Regression Methods

ECON 5753 — University of Arkansas

Prof. Kyle Butts

April 2025

Introduction to Time-Series

Learning from Time-Series

Time-series Statistics

Time-series

Time-series data is a set of observations y_t that occur for a single unit measured over the course of time

$\rightarrow$ In general, we call t the 'period'

$\rightarrow$ If the time-series is spaced evenly over time without missing, it is called regular.

- Some methods require regular time intervals, and will do weird things if you give it irregular time series

$\rightarrow$ If you observe many units' time-series, this is called panel data

- Panel data that is regular is called a balanced panel

Examples

Examples include:

→ Annual data on the DGP of a country

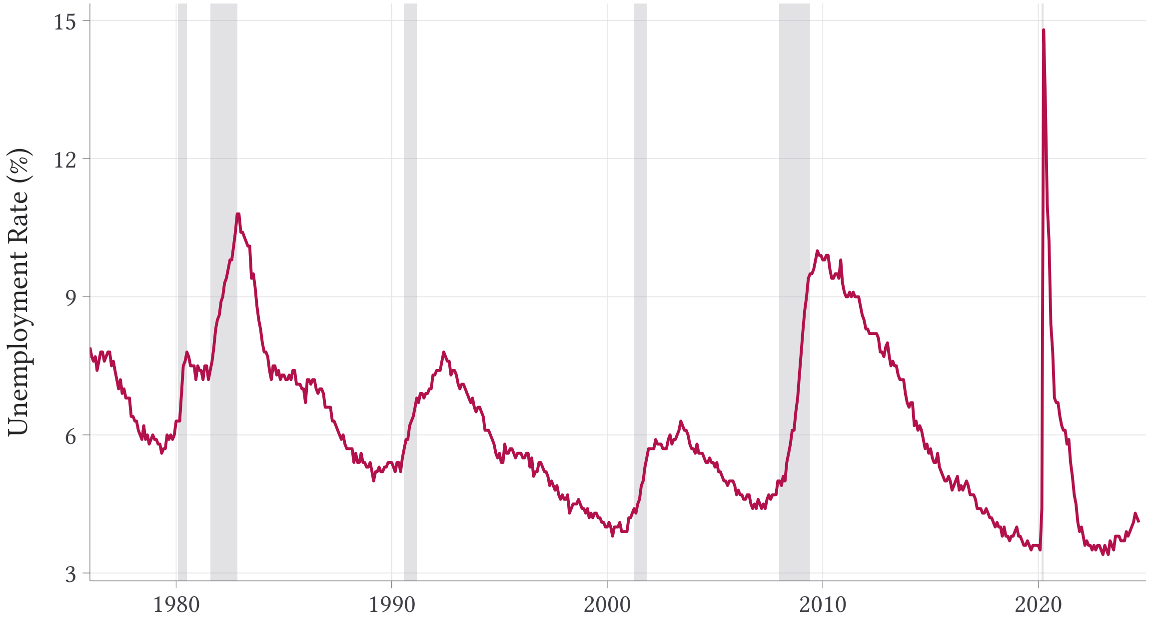
→ Hourly stock price for a company

→ Annual data on cigarette consumption per capita in a state

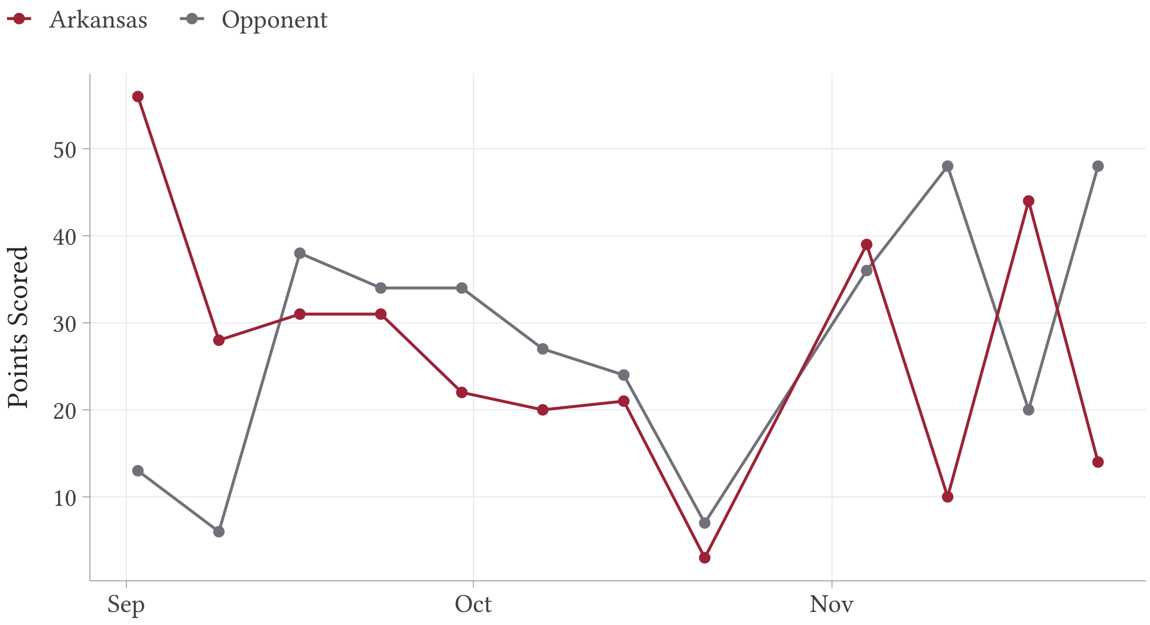
→ A sport's teams number of points scored in games (unequally spaced)

For each example, think through:

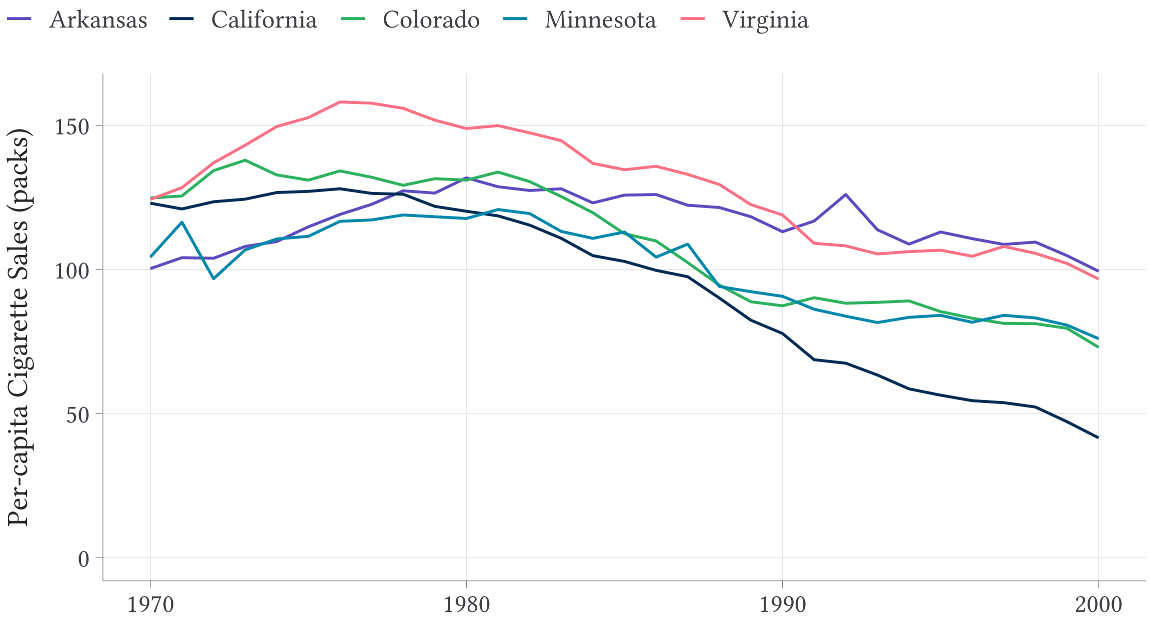
1. Is this a regular time-series?
2. Is this a panel dataset?



Unemployment Rate (%)
15
12
9
6
3
1980
1990
2000
2010
2020


Arkansas
Opponent
Points Scored
50
40
30
20
10
Sep
Oct
Nov


Apple Stock Price ($)
0
50
100
150
200
2010
2015
2020
2025


Arkansas
California
Colorado
Minnesota
Virginia
Per-capita Cigarette Sales (packs)
150
100
50
0
1970
1980
1990
2000

What is special about time-series?

In our previous topics, we have been thinking about cross-sectional data

→ Each person in your dataset is an independent draw that provides us with unique information

In cross-sectional data, knowing about one indivdiual does not really tell me much information about another

→ This is not *entirely true*; e.g. worker's in same firm have common experiences, kids in same school have same teacher quality, etc. (hence why we might cluster our standard errors)

What is special about time-series?

In time-series data, knowing last period's value of y_{t-1} is often very useful for this period's value of y_t

- $\rightarrow$ This property is essential in forecasting; following a variable over time might let us predict future values
- $\rightarrow$ Shocks that happened last period probably still impact me today(!)

What is special about time-series?

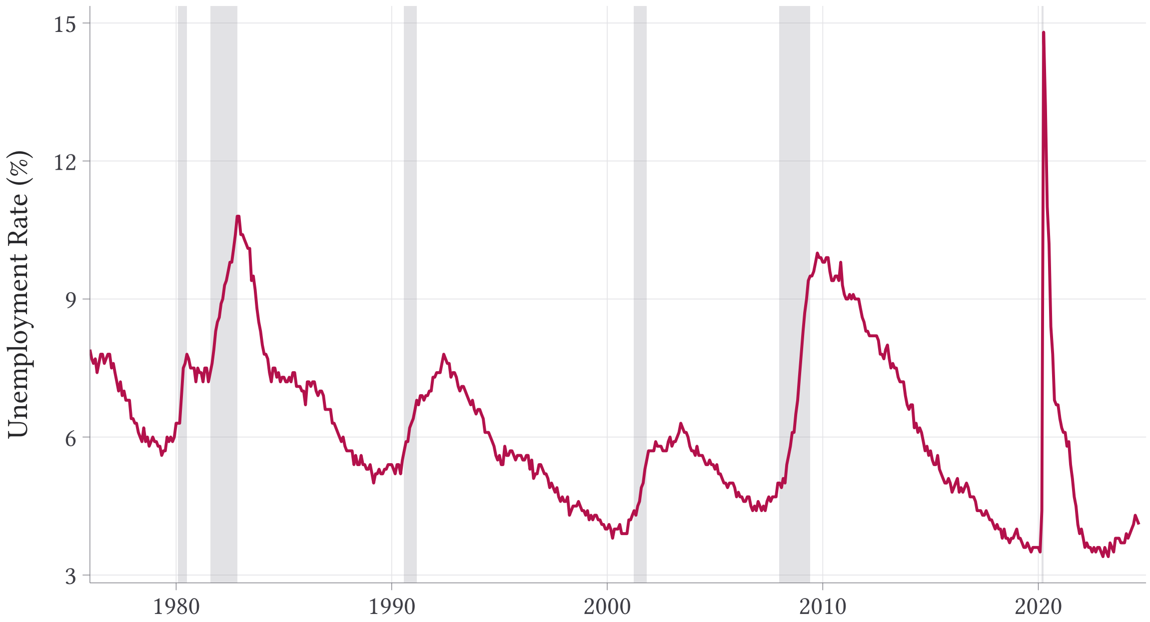
In time-series data, knowing last period's value of y_{t-1} is often very useful for this period's value of y_t

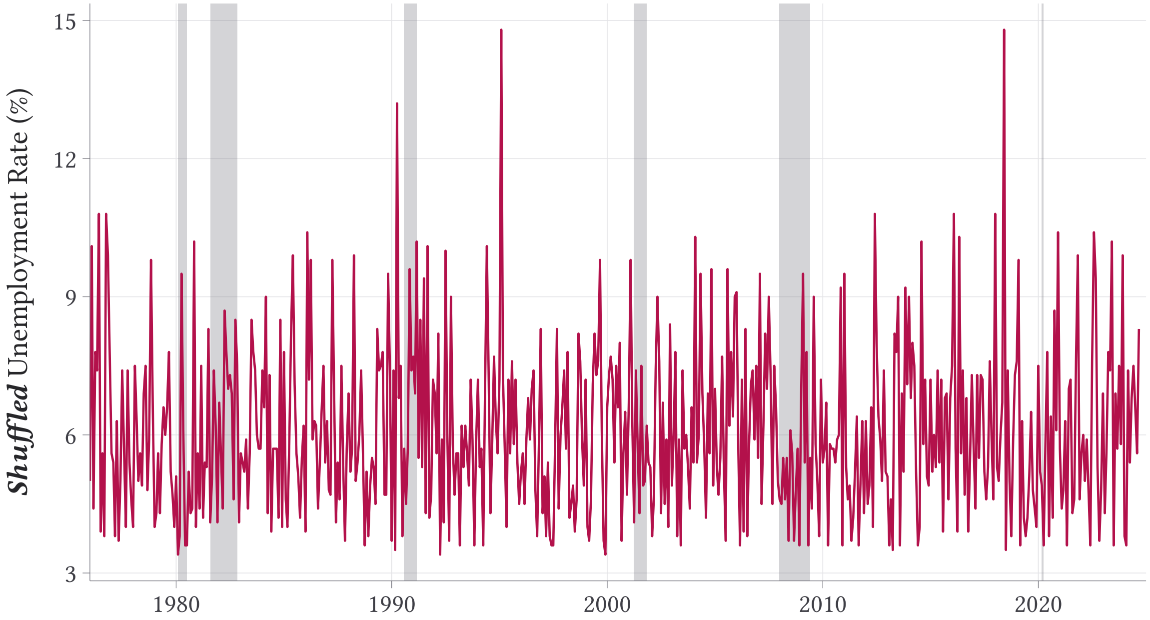
$\rightarrow$ This property is essential in forecasting; following a variable over time might let us predict future values

$\rightarrow$ Shocks that happened last period probably still impact me today(!)

Another way of saying this, is if we randomly shuffled time-series data, we would lose information!

$\rightarrow$ This is not true of a cross-sectional dataset; we can reshuffle rows without problem


Unemployment Rate (%)
15
12
9
6
3
1980
1990
2000
2010
2020



Shuffled Unemployment Rate (%)
15
12
9
6
3
1980
1990
2000
2010
2020

Thinking about inference with time-series

We have a bit of a problem with our time-series:

$$y_1, y_2, \ldots, y_T$$

→ y_1 is related to y_2

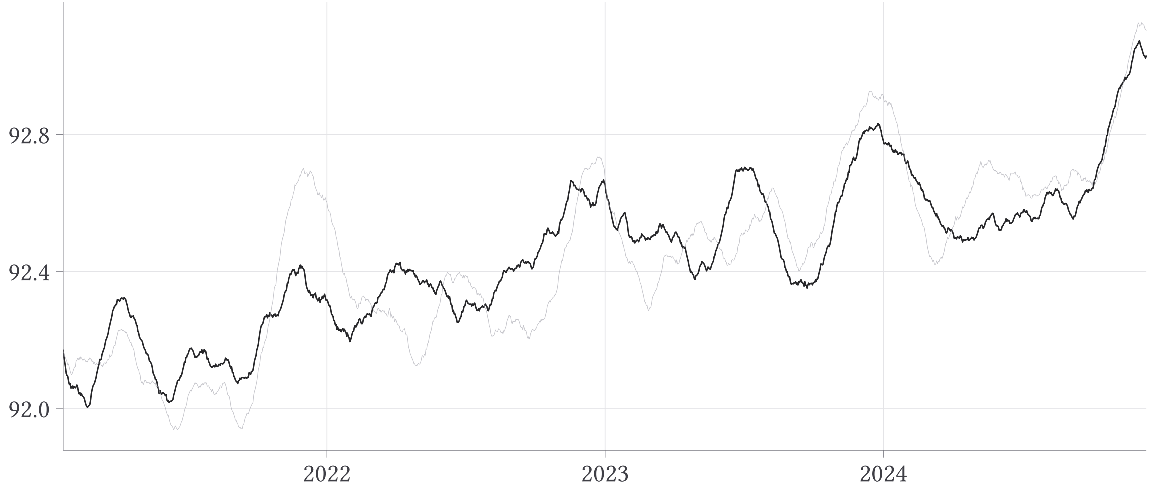
→ y_2 is related to y_3

→ and so on...

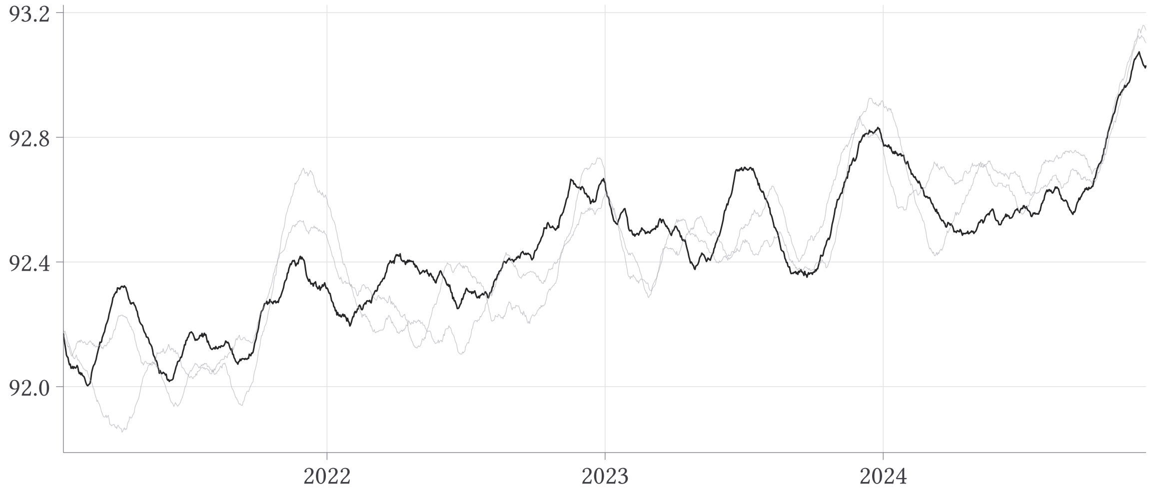
In some sense, we have only a 'single' observation

Observed Time Series

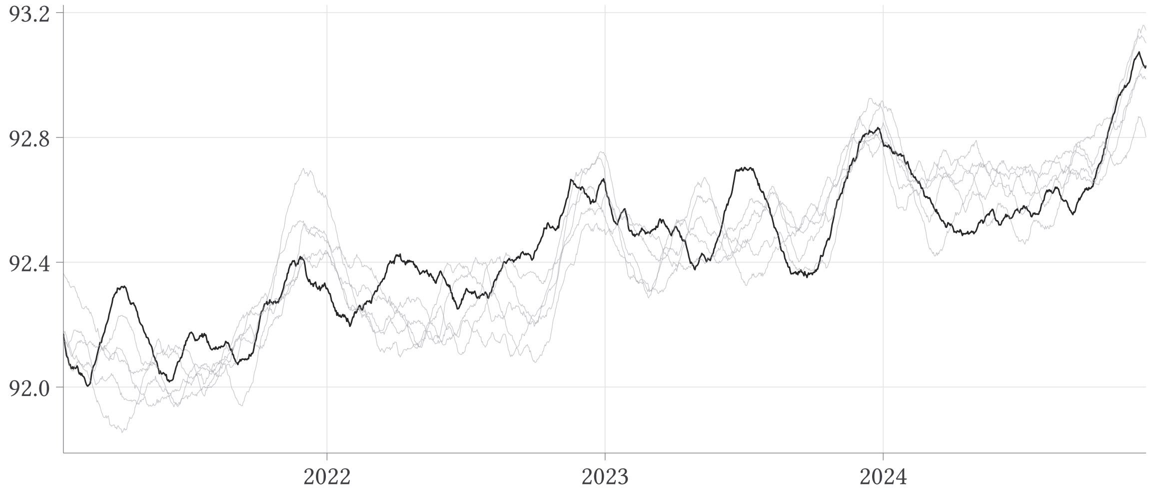
Observed Time Series + 1 Extra Sample



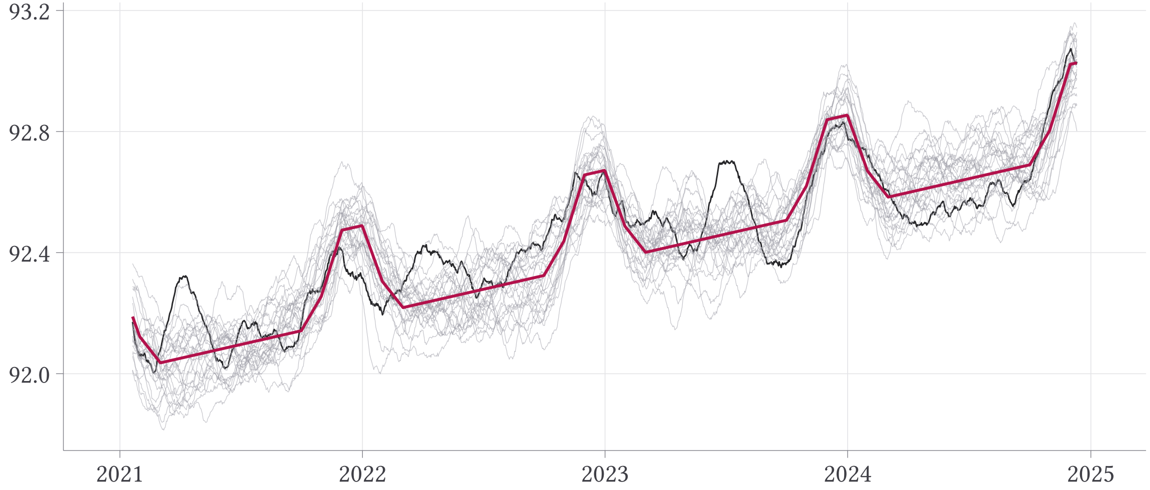
Observed Time Series + 2 Extra Samples



Observed Time Series + 5 Extra Samples



Observed Time Series + 25 Extra Samples; Systematic component: μ_t



Thinking about inference with time-series

$$y_1, y_2, \dots, y_T$$

While every observation might be related to one another, we are typically willing to assume that as you move away in time, observations become less and less correlated.

> “”

Thinking about inference with time-series

$$y_1, y_2, \cdots, y_T$$

Hence, statistical inference is quite a bit more challenging in time-series and requires weak dependency central limit theorems

$\rightarrow$ The intuition is that as you have larger and larger T, the information you have increases

Thinking about inference with time-series

$$y_1, y_2, \dots, y_T$$

Hence, statistical inference is quite a bit more challenging in time-series and requires weak dependency central limit theorems

$\rightarrow$ The intuition is that as you have larger and larger T, the information you have increases

But, we will not spend much time discussing that in this class:

$\rightarrow$ If you are running time-series regressions, default to Newey West standard errors; in `fixest`, use `vcov = NW(lag = #)`

Introduction to Time-Series

Learning from Time-Series

Time-series Statistics

What we can gain from using time-series

Time-series forecasting can be useful to:

→ Predict future values based on past data

→ Inform decision-making by anticipating changes over time

→ Identify patterns like trends or seasonality

Two goals of time-series

There are two possible goals that we can tackle when working with time-series data:

1. Learn about *persistent* patterns in how y_t evolves over time while ignoring random fluctuations (inference)
   - E.g. learn about seasonality, trends, etc.
2. Predict future values of y_t (forecasting)
   - The above step might be useful in predicting future y, but not necessary (only care about prediction)

Will try to clarify when we are discussing forecasting vs. describing time-series patterns (inference)

Learning from time-series

We observe a set of time-series observations y_t. Think of the observed y as being generated by

$$y_t = \mu_t + \varepsilon_t$$

→ μ_t is the 'typical' or 'systematic' value of y at time t

→ ε_t is a random fluctuation

Of course, we do not know which fluctuations are due to μ_t changing over time or ε_t changing over time

→ Without any more structure, this is is an impossible task

Learning from time-series

$$y_t = \mu_t + \varepsilon_t$$

Say we assume $\mathbb{E}[\varepsilon_t] = 0$

$\rightarrow$ On average over different draws of the time-series, the error term is on average 0

But, our observed time-series is a single draw, so it's not obvious that the noise will 'average away'

$\rightarrow$ Is the bump in the time-series just a shock that affected the unit for multiple periods or a systematic component

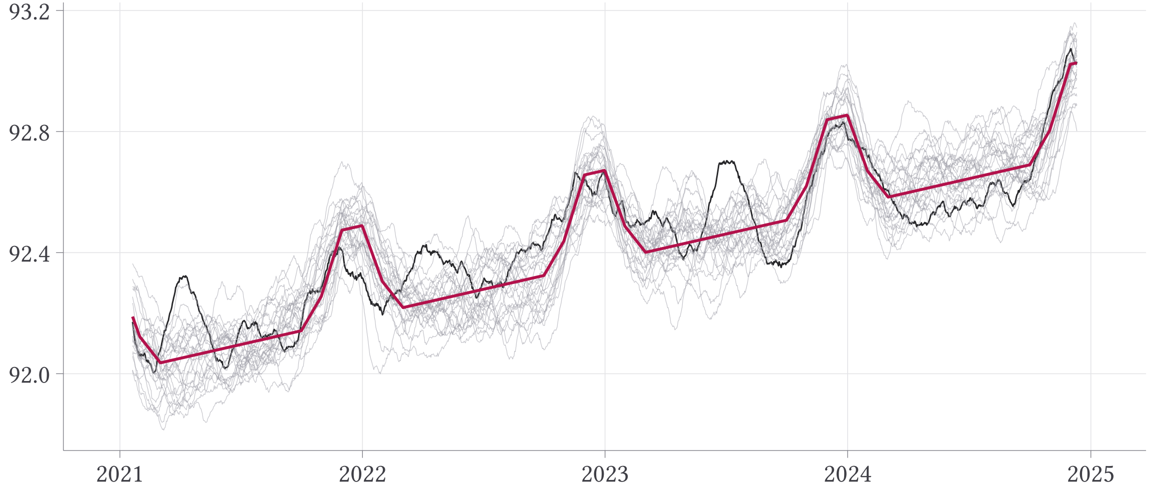
Learning from time-series

$$y_t = \mu_t + \varepsilon_t$$

Two options for solving this:

1. Assume that ε_t is not too persistent, and use some kind of 'smoothing' method to smooth out noise
2. Rely on functional form assumptions and run a time-series regression
   - By pooling over time, we are averaging out noise (but requires us to model μ_t well)

Observed Time Series + 25 Extra Samples; Systematic component: μ_t



$$y_t = \mu_t + \varepsilon_t$$

Here are some examples of what we can hope to learn using time-series data:

1. Identify seasonality in data
   - Does the change in μ_t over the year follow a standard pattern?
   - E.g. retail sales increasing in December
2. Detect long-term trends and short-term shocks
   - How does μ_t change over time?
   - E.g. trends in GDP changing over time?
   - E.g. recessions
3. Assess how strongly autocorrelated the data is
   - How ‘sticky’ shocks are from past periods are

Key insight in time-series forecasting

Key Insight: By analyzing the changes across time, we reveal structure and patterns that help in making better predictions. For example:

→ Does yesterday's sales help us learn about what products people will buy today?

→ Do we see an up-swing in jacket sales every October?

Key insight in time-series forecasting

Key Insight: By analyzing the changes across time, we reveal structure and patterns that help in making better predictions. For example:

→ Does yesterday's sales help us learn about what products people will buy today?

→ Do we see an up-swing in jacket sales every October?

Of course, this can fail if the underlying structure of the world changes over time

→ If we are using data from early 2000s on homes, we will surely fail at forecasting during the Great Recession

→ Assumptions on the stability of the time-series is called stationarity

Evaluating forecasting methods

As usual, we can use the mean-squared prediction error to evaluate our models:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$$

$\rightarrow$ Typically, will evaluate on the time-series data you do observe

Plotting residuals

It is also common in time-series methods to plot the residuals over time: t on x-axis and $y_t - \hat{y}_t$ on y-axis.

$\rightarrow$ If your forecast is doing a good job, then you should see no pattern in the residuals ("eyeball test")

Beware!! Just because there's no remaining patterns in the residuals, does not mean your model is well fit. You could be overfitted!!

Evaluating forecasting methods

Time-series forecasting is particularly difficult to evaluate

$\rightarrow$ Our training data is past-values up until today

$\rightarrow$ Our testing data is values in the future

If the structure of the world changes over time, then our testing data *can* look fundamentally different over time

$\rightarrow$ Consumer preferences change over time can make predicting future sales hard

Over-fitting

For this reason, we have to be *very* careful when using forecasting methods on time-series

$\rightarrow$ Over-fitting the past data makes us learn 'false' time-series relationships

Introduction to Time-Series

Learning from Time-Series

Time-series Statistics

Statistics of Time-series

For the next few slides, we will discuss some statistics of time-series data that we might be interested in

To review, in cross-sectional data, we mainly cared about:

- $\rightarrow$ the mean and the variance of a single variable, and
- $\rightarrow$ the correlation between two variables

Autocovariance

Autocovariance measures the covariance between a variable and a lagged version of itself over successive time periods.

In formal terms, the autocovariance at lag k is defined as:

$$\gamma_k = \text{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$$

where:

$\rightarrow$ μ is the mean of y_t,

$\rightarrow$ $\text{Cov}(y_t, y_{t-k})$ is the covariance between y_t and y_{t-k}.

Autocovariance

$$\gamma_k = \text{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$$

Intuition: Autocovariance helps quantify how much the past values of y move together with its current value.

$\rightarrow$ When y_{t-k} was above the mean, was y_t typically above it's mean?

Autocovariance

$$\gamma_k = \text{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$$

Intuition: Autocovariance helps quantify how much the past values of y move together with its current value.

$\rightarrow$ When y_{t-k} was above the mean, was y_t typically above it's mean?

In most settings, it is likely that $\gamma_1 \geq \gamma_2 \geq \dots$

$\rightarrow$ More-recent 'shocks' (in say $t - 1$) tend to persist for a little and then fade-out

Autocovariance

$$\gamma_k = \text{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$$

As an aside, note that when $k = 0$,

$$\gamma_0 = \text{Cov}(y_t, y_t) = \text{Var}(y_t)$$

Autocorrelation

Autocorrelation is the normalized version of autocovariance. It measures the correlation of a variable with its lagged values.

The autocorrelation at lag k is defined as:

$$\rho_k = \frac{\gamma_k}{\mathrm{Var}(y_t)} = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\mathrm{Var}(y_t)}$$

where:

- $\rightarrow$ γ_k is the autocovariance at lag k,
- $\rightarrow$ γ_0 is the variance of y_t (i.e., autocovariance at lag 0).

Autocorrelation

$$\rho_k = \frac{\gamma_k}{\text{Var}(y_t)} = \frac{\text{Cov}(y_t, y_{t-k})}{\text{Var}(y_t)}$$

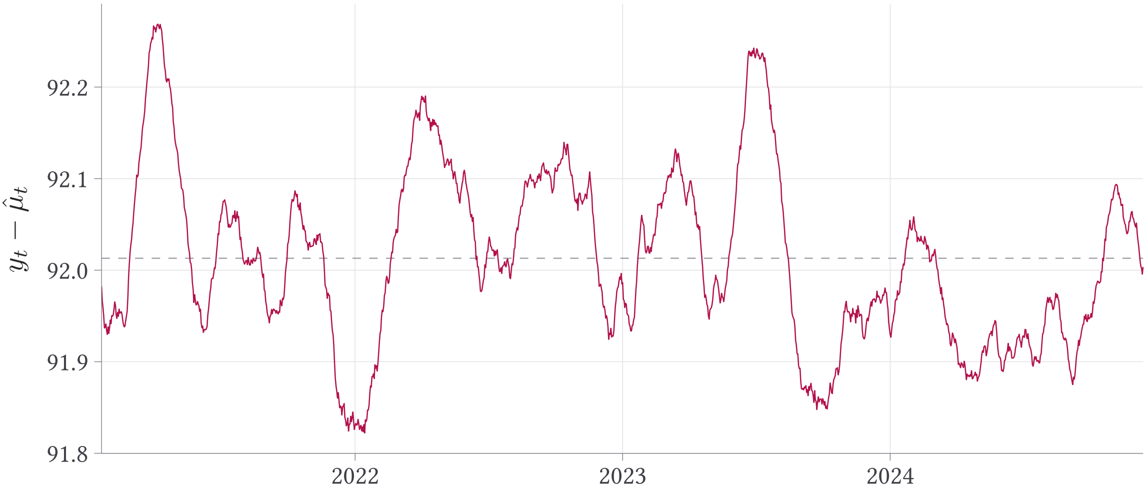
Intuition: Autocorrelation tells us the strength of the relationship between y_t and its past values. It ranges between -1 and 1.

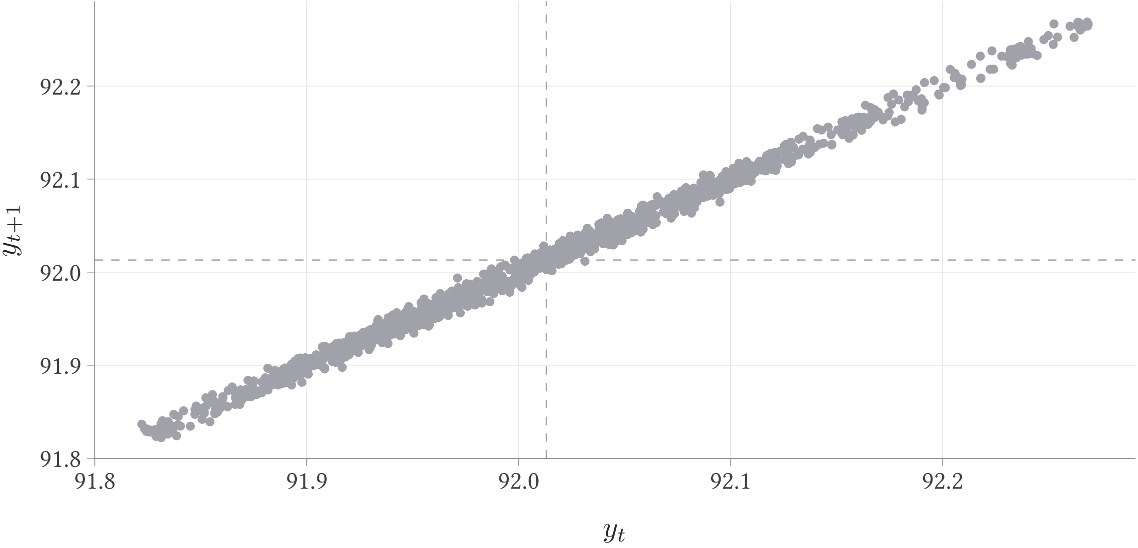
Examples of Covariance

Let's give two examples to help build intuition:

1. The first time-series will have very significant autocorrelation
2. The second time-series will have near zero autocorrelation

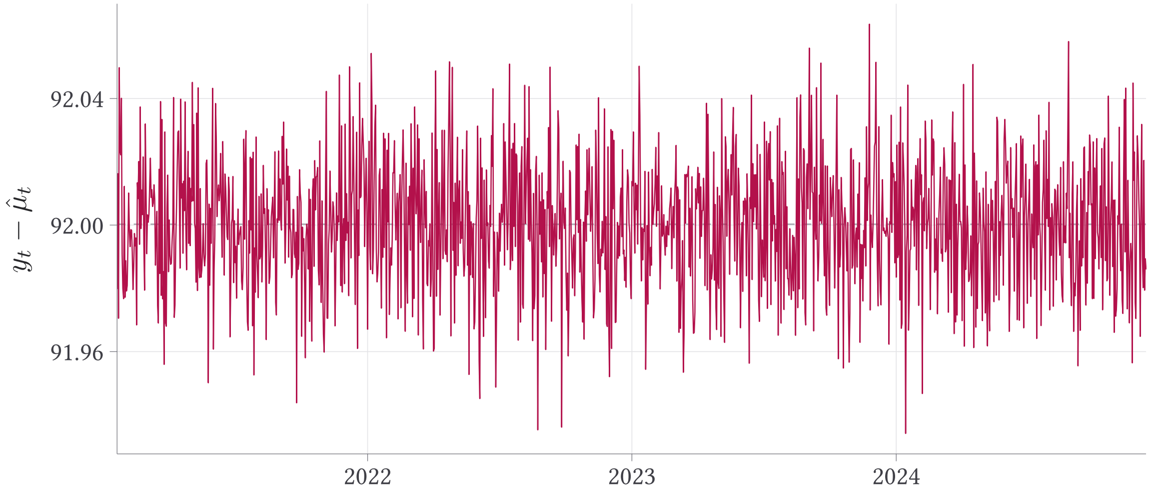
Observed Time Series

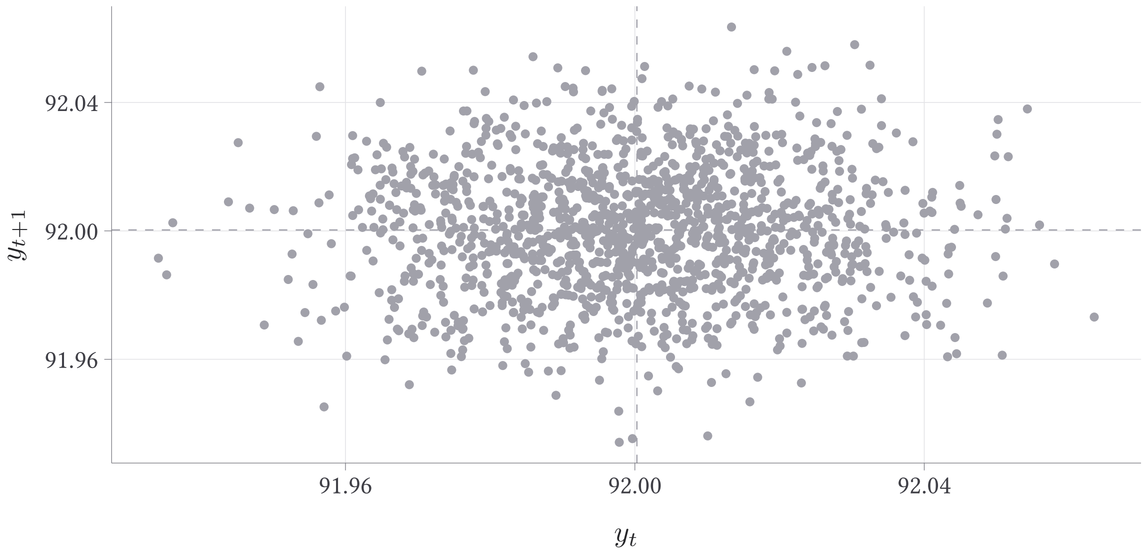




92.2
92.1
92.0
91.9
91.8
y_{t+1}
91.8
91.9
92.0
92.1
92.2
y_t

Observed Time Series




92.04
92.00
91.96
y_{t+1}
91.96
92.00
92.04
y_t

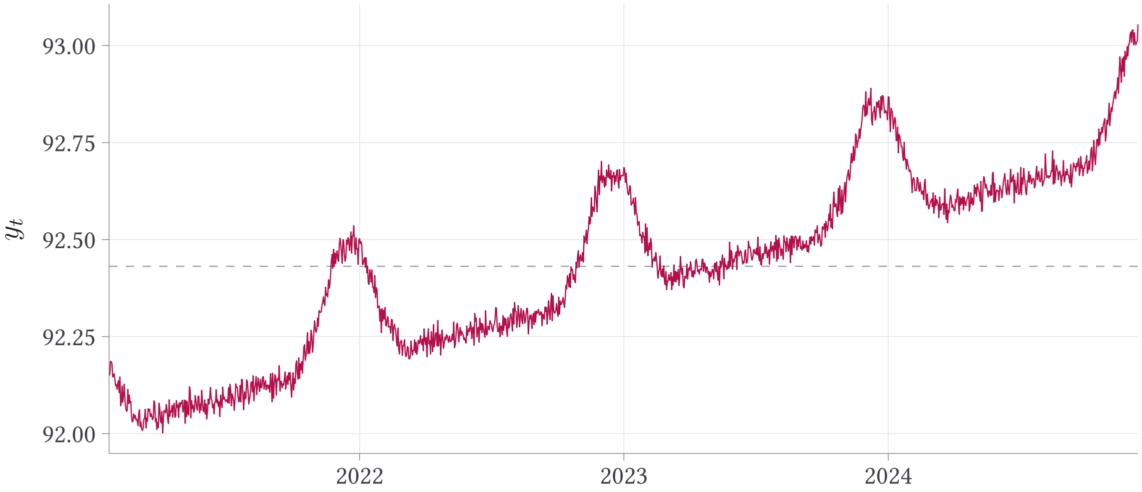
Autocorrelation with trends

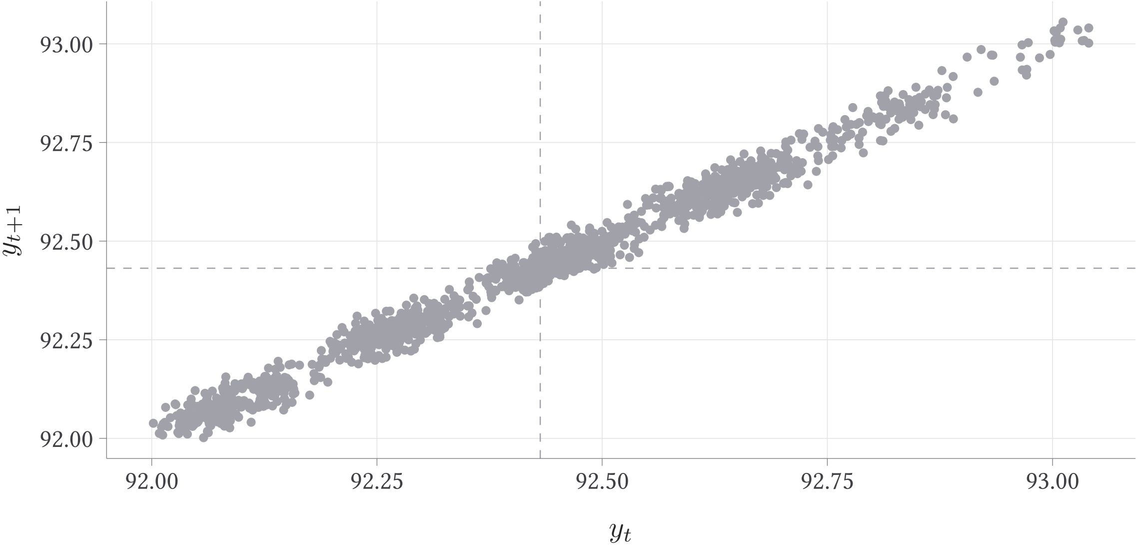
When the data is trending in a direction (e.g. up over time), the data will exhibit a strong autocorrelation.

E.g. we generate data with a trend and a winter seasonal effect plus an error term ε_t that is normal and independent in each period

$\rightarrow$ The error term ε_t exhibitis zero autocorrelation

Observed Time Series





93.00
92.75
92.50
92.25
92.00
y_{t+1}
92.00
92.25
92.50
92.75
93.00
y_t

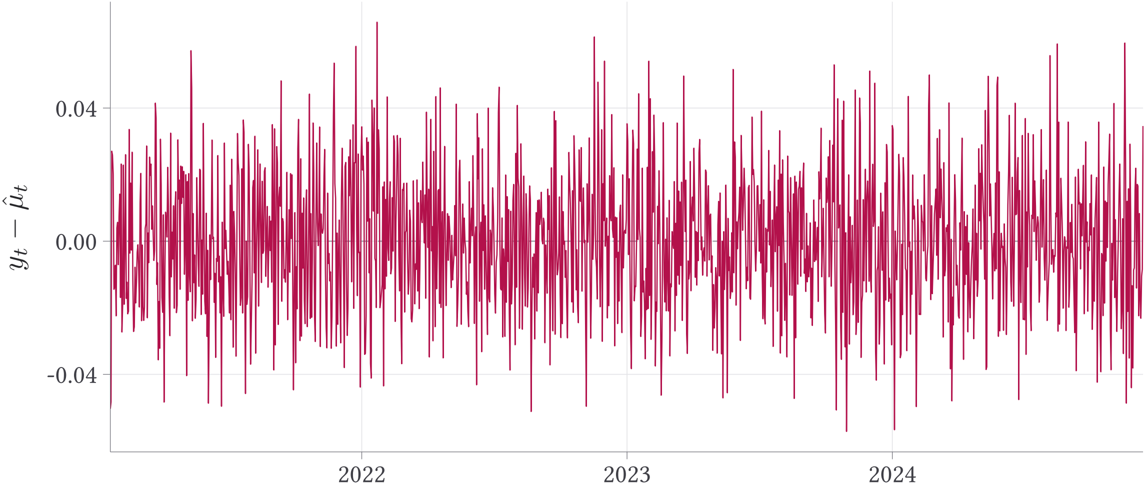
Autocorrelation with trends

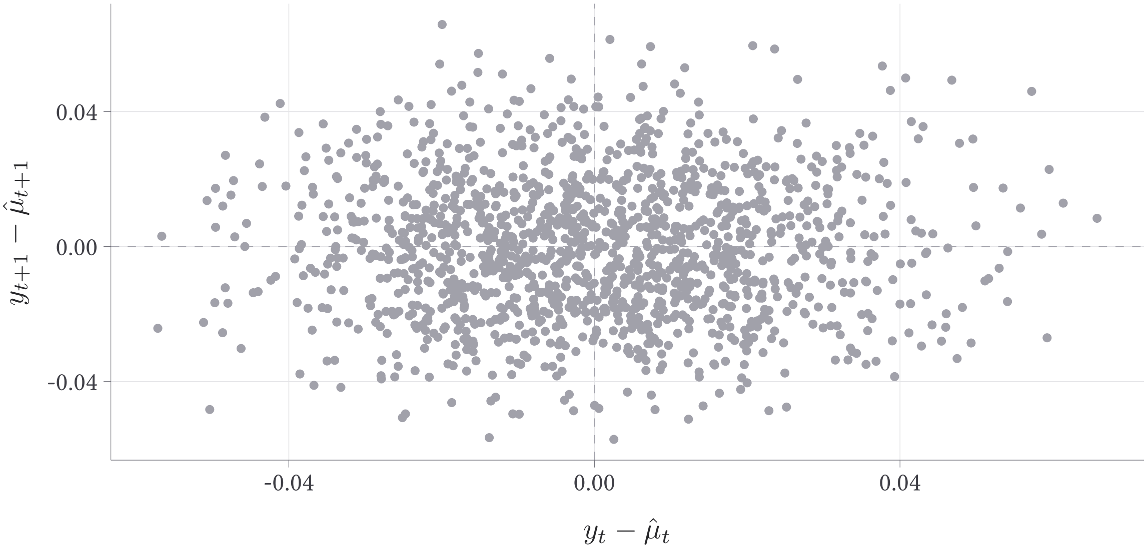
Now, let's estimate a time-series regression (we will see how in the future) that estimates the trend and seasonal effects and subtracts them off

Then, we can evaluate the autocorrelation of the "de-trended data": $y_t - \hat{y}_t$

$\rightarrow$ In our example, we generated ε_t with zero autocorrelation, so let's see how that looks

Residualized Time Series




0.04
0.00
-0.04
-0.04
0.00
0.04
$y_{t+1} - \hat{\mu}_{t+1}$
$y_t - \hat{\mu}_t$

Unemployment Rate Example

In the unemployment example, the time-series

$$\hat{\gamma}_1 = \text{Cov}(y_t, y_{t-1}) = 2.968 \quad \text{and} \quad \hat{\rho}_1 = 0.961$$

$\rightarrow$ Unsurprisingly the correlation of unemployment from 1-month to the next is very strong

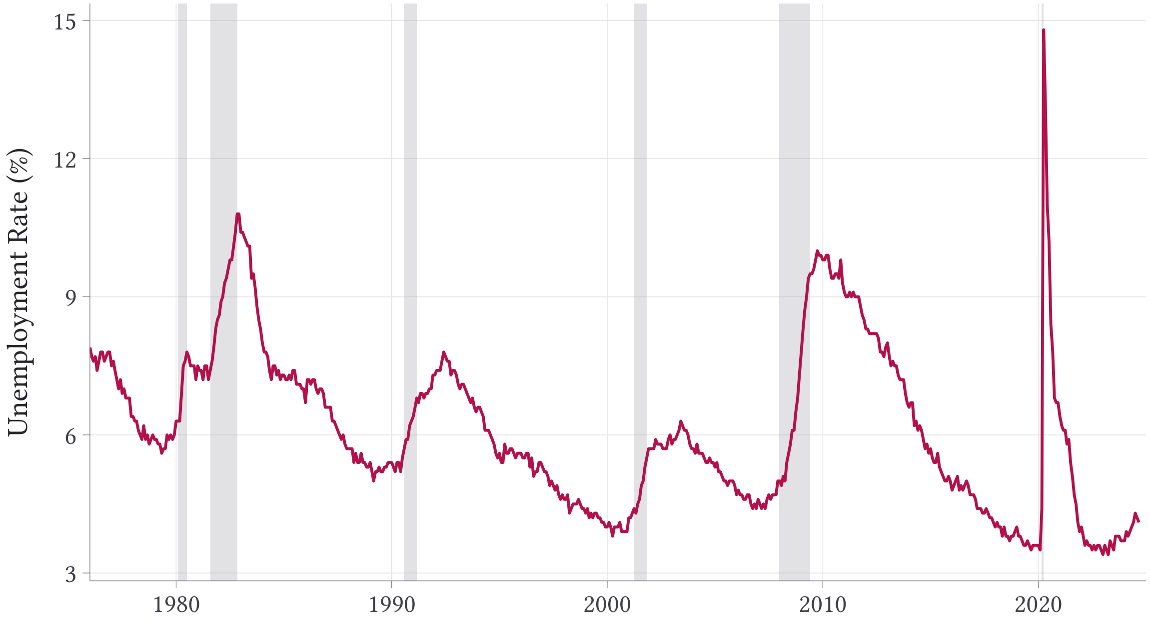
Unemployment Rate Example

In the unemployment example, the time-series

$$\hat{\gamma}_1 = \text{ Cov}(y_t, y_{t-1}) = 2.968 \quad \text{and} \quad \hat{\rho}_1 = 0.961$$

→ Unsurprisingly the correlation of unemployment from 1-month to the next is very strong

This is useful for forecasting; a very strong autocorrelation tells us that recent values of y should be useful for predicting future values of y



Unemployment Rate (%)
15
12
9
6
3
1980
1990
2000
2010
2020

Unemployment Rate Example

Let's look at the correlation unemployment over 12 periods (year to year)

$$\hat{\rho}_{12} = 0.659$$

$\rightarrow$ Shocks to last year's unemployment seem to 'persist' into the current period

Unemployment Rate Example

If we use the reshuffled gdp data, what do we think the autocorrelation may be?

Unemployment Rate Example

If we use the reshuffled gdp data, what do we think the autocorrelation may be?

$$\hat{\rho}_{1,\texttt{reshuffled}} = -0.03081183$$

When we completely randomly shuffled the data, we have destroyed any autocorrelation!

$\rightarrow$ This makes sense. If I reshuffled the data, knowing last month's (reshuffled) unemployment is no longer useful for predicting this month's (reshuffled) unemployment rate

How to calculate in R

The first thing we need to do is calculate the sample mean $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$ using `mean(y)`.

We want two vectors

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{T-1} \\ y_T \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} y_2 \\ \vdots \\ y_T \\ \texttt{NA} \end{bmatrix}$$

The first one is our original vector `y` and we need `L1_y` ("lag 1 y").

Then, calculate $\frac{1}{T}\sum_{t=2}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y})$

How to calculate in R

First, we will do it by hand to make sure we follow all the steps

$\rightarrow$ Note this requires the time-series to be sorted in order!

```r
y <- 1:10
T <- length(y)
y_dm <- y - mean(y)
sum(y_dm[1:(T - 1)] * y_dm[2:T]) / T
#> [1] 5.775
```

Or we can use the acf function to make things way easier

```
# Or, using a function
acf(y, lag.max = 1, type = "covariance", plot = FALSE)
#>    0    1
#> 8.25 5.78

acf(y, lag.max = 1, plot = FALSE)
#> Autocorrelations of series 'y', by lag
#>
#>   0   1
#> 1.0 0.7
```