STA 542: Introduction to Time Series Analysis

Lecture 8: Choosing a transformation

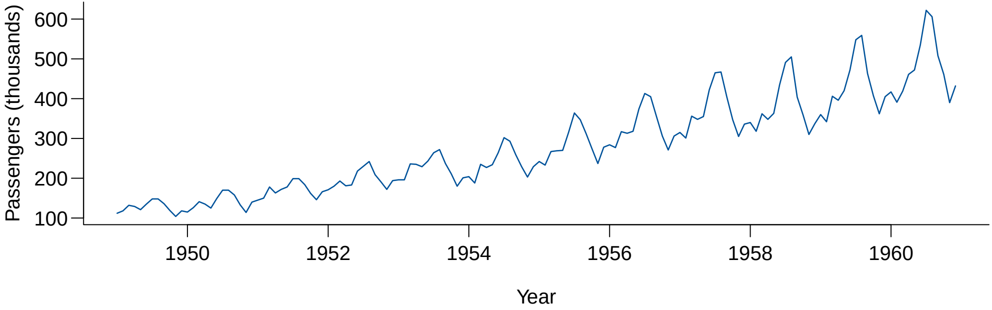
Lasse Vuursteen

Duke University

Mon Sep 21, 2026

What do we see in the passenger record?

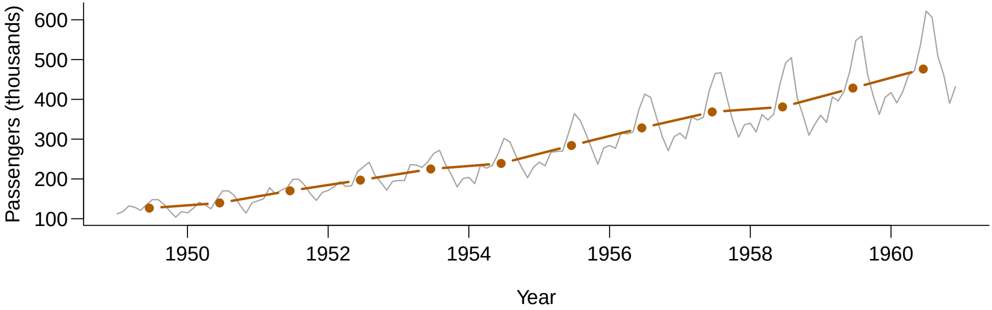
Monthly international airline passenger totals, 1949–1960.



Which features would a model need to describe?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

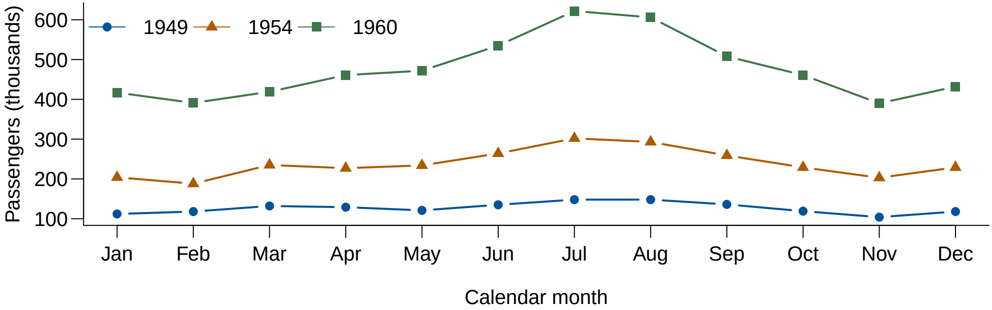
Passenger totals rise across years



The orange line joins the average monthly passenger count in each year. How should we describe this trend?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

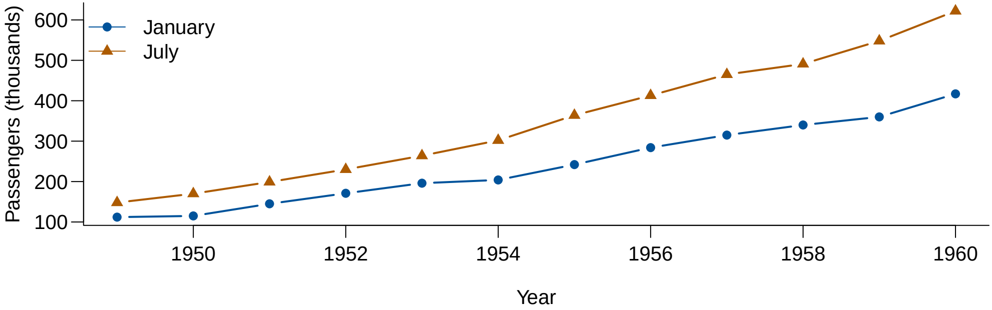
Compare the months within different years



The summer peak recurs. Its height above the winter months grows with the level.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Compare January with January, and July with July

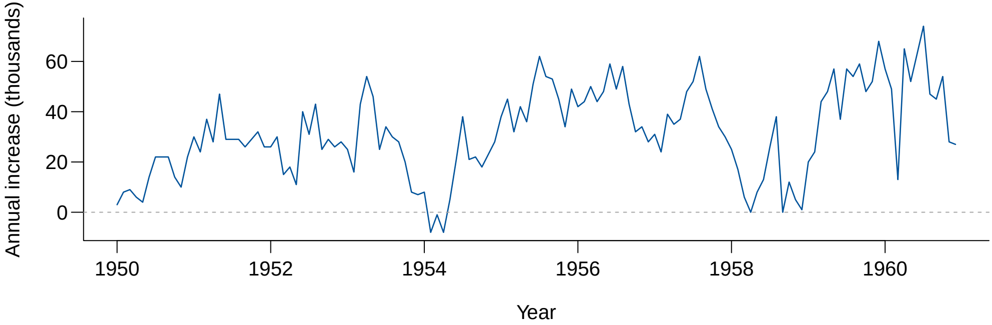


Even when we hold the month fixed, passenger totals rise. How much do they increase from one year to the next?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Measure growth by the increase in passenger numbers

Compare the same month in consecutive years: $X_t - X_{t-12}$.

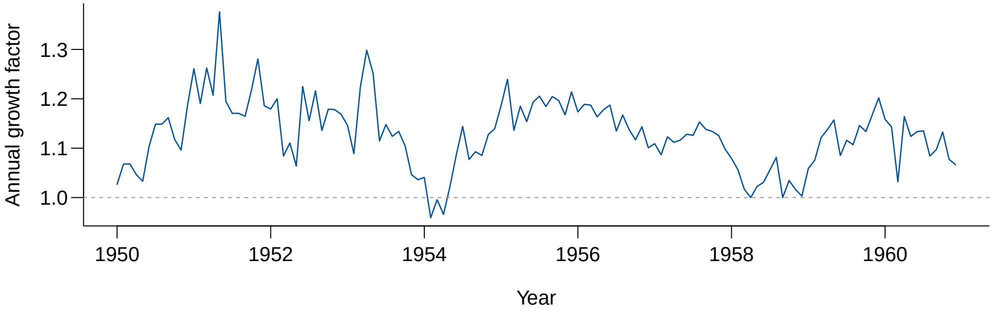


Do these annual increases fluctuate around a common level?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Or measure growth by a factor

The annual growth factor is $G_t := X_t / X_{t-12}$.



A factor of 1.1 means 10% growth, whatever the starting passenger count. Is a roughly constant growth factor plausible?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

What would constant 10% annual growth mean?

Suppose, as a simple model, that each month's total is 10% above the same month a year earlier:

$$X_t = 1.1X_{t-12}.$$

For a given calendar month, a starting value of 100 would become 110, then 121, then 133.1.

The absolute increases grow; the growth factor stays at 1.1.

On the log scale, the annual increase is constant

Under the hypothetical relation $X_t = 1.1X_{t-12}$, taking logs gives

$$\log X_t = \log(1.1) + \log X_{t-12}.$$

On the log scale, the annual increase is constant

Under the hypothetical relation $X_t = 1.1X_{t-12}$, taking logs gives

$$\log X_t = \log(1.1) + \log X_{t-12}.$$

Set $L_t := \log X_t$. Then $L_t - L_{t-12} = \log(1.1)$.

Each calendar month's log total increases by the same amount each year. This motivates trying a linear trend on the log scale.

Apply the same transformation at every time

Definition (Point transformation)

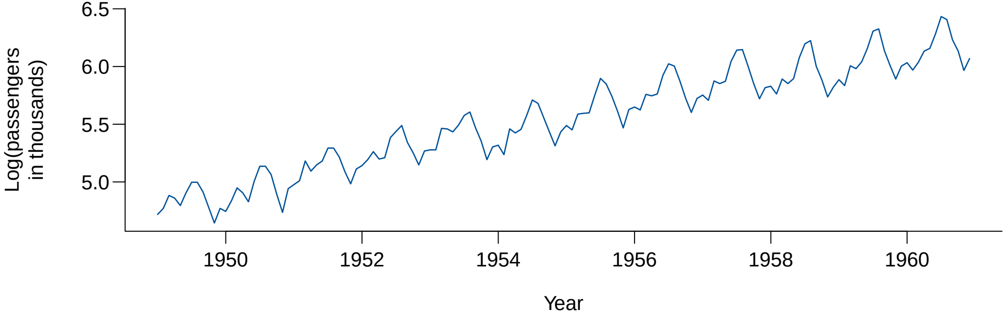
Let $(X_t)_{t\in\mathbb{Z}}$ be a process taking values in $D \subseteq \mathbb{R}$. A *point transformation* applies the same function $h : D \to \mathbb{R}$ at every time, giving

$$L_t = h(X_t), \qquad t \in \mathbb{Z}.$$

Each transformed value depends only on the observation at that time.

For positive passenger totals, $h(x) = \log x$ gives $L_t = \log X_t$.

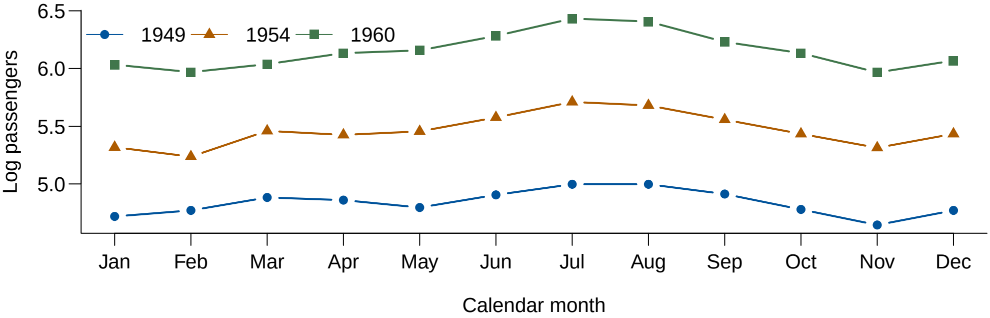
Now take logs of the observed passenger totals



The rising level looks approximately linear. The annual pattern remains.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

The seasonal swings are more comparable on the log scale



The annual patterns have more similar ranges. We still need to account for the rising level and the recurring seasonal mean.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Postulate a trend in log passenger totals

Let $L_t := \log X_t$, with t counting months. We propose

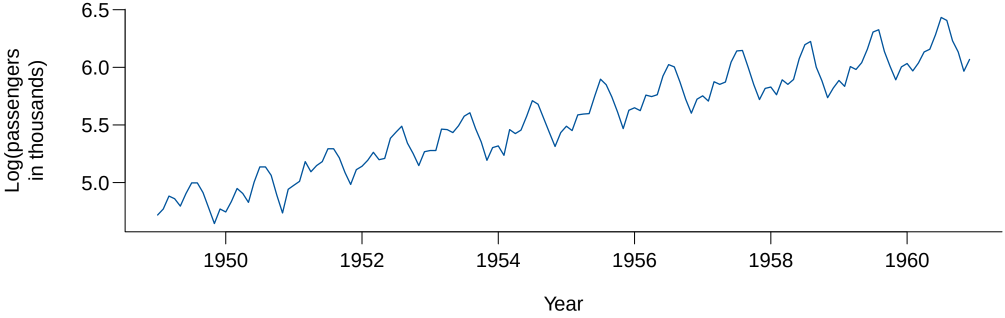
$$L_t = m(t) + Y_t.$$

- $m(t)$ is a deterministic function describing the trend.
- Y_t is what remains after subtracting that trend.

At this stage, Y_t still includes the seasonal pattern. We have not assumed that this remainder is stationary.

A linear trend is one possible choice

The log plot motivates $m(t) = a + bt$, giving $L_t = a + bt + Y_t$.



The intercept a and monthly slope b are unknown. We have chosen a form for the trend; we have not fitted it yet.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

A curved trend is another possible choice

We could instead propose

$$m(t) = a + bt + ct^2, \qquad L_t = a + bt + ct^2 + Y_t.$$

The coefficient c allows curvature in the log trend.

The known functions $1, t, t^2$ are the regressors. The unknown coefficients a, b, c enter linearly, so this is still a linear regression model.

What does the log-trend model say about passenger totals?

Starting from $L_t = m(t) + Y_t$ and $L_t = \log X_t$, exponentiation gives

$$X_t = \underbrace{e^{m(t)}}_{\text{trend component}} \times \underbrace{e^{Y_t}}_{\text{multiplicative remainder}}.$$

An additive remainder on the log scale becomes a multiplicative factor on the passenger scale.

What annual growth factor does this model imply?

The trend and the remainder both affect observed growth

Compare the same calendar month in consecutive years:

$$\frac{X_t}{X_{t-12}} = \underbrace{e^{m(t)-m(t-12)}}_{\text{growth factor of the trend;}} \times \underbrace{e^{Y_t - Y_{t-12}}}_{\text{change in the remainder}}.$$

A constant growth factor for the trend does not force the observed annual growth factors to be constant.

A fixed repeating seasonal effect cancels in $Y_t - Y_{t-12}$; other fluctuations can remain.

A linear log trend gives constant proportional growth

For $m(t) = a + bt$, the annual change in the log trend is

$$m(t) - m(t-12) = (a + bt) - \{a + b(t-12)\} = 12b.$$

The trend's annual growth factor is therefore e^{12b}.

A linear log trend gives constant proportional growth

For $m(t) = a + bt$, the annual change in the log trend is

$$m(t) - m(t-12) = (a + bt) - \{a + b(t-12)\} = 12b.$$

The trend's annual growth factor is therefore e^{12b}.

Our 10% example corresponds to $e^{12b} = 1.1$, or $b = \log(1.1)/12$. The observed growth factor also contains $e^{Y_t - Y_{t-12}}$.

A quadratic log trend allows the growth rate to change

For $m(t) = a + bt + ct^2$,

$$\begin{aligned} m(t) - m(t-12) &= 12b + c\{t^2 - (t-12)^2\} \\ &= 12b + 24ct - 144c. \end{aligned}$$

The trend's annual growth factor is $e^{12b+24ct-144c}$.

It increases with time if $c > 0$ and decreases if $c < 0$. Setting $c = 0$ returns the constant-growth model.

We have chosen a family of trends. Which one should we use?

Return to the linear model

$$L_t = a + bt + Y_t.$$

Each choice of a and b gives a different trend $m(t) = a + bt$. We have chosen the form of the trend, but its coefficients are unknown.

We observe the log passenger totals

$$\ell_1, \ldots, \ell_{144}, \qquad \ell_t := \log x_t.$$

How can we use these observations to choose a and b?

A wrong slope leaves a changing mean

Suppose $L_t = a + bt + Y_t$, with (Y_t) weakly stationary and centered. Subtract a fixed candidate line $u + vt$:

$$L_t - u - vt = (a - u) + (b - v)t + Y_t.$$

Which coefficient has to be correct for the result to be stationary?

A wrong slope leaves a changing mean

Suppose $L_t = a + bt + Y_t$, with (Y_t) weakly stationary and centered. Subtract a fixed candidate line $u + vt$:

$$L_t - u - vt = (a - u) + (b - v)t + Y_t.$$

Which coefficient has to be correct for the result to be stationary?

We need $v = b$. An incorrect intercept changes only the constant mean.

Fitted coefficients are random. Their residuals estimate Y_t; they need not have exactly stationary covariance within a finite record.

Why minimize squares when the fluctuations are correlated?

Suppose $L_t = m(t) + Y_t$, where (Y_t) is weakly stationary with $\mathbb{E}[Y_t] = 0$ and variance $\gamma_Y(0)$. For any fixed candidate mean g,

$$\mathbb{E}[\{L_t - g(t)\}^2] = \{m(t) - g(t)\}^2 + 2\{m(t) - g(t)\}\underbrace{\mathbb{E}[Y_t]}_{0} + \mathbb{E}[Y_t^2]$$

$$= \{m(t) - g(t)\}^2 + \gamma_Y(0).$$

Sum over t. The actual mean minimizes the expected loss if it belongs to the proposed class.

Why minimize squares when the fluctuations are correlated?

Suppose $L_t = m(t) + Y_t$, where (Y_t) is weakly stationary with $\mathbb{E}[Y_t] = 0$ and variance $\gamma_Y(0)$. For any fixed candidate mean g,

$$\mathbb{E}[\{L_t - g(t)\}^2] = \{m(t) - g(t)\}^2 + 2\{m(t) - g(t)\} \underbrace{\mathbb{E}[Y_t]}_{0} + \mathbb{E}[Y_t^2]$$

$$= \{m(t) - g(t)\}^2 + \gamma_Y(0).$$

Sum over t. The actual mean minimizes the expected loss if it belongs to the proposed class.

This identifies a target. Accuracy of the fitted coefficients still depends on the observations' dependence and the chosen regressors.

Fit a trend, then examine what remains

We suspect that part of the nonstationarity comes from a changing mean.

Fit a trend, then examine what remains

We suspect that part of the nonstationarity comes from a changing mean.

If this description is appropriate and the fluctuations average out sufficiently, least squares can estimate the line.

Fit a trend, then examine what remains

We suspect that part of the nonstationarity comes from a changing mean.

If this description is appropriate and the fluctuations average out sufficiently, least squares can estimate the line.

We do not yet know whether subtracting it will leave a stationary process. Fitting gives us residuals with which to investigate that question.

Minimize the observed squares over a parametrized mean

Restrict the candidate to $m_\beta(t) = \sum_{j=0}^{p} \beta_j r_j(t)$, with known regressors r_j. For observations $L_1, \ldots, L_T$, ordinary least squares chooses

$$\widehat{\beta}_T := \arg\min_{\beta} \sum_{t=1}^{T} \{L_t - m_\beta(t)\}^2.$$

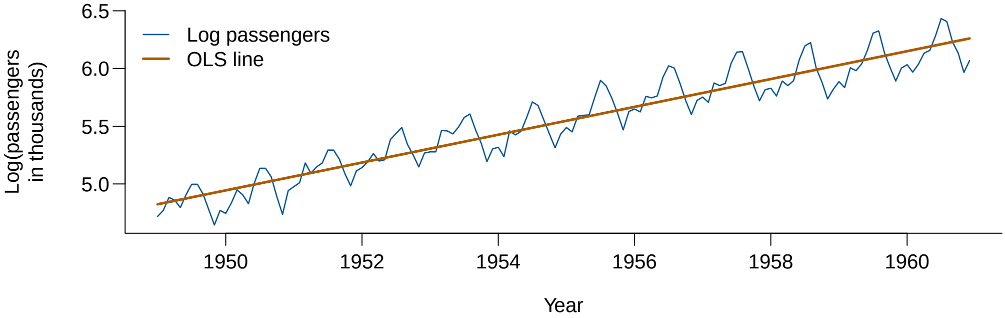
This is the empirical counterpart of the expected-loss identity.

Estimate the rising level by a fitted line

Return to the linear candidate $L_t = a + bt + Y_t$. Here $m_\beta(t) = a + bt$, so OLS is

$$(\hat{a}, \hat{b}) := \arg\min_{a,b} \sum_{t=1}^{144} (\ell_t - a - bt)^2.$$

The fitted line follows the rise in log passengers

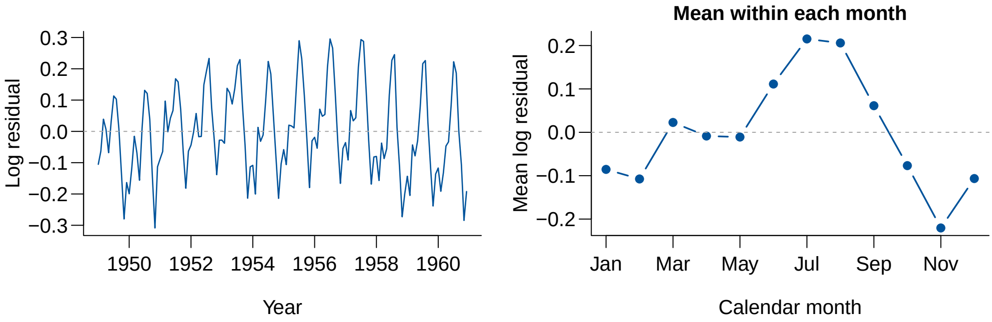


The fit estimates the changing level. We now inspect its residuals.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Subtract the fitted line: what remains?

The residual is $e_t := \ell_t - \hat{a} - \hat{b}t$.



The calendar-month means still show a recurring annual pattern.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Fit the annual pattern together with the trend

A smooth seasonal mean uses known sine and cosine regressors:

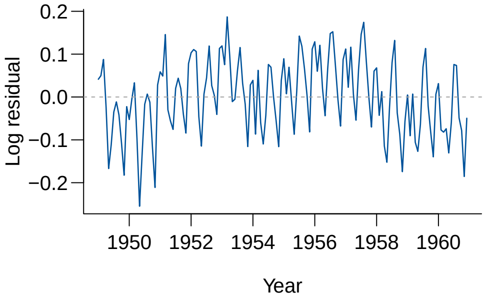
$$m(t) = a + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12).$$

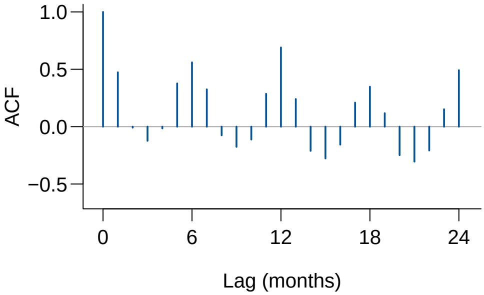
- The period is twelve months; a, b, c, d are estimated by OLS.
- Both sine and cosine allow the annual peak to occur at an unknown phase.
- Fit all four coefficients jointly, then subtract $\widehat{m}(t)$.

Adding t^2 or t^3 works the same way: known regressors, unknown coefficients entering linearly.

After subtracting the fitted trend and annual harmonic

Plot the residuals $e_t := \ell_t - \widehat{m}(t)$ and their sample ACF.





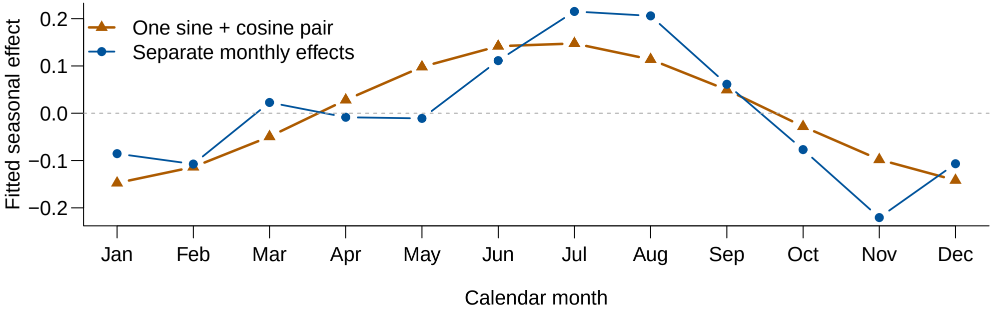
The sample ACF still has peaks at 6 and 12 months.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

ACF lags count months; sample ACVF uses divisor T. Reference bands are omitted throughout.

How much seasonal shape should the mean allow?

Fit a repeating monthly mean $m(t) = a + bt + c_{j(t)}$, where $j(t)$ is calendar month. Set $c_1 = 0$ and fit the trend and eleven monthly indicators jointly.

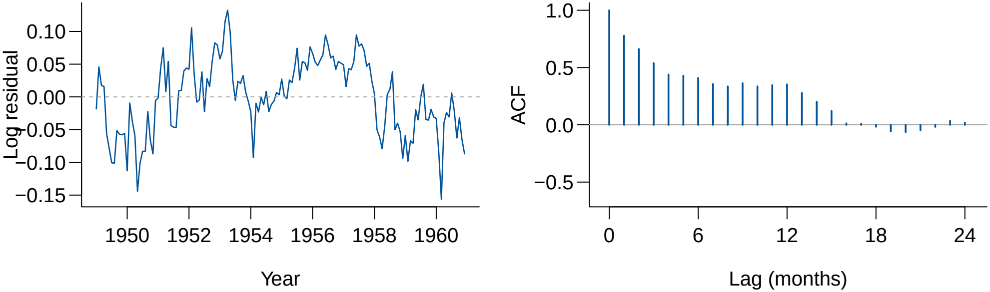


The plotted effects have been centered to have annual mean zero.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

After subtracting the fitted trend and monthly effects

These residuals measure deviations from the fitted mean log passenger count.



Could these be stationary dependence?

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Use the backshift operator to refer to earlier observations

Instead of fitting a mean, we can compare observations across time.

Definition (Backshift operator)

For a process $(X_t)_{t\in\mathbb{Z}}$, the *backshift operator* B and its powers are defined by

$$BX_t := X_{t-1}, \qquad B^k X_t := X_{t-k}, \qquad k = 0, 1, 2, \ldots.$$

The operator $B^0 = 1$ is the identity: $B^0 X_t = X_t$.

For monthly log passengers L_t, $B^{12} L_t = L_{t-12}$ is the value for the same month one year earlier.

First differences compare consecutive observations

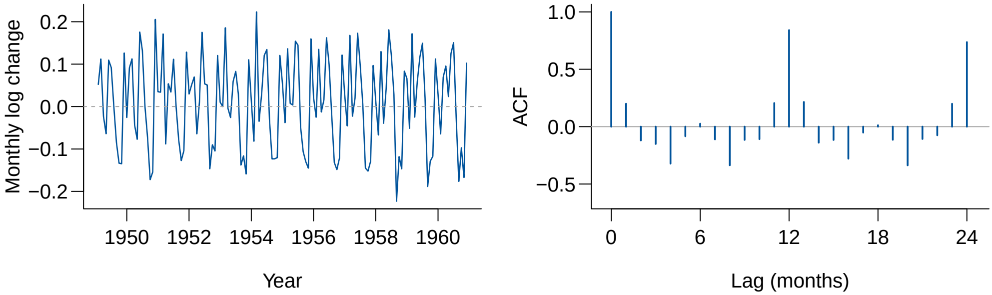
Definition (First-difference operator)

The *first-difference operator* is $\nabla := 1 - B$, where 1 denotes the identity operator. Applied to (X_t), it gives

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}.$$

For monthly log passengers, $\nabla L_t = L_t - L_{t-1}$ compares neighboring months.

$\nabla L_t = \log(X_t/X_{t-1})$ measures month-to-month log growth.



The ACF still has peaks at annual lags.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Seasonal differences compare observations one cycle apart

Definition (Seasonal difference)

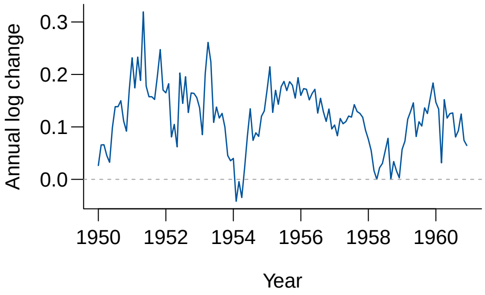
If one seasonal cycle contains s observations, where s is a positive integer, the *seasonal-difference filter* is

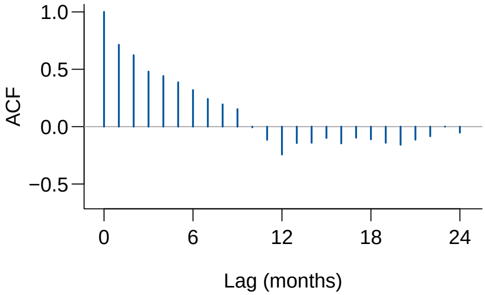
$$\nabla_s := 1 - B^s, \qquad \nabla_s X_t = X_t - X_{t-s}.$$

For monthly passengers with an annual cycle, $s = 12$. We compare the same calendar month in consecutive years.

Compare each month with the same month a year earlier

$W_t := \nabla_{12} L_t = \log(X_t / X_{t-12})$ measures annual log growth.





Growth has persistent stretches.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

We can also model changes in annual log growth

Take the monthly change in $W_t = L_t - L_{t-12}$:

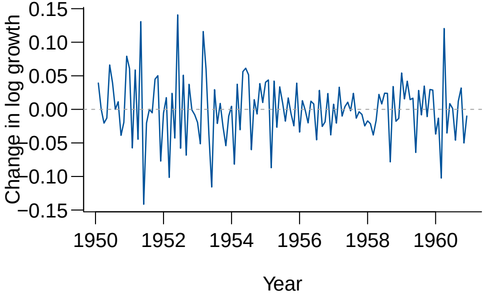
$$\begin{aligned} V_t := \nabla W_t &= (1 - B)(1 - B^{12})L_t \\ &= L_t - L_{t-1} - L_{t-12} + L_{t-13}. \end{aligned}$$

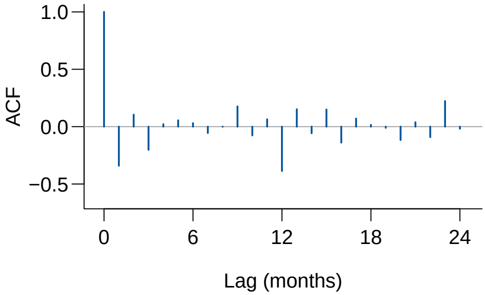
This compares this month's log growth with the same month's log growth a year earlier.

It is another candidate; the previous ACF does not establish that this extra difference is necessary.

After both differences, short and annual lags remain

$$V_t = (1 - B)(1 - B^{12}) \log X_t.$$

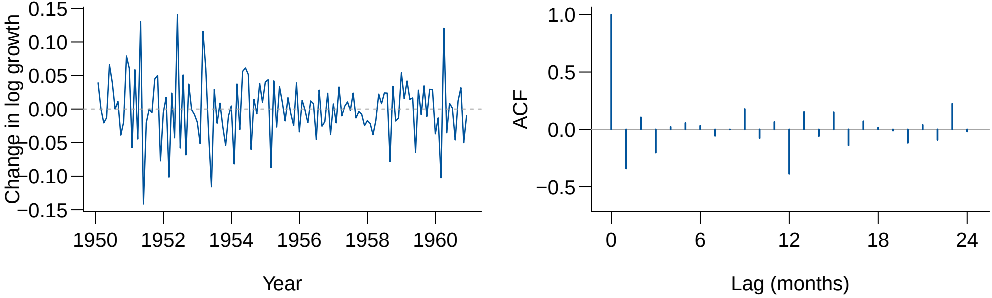




Are nonzero correlations a reason to keep differencing?

After both differences, short and annual lags remain

$$V_t = (1 - B)(1 - B^{12}) \log X_t.$$



Are nonzero correlations a reason to keep differencing?

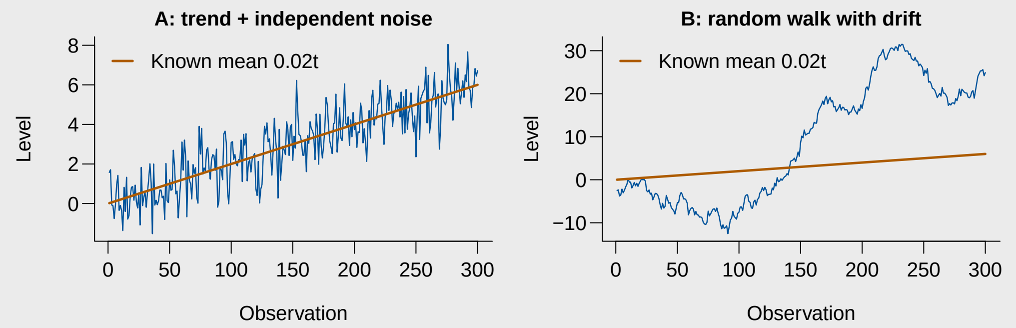
A stationary process can be correlated. These lags may need a dependence model.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

When does subtracting a trend work?

$X_t^{(D)} = bt + Z_t$ and $X_t^{(S)} = bt + \sum_{j=1}^{t} U_j$, for $t \geq 1$.

Independent i.i.d. $N(0,1)$ sequences $(Z_t), (U_t)$; $b = 0.02$, $X_0^{(S)} = 0$.

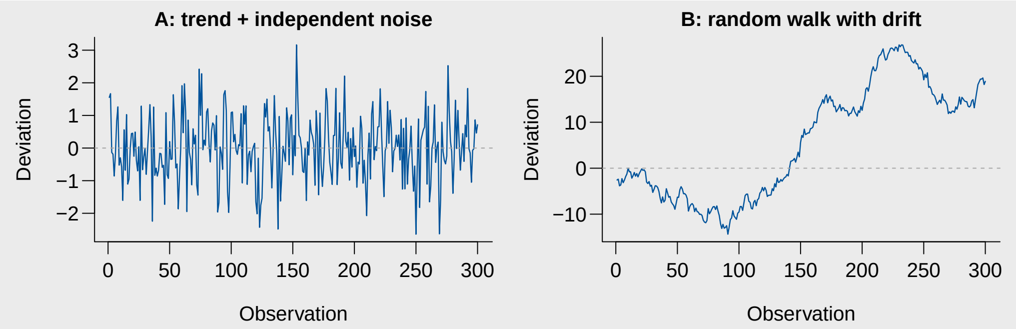


Both means equal bt. In the random walk, each shock affects all later levels.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

Subtract the true mean from both records

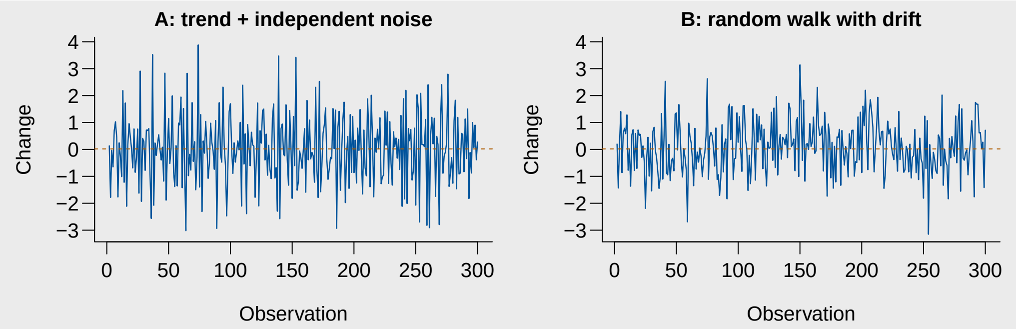
$X_t^{(D)} - bt = Z_t$, whereas $X_t^{(S)} - bt = \sum_{j=1}^{t} U_j$.



The variances are 1 and t, respectively. The random-walk remainder is nonstationary even when b is known.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

$\nabla X_t^{(D)} = b + Z_t - Z_{t-1}$, whereas $\nabla X_t^{(S)} = b + U_t$.



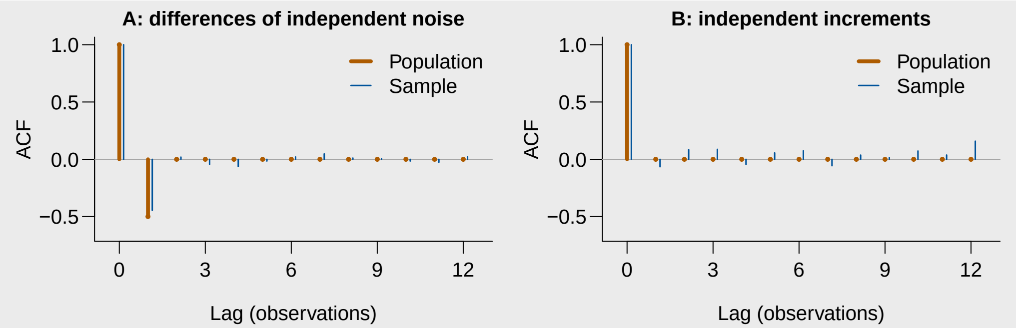
One output contains changes in noise; the other contains the original increments.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

Differencing stationary noise creates dependence

Here $D_t := Z_t - Z_{t-1}$ has variance 2. Its lag-one covariance is

$$\mathrm{Cov}(D_t, D_{t+1}) = \mathrm{Cov}(Z_t - Z_{t-1}, Z_{t+1} - Z_t) = -1.$$

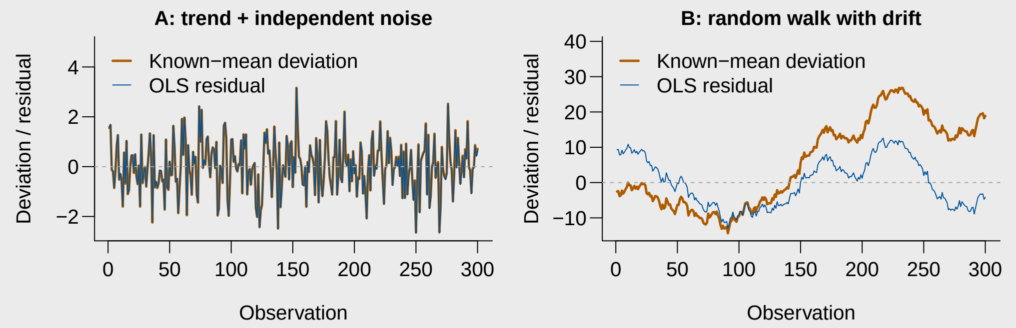


Consecutive changes share Z_t with opposite signs: $\rho_D(1) = -1/2$.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

Fitting a line does not resolve the distinction

Fit a separate line to each of the same two records and subtract it.



A fit can absorb part of a random walk's wandering. A zero residual average does not establish stationarity.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

A linear filter combines lagged observations

Definition (Finite linear filter)

Let $q \geq 0$ be an integer and $a_0, \ldots, a_q$ fixed real coefficients. A *finite linear filter* produces

$$a(B) := \sum_{j=0}^{q} a_j B^j, \qquad Y_t := a(B)X_t = \sum_{j=0}^{q} a_j X_{t-j}.$$

The same coefficients are applied at every time.

First differencing has $q = 1$, $a_0 = 1$, and $a_1 = -1$.

Which quantity do we want to retain?

Suppose $X_t = a + bt + Y_t$, with (Y_t) weakly stationary and centered.

$$
\begin{aligned}
X_t - a - bt &= Y_t && \text{deviation from the expected level,} \\
\nabla X_t &= b + Y_t - Y_{t-1} && \text{change between successive observations.}
\end{aligned}
$$

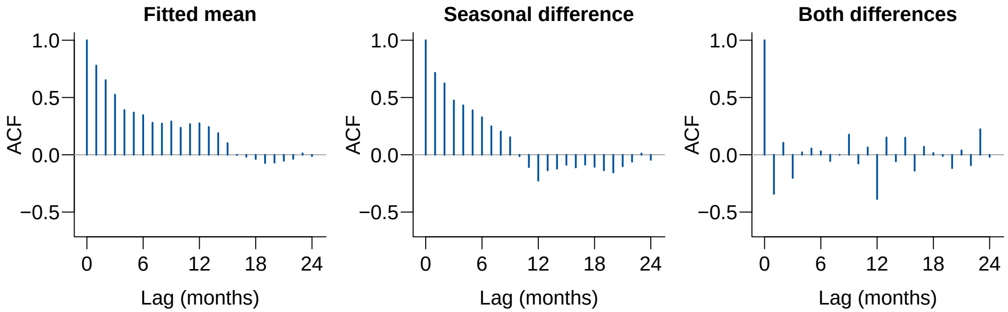
Both are stationary; their covariance structures and interpretations differ.

More generally, a finite linear filter applied to a weakly stationary process preserves weak stationarity.

Workbook: *How Does Filtering Change Stationary Dependence? — moment calculation*

Return to the airline candidates on the same dates

Compare February 1950–December 1960: 131 observations for each output.

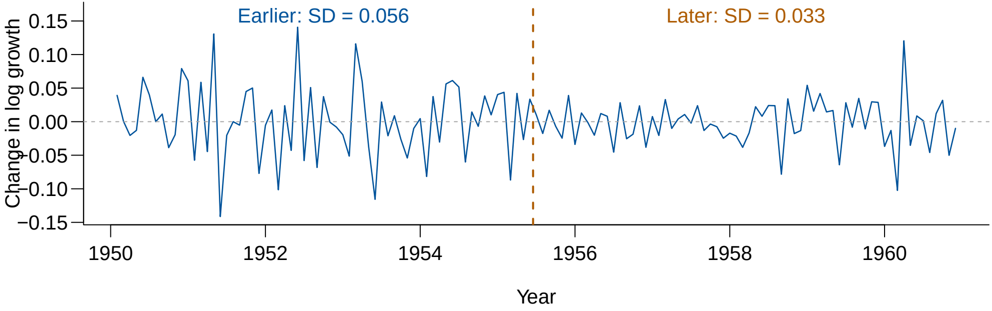


The combined differences leave a simpler lag pattern to model. This does not prove that the other candidates are nonstationary.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

Use combined log differences as our working choice

Model $V_t = (1-B)(1-B^{12})\log X_t$: changes in annual log growth.



The later stretch is quieter. A common variance remains a concern.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

The portions contain 65 and 66 months; differences in their summaries also include sampling variation.

Can a transformation be undone?

The records $(2, 5, 4)$ and $(12, 15, 14)$ both give first differences $(3, -1)$. The differences alone do not tell us which record we started with.

Definition (Inverse transformations)

An *inverse transformation* is a rule that recovers every original series from its transformed series. A transformation is *invertible* if an inverse exists.

If distinct original series produce the same transformed series, the transformation is not invertible.

Can we recover the original passenger totals?

- Logs: exponentiation recovers each positive passenger total.
- Fitted-mean residuals: reconstruction also needs the stored fitted mean.
- First differences: reconstruction also needs one initial level.
- Combined airline differences: retain the first thirteen log levels.

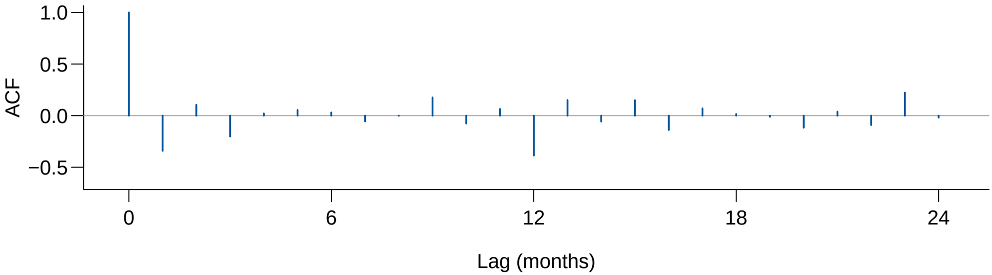
$$x_t = x_1 + \sum_{j=2}^{t} (x_j - x_{j-1}).$$

Changes alone cannot reveal the original level of passenger traffic.

Workbook: *Can We Recover the Original Observations?*

What structure is left to explain?

The chosen airline output still has correlations at short and annual lags.

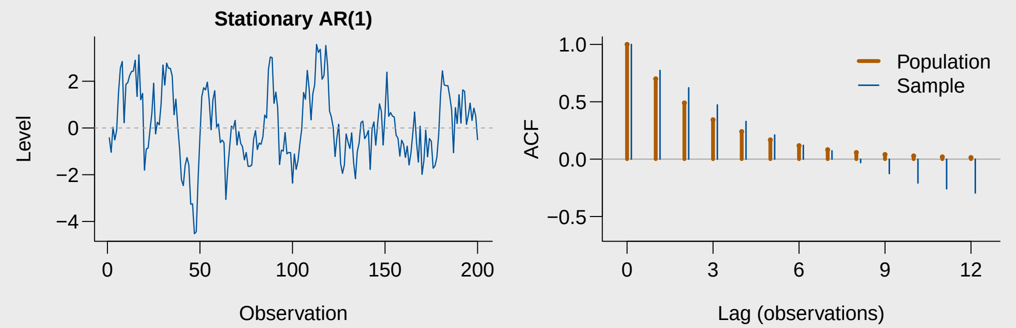


A dependence model describes how fluctuations persist or share shocks. It can improve prediction and quantify uncertainty.

Real data: `datasets::AirPassengers`, monthly totals in thousands, 1949–1960.

An AR model describes persistence of deviations

Stationary causal AR(1): $Y_t = 0.7Y_{t-1} + Z_t,\ Z_t \overset{\text{iid}}{\sim} N(0,1)$.



A shock persists with weights $1, 0.7, 0.7^2, \ldots$; $\rho(h) = 0.7^{|h|}$.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

Further filtering targets the innovations

For the stationary AR(1) $Y_t = \phi Y_{t-1} + Z_t$, define, for fixed ψ,

$$R_t(\psi) := Y_t - \psi Y_{t-1} = Z_t + (\phi - \psi) Y_{t-1}.$$

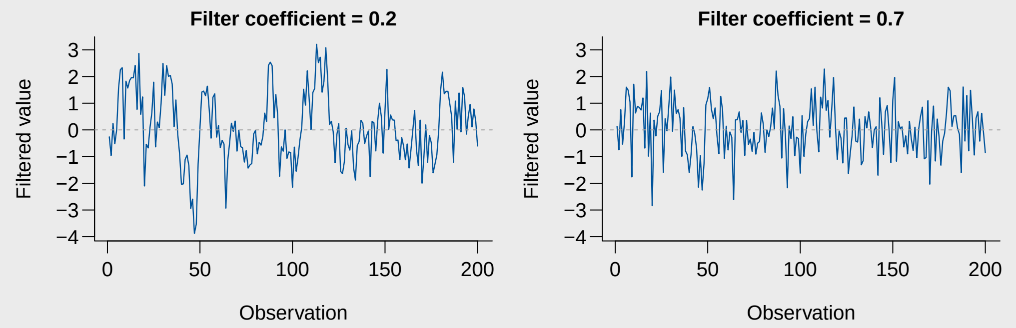
- Every fixed ψ leaves a weakly stationary process: this is a finite linear filter.
- Choosing $\psi = \phi$ recovers white noise.

With a trend, the correct slope removes a changing mean. Here the correct coefficient removes the remaining linear dependence.

Workbook: *How Does Filtering Change Stationary Dependence?*

After filtering the stationary AR(1)

Apply $R_t(\psi) = Y_t - \psi Y_{t-1}$ to the same stationary AR(1), $\phi = 0.7$.

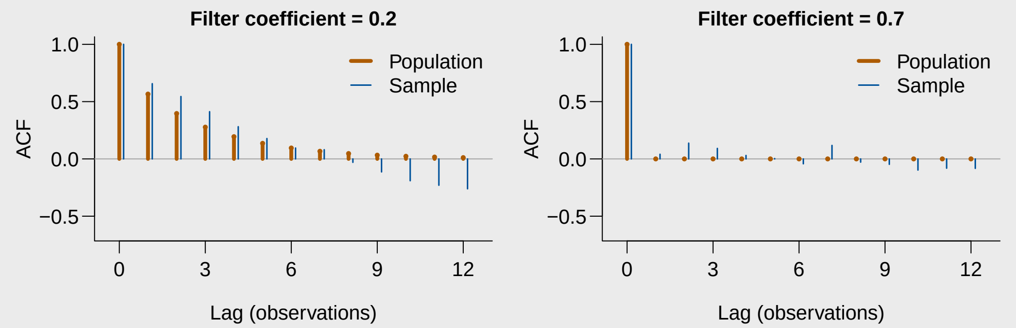


Both outputs are stationary. Which one retains dependence?

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

Does the residual filter recover white noise?

Apply $R_t(\psi) = Y_t - \psi Y_{t-1}$ to the same stationary AR(1), $\phi = 0.7$.



We would estimate ϕ from observations. Estimated residuals still need to be checked.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

What do we gain by estimating the remaining dependence?

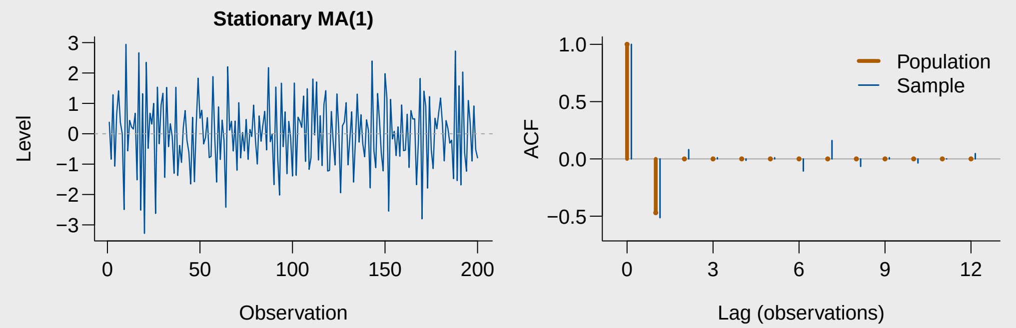
In the stationary causal AR(1) with $|\phi| < 1$ and i.i.d. centered innovations of variance σ^2, the best mean-square one-step predictor is

$$\mathbb{E}[Y_{T+1} \mid Y_1, \ldots, Y_T] = \phi Y_T.$$

- Its mean-square error is σ^2; predicting zero has error $\gamma(0) = \sigma^2/(1-\phi^2)$.
- For uncertainty about the mean, the same model gives $\tau^2 = \sigma^2/(1-\phi)^2$.
- Residuals help check whether the assumed dependence was captured.

An MA model describes observations that share shocks

Let $Y_t = Z_t - 0.7 Z_{t-1}$, with $Z_t \overset{\text{iid}}{\sim} N(0,1)$.



Neighbors share a shock with opposite signs; observations two steps apart share none.

Simulated data; Gaussian innovations. Reproducible code in `figures.R`.

For the airline series, shared shocks may also span a year

Write $\mu_V := \mathbb{E}[V_t]$ for the proposed constant mean and $Y_t := V_t - \mu_V$. One possible model is

$$\begin{aligned} Y_t &= (1+\theta B)(1+\eta B^{12})Z_t \\ &= Z_t + \theta Z_{t-1} + \eta Z_{t-12} + \theta\eta Z_{t-13}, \end{aligned}$$

where $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$.

- Shared shocks permit covariance at lags $1, 11, 12, 13$.
- The coefficients control its size and sign; they would need to be estimated.
- This is a candidate to assess, not a conclusion from the sample ACF.

ARMA models combine lagged observations and lagged shocks.

How convincing is the stationary approximation?

For the airline candidate, we still need to assess whether a common mean, variance, and fixed-lag covariance reasonably describe the record.

- Compare different portions of the transformed series.
- Distinguish remaining dependence from a change in its structure over time.
- Identify the hypothesis before using a stationarity or unit-root test.

Workbook 9: assessing a stationary approximation.