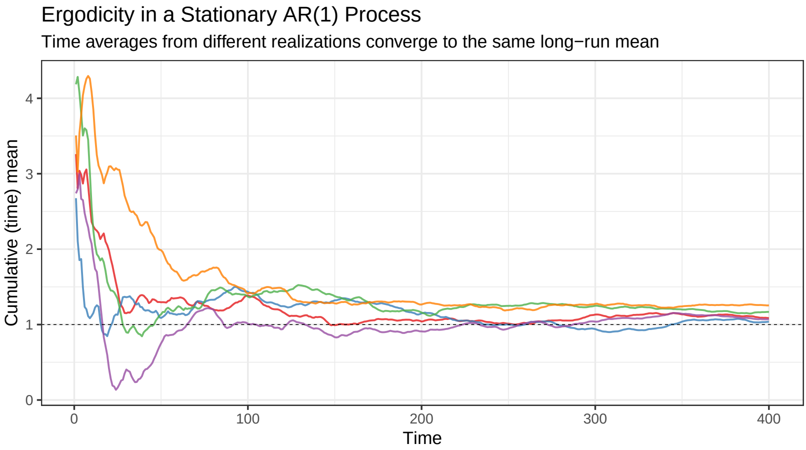
Why this matters:

- **Stationarity** $\Rightarrow$ moments exist and don't change over time.
- **Ergodicity** $\Rightarrow$ those moments can be consistently estimated from one time series.
- Justifies treating time averages as estimators of expected values.

Example:

For a stationary AR(1) process with $|\rho| < 1$, the influence of initial conditions fades away, so the sample mean converges to the true mean.

Illustration: Ergodicity



Example: Stationary but Non-Ergodic Process

Idea: A process can have constant unconditional moments (stationary) but still fail to "forget" its initial state (non-ergodic).

Example: Mixture of Two Mean Regimes

$$y_t = \mu_i + u_t, \qquad \mu_i = \begin{cases} +1, & \text{with prob. } 0.5 \\ -1, & \text{with prob. } 0.5 \end{cases} \quad \text{and} \quad u_t \sim \text{i.i.d. } N(0,1)$$

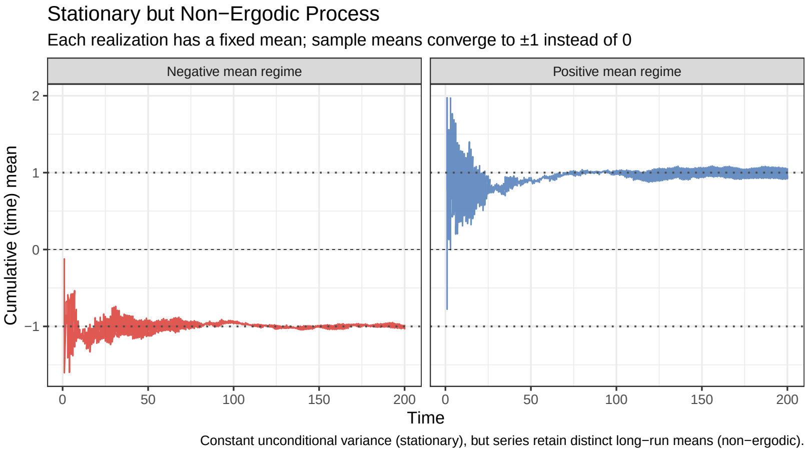
Each realization draws its own regime μ_i once and for all at $t = 0$.

Properties:

- $\mathbf{E}[y_t] = 0$, $\mathbf{var}(y_t) = 2 \Rightarrow$ **stationary.**
- But each sample path remains around its own mean $\mu_i = \pm 1$ forever.
- The sample mean $\bar{y}_T \to \mu_i$ instead of 0.
- $\Rightarrow$ Time averages $\neq$ Ensemble averages $\Rightarrow$ **non-ergodic.**

Intuition: The process has stable distributional moments, but it never "mixes" across its two mean regimes.

Illustration: Stationary but Non-Ergodic



Deriving the Covariance for AR(1)

By repeated substitution:

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \cdots$$

Variance:

$$Var(\varepsilon_t) = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \rho^6\sigma_u^2 + \cdots = \sigma_u^2(1 + \rho^2 + \rho^4 + \cdots)$$

A geometric series:

$$1 + \rho^2 + \rho^4 + \cdots = \frac{1}{1-\rho^2} \quad \text{for } |\rho| < 1$$

$$\Rightarrow Var(\varepsilon_t) = \frac{\sigma_u^2}{1-\rho^2}.$$

Covariances:

$$Cov(\varepsilon_t, \varepsilon_{t-s}) = \rho^s Var(\varepsilon_t) = \frac{\rho^s\sigma_u^2}{1-\rho^2}.$$

Hence,

$$Corr(\varepsilon_t, \varepsilon_{t-s}) = \rho^s,$$

which decays geometrically with the lag s.

Mean and Variance of a Stationary AR(1)

If $|\rho| < 1$:

$$\mathbf{E}[y_t] = \frac{\mu}{1-\rho}, \quad \mathbf{var}(y_t) = \frac{\sigma_u^2}{1-\rho^2}.$$

Interpretation:

- $\mu/(1-\rho)$ is the level around which y_t fluctuates.
- Higher $|\rho| \longrightarrow$ larger variance $\longrightarrow$ more persistence.

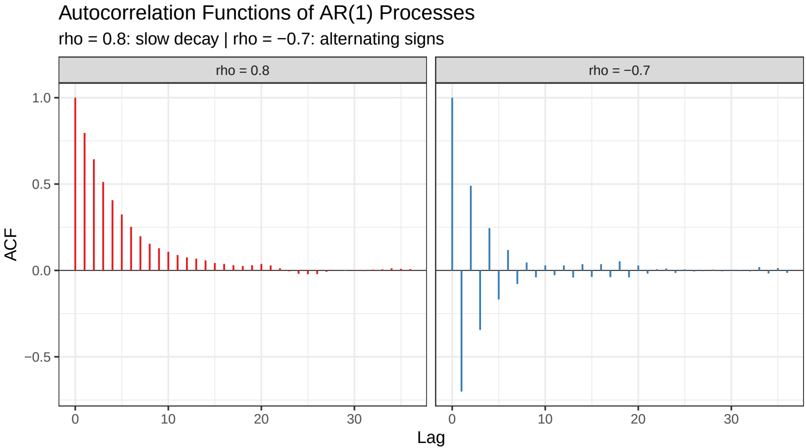
Autocorrelation Function (ACF) of AR(1)

$$\rho_k = \frac{\mathbf{cov}(y_t, y_{t-k})}{\mathbf{var}(y_t)} = \rho^k.$$

Implications:

- Correlation between y_t and y_{t-k} decays geometrically.
- The closer $|\rho|$ is to 1, the slower the decay $\longrightarrow$ strong persistence.
- Negative $\rho \longrightarrow$ alternating signs in ρ_k.

Illustration: Autocorrelation Function of AR(1)



The first bar (lag 0) equals 1 by definition, since any variable is perfectly correlated with itself.

Covariance Matrix of AR(1)

For T observations from a stationary AR(1):

$$\Sigma_y = \sigma_y^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}$$

$\Rightarrow$ This is exactly the Ω matrix structure we introduced earlier in GLS. It formalizes how serial correlation generates non-diagonal covariance.

OLS with AR(1) Errors: When Does It Work?

Consider the model:

$$y_t = x_t'\beta + \varepsilon_t, \qquad \varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad u_t \sim (0, \sigma_u^2)$$

Assume the process is:

- **Stationary:** mean, variance, and covariance constant over time
- **Ergodic:** long-run averages represent population moments

Then:

- OLS is **unbiased and consistent** if regressors are strictly exogenous:
$$\mathbf{E}[\varepsilon_t | X] = 0$$
- But if y_{t-1} appears as a regressor, it is correlated with ε_t:
$$Cov(y_{t-1}, \varepsilon_t) = \rho Var(\varepsilon_{t-1}) \neq 0$$

$\Rightarrow$ OLS becomes biased and inconsistent.

The AR(p) Process

Generalizing AR(1):

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + u_t, \quad u_t \sim (0, \sigma_u^2).$$

Stability condition: All roots of

$$1 - \rho_1 z - \rho_2 z^2 - \cdots - \rho_p z^p = 0$$

must lie **outside the unit circle** ($|z| > 1$).

Estimation:

- OLS is consistent if u_t is white noise and y_t is stationary.
- Information criteria (AIC, BIC) help choose lag order p.

The Lag Operator: Intuition

Definition:

$$Ly_t = y_{t-1}, \quad L^2 y_t = y_{t-2}, \quad L^k y_t = y_{t-k}.$$

The lag operator simply “shifts” a variable back in time.

Why we use it:

- Compactly writes dynamic models, e.g.:

$$(1 - \rho L) y_t = \mu + u_t.$$

- Treats lagged values like algebraic terms in a polynomial.
- Makes it easier to manipulate distributed lags or derive long-run effects.

Example: $(1 - 0.5L) y_t = u_t \Rightarrow y_t = u_t + 0.5u_{t-1} + 0.25u_{t-2} + \dots$

Lag Polynomials and Rational Lag Models

Lag polynomial notation:

$$A(L) = 1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_p L^p.$$

Autoregressive model:

$$A(L)y_t = u_t.$$

Distributed-lag model:

$$y_t = B(L)x_t + u_t, \quad B(L) = b_0 + b_1 L + b_2 L^2 + \cdots$$

Rational-lag form:

$$y_t = \frac{B(L)}{A(L)} x_t + u_t$$

combines dynamic adjustment (denominator) and delayed responses (numerator).

Key takeaway:

Multiplying or inverting lag polynomials works like algebra—just remember L means “one step back in time.”

Mean and Long-Run Effects in Lag Models

For a distributed-lag model

$$y_t = b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \cdots + u_t,$$

the coefficients b_i describe how past values of x influence y_t.

Key summary measures:

$$\text{Short-run effect:} \quad b_0$$

$$\text{Long-run multiplier:} \quad \sum_{i=0}^{\infty} b_i$$

$$\text{Mean lag:} \quad \frac{\sum_{i=0}^{\infty} i b_i}{\sum_{i=0}^{\infty} b_i}$$

Intuition:

- b_0: immediate impact of a change in x_t.
- Long-run multiplier: total accumulated response over time.
- Mean lag: average delay until the effect of x_t is felt.

9.4: Inference and Robust Estimation

Why the OLS Covariance Formula Fails

OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

$$Var(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'Var(\varepsilon|X)X(X'X)^{-1}$$

Under serial correlation:

$$Var(\varepsilon|X) = \sigma^2\Omega, \quad \text{with} \quad \Omega_{ij} = \rho^{|i-j|}$$

$\Rightarrow$

$$Var(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Problem: Using $\sigma^2(X'X)^{-1}$ (the "usual" formula) implicitly assumes $\Omega = I_T$.
$\Rightarrow$ OLS SEs are **too small** when residuals are positively autocorrelated.

What Standard Errors Do We Want?

If errors follow an AR(1) process, we want $Var(\hat{\beta}_{OLS}|X)$ to be robust to serial correlation:

$$\widehat{Var}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{t=1}^{T} \sum_{s=1}^{T} x_t x_s' \hat{\varepsilon}_t \hat{\varepsilon}_s \right) (X'X)^{-1}.$$

Simpler representation:

$$(X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1},$$

where $\hat{\Omega}$ estimates the serial covariance of residuals.

Two main approaches:

- **(1) Newey–West (HAC)** SEs — consistent under general autocorrelation and heteroskedasticity.
- **(2) Feasible GLS (Prais–Winsten or Cochrane–Orcutt)** — if AR(1) structure is approximately known.

Testing for Autocorrelation

- **Quick diagnostic:** Regress residuals $\hat{e}_t$ on their lagged values $\hat{e}_{t-1}$:

$$\hat{e}_t = \hat{\rho}\hat{e}_{t-1} + v_t$$

  A significant $\hat{\rho}$ indicates autocorrelation. (Use robust SEs if heteroskedasticity is suspected.)

Durbin–Watson (DW) test:

$$DW = \frac{\sum_{t=2}^{T}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T}\hat{e}_t^2} = 2(1 - \hat{\rho})$$

- $DW \approx 2$: no autocorrelation ($\hat{\rho} \approx 0$)
- $DW < 2$: positive autocorrelation ($\hat{\rho} > 0$)
- $DW > 2$: negative autocorrelation ($\hat{\rho} < 0$)

Limitations: DW is mainly valid for AR(1) errors and models without lagged dependent variables.

Lagrange Multiplier (Breusch–Godfrey) Test

General and flexible test for serial correlation:

1. Estimate the model and obtain residuals $\hat{e}_t$.
2. Regress $\hat{e}_t$ on its own p lags and the original regressors x_t:

$$\hat{e}_t = \alpha_0 + \sum_{s=1}^{p} \rho_s \hat{e}_{t-s} + x_t' \gamma + u_t$$

3. Test joint significance of $\rho_1, \ldots, \rho_p$.

Test statistic:

$$TR^2 \sim \chi^2(p)$$

under the null hypothesis $H_0 : \rho_1 = \cdots = \rho_p = 0$.

Interpretation:

- High $R^2 \Rightarrow$ strong evidence of autocorrelation.
- Works with lagged dependent variables and higher-order AR processes.

Q-Test for Serial Correlation (Ljung–Box Statistic)

Often used when there are no regressors x_t (pure time series):

$$Q' = T(T+2)\sum_{s=1}^{P} \frac{r_s^2}{T-s} \quad \sim \quad \chi^2(P)$$

where

$$r_s = \frac{\sum_{t=s+1}^{T} \hat{e}_t \hat{e}_{t-s}}{\sum_{t=1}^{T} \hat{e}_t^2}$$

Interpretation:

- Q' aggregates autocorrelations up to lag P.
- Significant $Q' \Rightarrow$ reject H_0: residuals are white noise.
- Especially useful for diagnosing ARMA model residuals.

Remark: The Ljung–Box version improves small-sample accuracy over the original Box–Pierce test.

Adding Regressors: What Changes?

So far, we studied univariate processes:

$$y_t = \mu + \rho y_{t-1} + u_t, \qquad u_t \sim \text{white noise.}$$

Now let's allow y_t to depend on external variables x_t:

$$y_t = x_t'\beta + u_t.$$

New issues in a time–series context:

- **Time ordering:** x_t must be known at time t. $\Rightarrow$ requires strict exogeneity: $E[u_t|x_s] = 0$ for all s.
- **Persistent regressors:** Serially correlated x_t can make residuals appear autocorrelated.
- **Dynamic misspecification:** Autocorrelation in $\hat{u}_t$ may signal missing lags:

$$y_t = \rho y_{t-1} + x_t'\beta + \varepsilon_t.$$

Review: Lagged Dependent Variables

In dynamic models, the lagged outcome y_{t-1} often appears as a regressor:

$$y_t = \rho y_{t-1} + x_t'\beta + \varepsilon_t.$$

Problem: If ε_t is serially correlated, y_{t-1} becomes correlated with ε_t:

$$\mathbf{cov}(y_{t-1}, \varepsilon_t) \neq 0.$$

Consequence:

- OLS is biased and inconsistent.
- The bias depends on ρ and the degree of autocorrelation.

Remedies:

- Use **instrumental variables** (e.g. y_{t-2} or external instruments).
- Estimate by **Arellano–Bond GMM** if in panel context.

9.5: Forecasting and ARMA Models

Forecasting in an AR(1) Model

Consider the stationary AR(1) process:

$$y_t = \mu + \rho y_{t-1} + u_t, \qquad u_t \sim (0, \sigma_u^2),\ |\rho| < 1.$$

The optimal *h*-step-ahead forecast (under squared-error loss):

$$\hat{y}_{t+h|t} = \mu(1 + \rho + \cdots + \rho^{h-1}) + \rho^h y_t = \mathbf{E}_t[y_{t+h}].$$

Forecast properties:

- As h increases, $\rho^h \to 0 \Rightarrow$ forecasts converge to the unconditional mean $\mu/(1-\rho)$.
- Forecasts are linear in the most recent value y_t.
- The smaller $|\rho|$, the faster convergence to the long-run mean.

Intuition: Persistence ($|\rho|$) governs how quickly shocks lose predictive power.

Forecast Error Variance in AR(1)

For an h–step–ahead forecast:

$$y_{t+h} - \hat{y}_{t+h|t} = \sum_{i=0}^{h-1} \rho^i u_{t+h-i}.$$

Hence

$$Var(y_{t+h} - \hat{y}_{t+h|t}) = \sigma_u^2 \frac{1-\rho^{2h}}{1-\rho^2}.$$

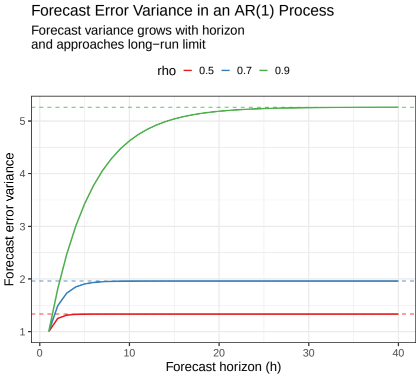
Interpretation:

- Forecast uncertainty grows with the horizon h.
- As $h \to \infty$, it approaches the unconditional variance $\sigma_u^2/(1-\rho^2)$.
- More persistence $\Rightarrow$ slower convergence to long-run variance.

Forecast Error Variance in an AR(1) Process

$$y_{t+h} - \hat{y}_{t+h|t} = \sum_{i=0}^{h-1} \rho^i u_{t+h-i}$$

$$Var(y_{t+h} - \hat{y}_{t+h|t}) = \sigma_u^2 \frac{1-\rho^{2h}}{1-\rho^2}$$



- Forecast uncertainty increases with horizon h.
- Dashed lines show the long-run limit:

$$\lim_{h\to\infty} Var = \frac{\sigma_u^2}{1-\rho^2}$$

- Higher persistence $\rho \Rightarrow$ slower convergence.

From Persistence to Short Memory

So far, we've modeled dependence as coming from past values of y_t:

$$y_t = \mu + \rho y_{t-1} + u_t \quad \text{(AR(1))}$$

But time dependence can also come from **overlapping shocks:**

$$y_t = \mu + u_t + \theta u_{t-1}.$$

This is a Moving Average (MA) model.

Intuition:

- Each period's value reflects not only today's shock but also yesterday's.
- Creates short-lived correlation even though u_t are i.i.d.
- Unlike AR(1), dependence **dies out quickly.**

The Moving Average Model

A first-order Moving Average process:

$$y_t = \mu + u_t + \theta u_{t-1}, \quad u_t \sim \text{i.i.d. } (0, \sigma_u^2)$$

Interpretation:

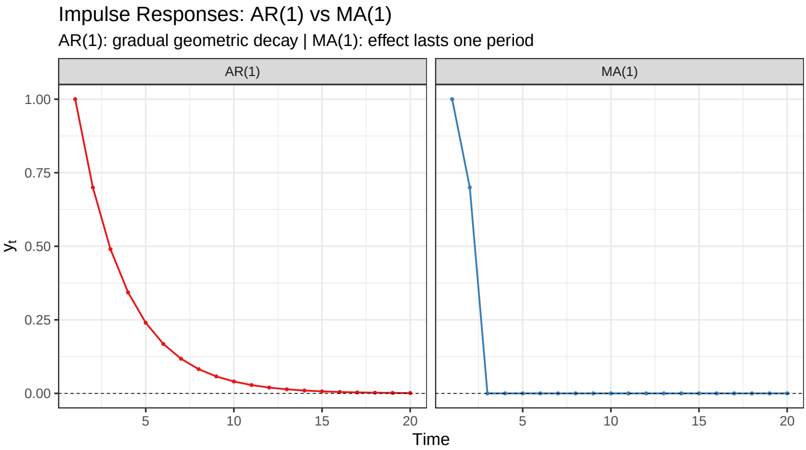
- u_t: new shock today
- θu_{t-1}: echo of yesterday's shock
- Correlation comes from overlapping terms u_{t-1}

Autocovariances:

$$\gamma_0 = (1 + \theta^2)\sigma_u^2,$$
$$\gamma_1 = \theta\sigma_u^2,$$
$$\gamma_k = 0 \quad \text{for } k \geq 2.$$

Implication: Serial correlation is real but short-lived: one period only.

Impulse Response: AR(1) vs. MA(1)



AR(1): Shock effects decay geometrically.

MA(1): Shock lasts one period only, the system "forgets" instantly.

AR(1) vs. MA(1): Two Ways to Create Dependence

Autoregressive (AR(1))	Moving Average (MA(1))
$y_t = \mu + \rho y_{t-1} + u_t$	$y_t = \mu + u_t + \theta u_{t-1}$
Dependence through past y_t	Dependence through past shocks u_t
Persistence can last many periods	Correlation dies out after one lag
Models inertia or state dependence	Models short-lived noise or adjustment

Both generate autocorrelation, but with very different persistence patterns.

From AR to ARMA Models

Observation: Pure AR models capture persistence, but sometimes shocks have short-run patterns not explained by past y_t alone.

Idea: Allow the error term itself to follow a moving-average (MA) process:

$$y_t = \mu + \rho_1 y_{t-1} + \cdots + \rho_p y_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}.$$

$\Rightarrow$ This is an **ARMA(p, q)** model.

Interpretation:

- AR part: captures state persistence ($y_{t-1}, y_{t-2}, \ldots$)
- MA part: captures shock persistence ($u_{t-1}, u_{t-2}, \ldots$)
- Together they describe dynamics using a compact combination of both effects.

Next: Estimating ARMA models and using them for multi-step forecasts.

The Partial Autocorrelation Function (PACF)

Motivation: After learning about ARMA models, we need tools to identify their order. The ACF alone cannot tell us whether persistence comes from AR or MA terms.

Definition: The PACF measures the correlation between y_t and y_{t-k} after controlling for all intermediate lags:

$$\phi_k = \text{corr}(y_t, y_{t-k} \mid y_{t-1}, \ldots, y_{t-k+1}).$$

Typical patterns:

- **AR(p)**: PACF **cuts off** after lag p, ACF decays geometrically.
- **MA(q)**: ACF **cuts off** after lag q, PACF decays gradually.
- **ARMA(p, q)**: both ACF and PACF decay.

Use: Compare sample ACF and PACF plots to identify candidate model orders.

ACF vs. PACF: Math and Intuition

Autocorrelation Function (ACF):

$$\rho_k = corr(y_t, y_{t-k})$$

Captures all correlation between y_t and y_{t-k}, including indirect paths through intermediate lags ($y_{t-k} \to y_{t-k+1} \to \cdots \to y_{t-1} \to y_t$).

Partial Autocorrelation Function (PACF):

$$\phi_k = corr\big(y_t, y_{t-k} \,\big|\, y_{t-1}, \ldots, y_{t-k+1}\big)$$

Removes those indirect paths; measures the net/direct k-step link.

Operational definitions:

- **ACF**: sample correlation of (y_t, y_{t-k}).
- **PACF**: coefficient on y_{t-k} in the OLS of y_t on $(y_{t-1}, \ldots, y_{t-k})$.

PACF: A Clean Computation View

Let P_{k-1} be the linear projection on $(y_{t-1}, \dots, y_{t-k+1})$. Define residuals

$$\tilde{y}_t \;=\; y_t - P_{k-1} y_t, \qquad \tilde{y}_{t-k} \;=\; y_{t-k} - P_{k-1} y_{t-k}.$$

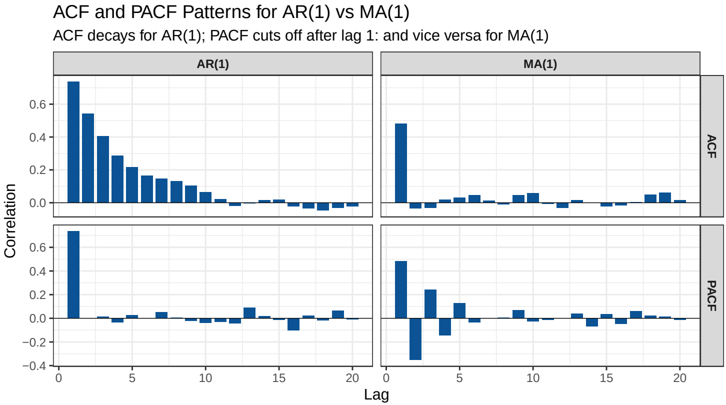
Then the **partial autocorrelation at lag** k is simply

$$\phi_k \;=\; corr(\tilde{y}_t, \tilde{y}_{t-k}),$$

i.e. the correlation of the residualized current value with the residualized k-lag.

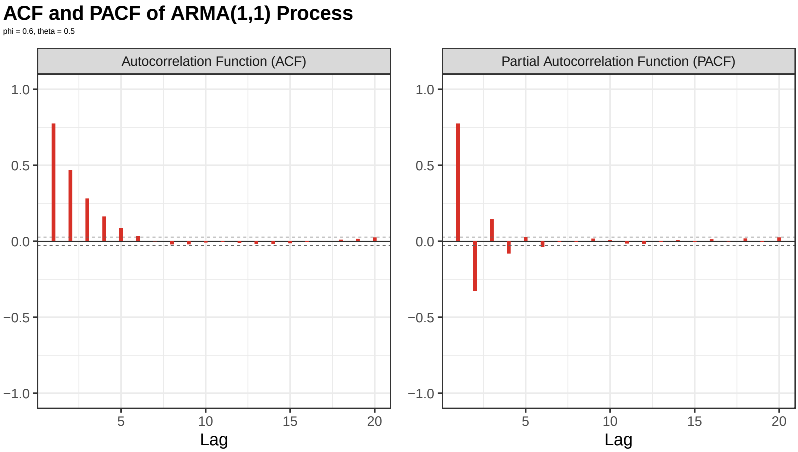
Intuition in one line:
ACF mixes direct and indirect persistence; PACF strips out the middlemen and keeps only the direct k-step link.

Illustration: ACF and PACF Patterns



- **AR(1):** ACF decays gradually; PACF cuts off after lag 1.
- **MA(1):** PACF decays gradually; ACF cuts off after lag 1.
- Both panels help identify whether persistence arises from autoregressive or moving-average terms.

ACF and PACF of ARMA(1,1)



- Both ACF and PACF decay gradually — neither cuts off sharply.
- This “tailing off” pattern is characteristic of mixed ARMA processes.

Differencing and ARIMA (One-Slide Summary)

Why difference? Make a nonstationary series stationary in mean.

Operators: $\Delta y_t \equiv y_t - y_{t-1} = (1-L)y_t, \qquad \Delta^d y_t = (1-L)^d y_t$
Seasonal differencing: $\Delta_s y_t \equiv y_t - y_{t-s} = (1-L^s)y_t, \quad \Delta_s^D \Delta^d y_t$ for seasonal & nonseasonal nonstationarity.

ARIMA model (backshift L):

$$\underbrace{\phi(L)}_{\text{AR}(p)}\ \underbrace{(1-L)^d(1-L^s)^D}_{\text{differencing}}\ y_t = c + \underbrace{\theta(L)}_{\text{MA}(q)}\ u_t, \qquad u_t \sim \text{i.i.d. } (0,\sigma^2)$$

Nonseasonal ARIMA(p,d,q); with seasonality:
ARIMA$(p,d,q) \times (P,D,Q)_s$.

9.6: Conditional Heteroskedasticity

Motivation for ARCH Models

Until now we modeled correlation in the **mean**. But economic and financial data often show dependence in the **variance**:

- Periods of calm alternate with periods of high volatility.
- Large shocks tend to cluster together.

Idea: Let conditional variance depend on past squared shocks:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2, \qquad \alpha_0 > 0,\ \alpha_1 \geq 0.$$

This is the **ARCH(1)** model (Autoregressive Conditional Heteroskedasticity)

The ARCH(1) Model

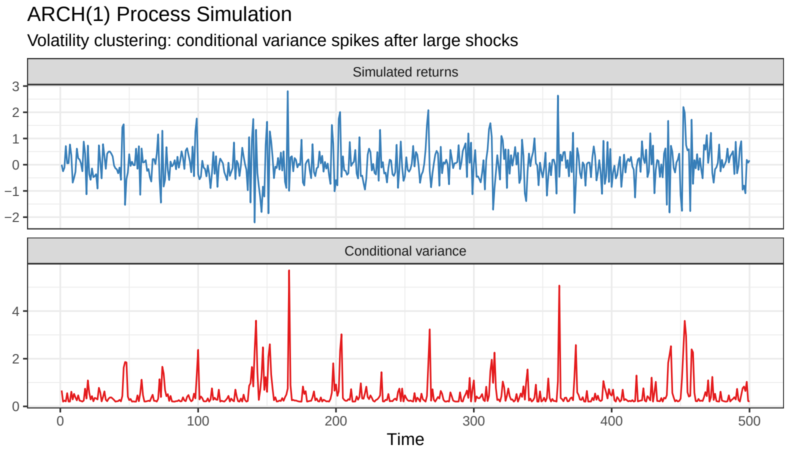
Model:

$$y_t = x_t'\beta + \varepsilon_t, \qquad \varepsilon_t = u_t, \quad u_t \sim N(0, \sigma_t^2),$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2.$$

Properties:

- $E[u_t] = 0$, $Var(u_t) = \frac{\alpha_0}{1-\alpha_1}$ if $\alpha_1 < 1$.
- Captures volatility clustering: high variance follows large shocks.
- OLS remains unbiased but inefficient; MLE is preferred.

Illustration: Volatility Clustering in ARCH(1)



Periods of high volatility follow large shocks, even though $E[u_t] = 0$.

ARCH-in-Mean and Summary

ARCH-in-Mean (ARCH-M):

$$y_t = \mu + \lambda\sigma_t^2 + u_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2.$$

Interpretation:

- Risk (variance) directly affects the conditional mean.
- Useful in finance (risk–return tradeoff) and macro uncertainty models.

Summary Table:

	Mean Equation	**Variance Equation**
ARMA	$y_t = \phi(L)y_t + \theta(L)u_t$	Constant variance
ARCH	Same mean	$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$
ARCH-M	$\mathbf{E}[y_t \vert \mathcal{F}_{t-1}]$ depends on σ_t^2	same as above

9.7: Vector Autoregression (VAR) Models

The VAR(p) Model

When multiple time series influence each other:

$$\mathbf{y}_t = c + A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_p\mathbf{y}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim (0, \Sigma_u)$$

where $\mathbf{y}_t$ is an $m \times 1$ vector of jointly determined variables.
Example:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + A_1 \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

Interpretation:

- Captures feedback and interdependence between variables.
- Each equation resembles an AR model but includes lags of all variables.

The VAR(p) Model

When multiple time series influence each other:

$$\mathbf{y}_t = c + A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_p\mathbf{y}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim (0, \Sigma_u)$$

where $\mathbf{y}_t$ is an $m \times 1$ vector of jointly determined variables.
Example:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + A_1 \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

Interpretation:

- Captures feedback and interdependence between variables.
- Each equation resembles an AR model but includes lags of all variables.

Impulse–Response Analysis

From the VAR(p) we can write a moving–average representation:

$$\mathbf{y}_t = \mu + \sum_{i=0}^{\infty} \Psi_i \mathbf{u}_{t-i}.$$

Impulse–Response Function (IRF):

$$\Psi_h = \frac{\partial \mathbf{y}_{t+h}}{\partial \mathbf{u}_t}.$$

Interpretation: How a one-unit shock to variable j today affects variable i in the next h periods.

Further Topics in Time-Series Econometrics

A Glimpse Beyond This Lecture

We will **not cover the following topics in detail**, but they are central to modern time-series econometrics and forecasting. Consider them as directions to explore if you are intersted in time-series.

- **Forecasting beyond ARMA:** ARIMA, exponential smoothing, and Kalman filtering; practical tools for prediction.
- **Structural change and model stability:** Chow, QLR, and Bai–Perron tests for detecting breaks in economic relationships.
- **Conditional heteroskedasticity:** ARCH and GARCH models for time-varying volatility in financial and macro data.

Forecasting Beyond ARMA: Practical Methods

Goal: Predict future values of y_t given past data and limited model structure.

Traditional econometric forecasting:

- **ARIMA models:** Combine differencing (I) with AR and MA terms to handle trends and persistence.
- **Exponential Smoothing (ETS):** Adaptive weights that give more importance to recent observations.
- **State-space and Kalman filtering:** Recursively update forecasts as new data arrive.

Modern forecasting:

- **Machine learning methods:** Random forests, gradient boosting: strong at short-horizon forecasting.
- **Combination forecasts:** Average multiple model forecasts; improves robustness to misspecification.
- **Rolling-origin evaluation:** Estimate model on $t \leq T$, forecast $t + 1$, roll forward, compute RMSE.

But: Forecasting is about **capturing persistence and adaptability**. Simple models often perform surprisingly well out-of-sample.

Detecting Structural Change and Breaks

Motivation: Economic relationships may shift due to policy, technology, or crises. A model estimated on past data may no longer describe today's world.

Tests for breaks:

- **Chow test:** Compares fit before and after a known break date. *Use when the date of change is suspected (e.g., policy reform).*
- **Quandt Likelihood Ratio (QLR) test:** Searches over possible breakpoints; identifies most likely break.
- **Bai-Perron test:** Allows multiple breaks at unknown dates; widely used in macro and finance.

Intuition:

- If model residuals or coefficients shift abruptly, the process may have new parameters.
- Structural breaks are about **parameter instability** not model failure.

Practical tip: Always check model stability when working with long time spans or major events (e.g., COVID, Euro introduction).

From ARCH to GARCH and Beyond

Why move beyond ARCH? ARCH(1) captures volatility clustering but often needs many lags to fit real data. **GARCH (Generalized ARCH)** makes this more parsimonious:

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Interpretation:

- Combines **shock persistence** (u_{t-1}^2) and **volatility persistence** (σ_{t-1}^2).
- Implies that volatility responds slowly to shocks; exactly what we observe in markets.

Common extensions:

- **EGARCH (Exponential GARCH):** captures asymmetry (“bad news increases volatility more”).
- **TGARCH (Threshold GARCH):** allows separate parameters for positive/negative shocks.
- **Multivariate GARCH:** tracks time-varying covariances across assets.