Time series analysis

EDS 222

Tamma Carleton
Fall 2023

Announcements/check-in

- Assignment 04 posted, due 12/08

Announcements/check-in

- Assignment 04 posted, due 12/08
- A note on depth in coming lectures

Announcements/check-in

- Assignment 04 posted, due 12/08
- A note on depth in coming lectures
- **No class** 11/23

Announcements/check-in

- Assignment 04 posted, due 12/08
- A note on depth in coming lectures
- **No class** 11/23
- Final projects: due in 3.5 weeks!
    - Presentations: 12/12 4:00-7:00pm (Bren Hall 1424)
    - Blog posts: 12/15

Today

What are time series data?

Today

What are time series data?

Decomposition

Today

What are time series data?

Decomposition

Autocorrelation

Today

What are time series data?

Decomposition

Autocorrelation

Time series and OLS

What are time series data?

What are time series data?

Up to this point, we focused on **cross-sectional data**.

- Sampled *across* a population (*e.g.*, people, counties, countries).
- Sampled at *one moment* in time (*e.g.*, Jan. 1, 2015).
- We had n *individuals*, each indexed i in $\{1, \ldots, n\}$.

What are time series data?

Up to this point, we focused on **cross-sectional data**.

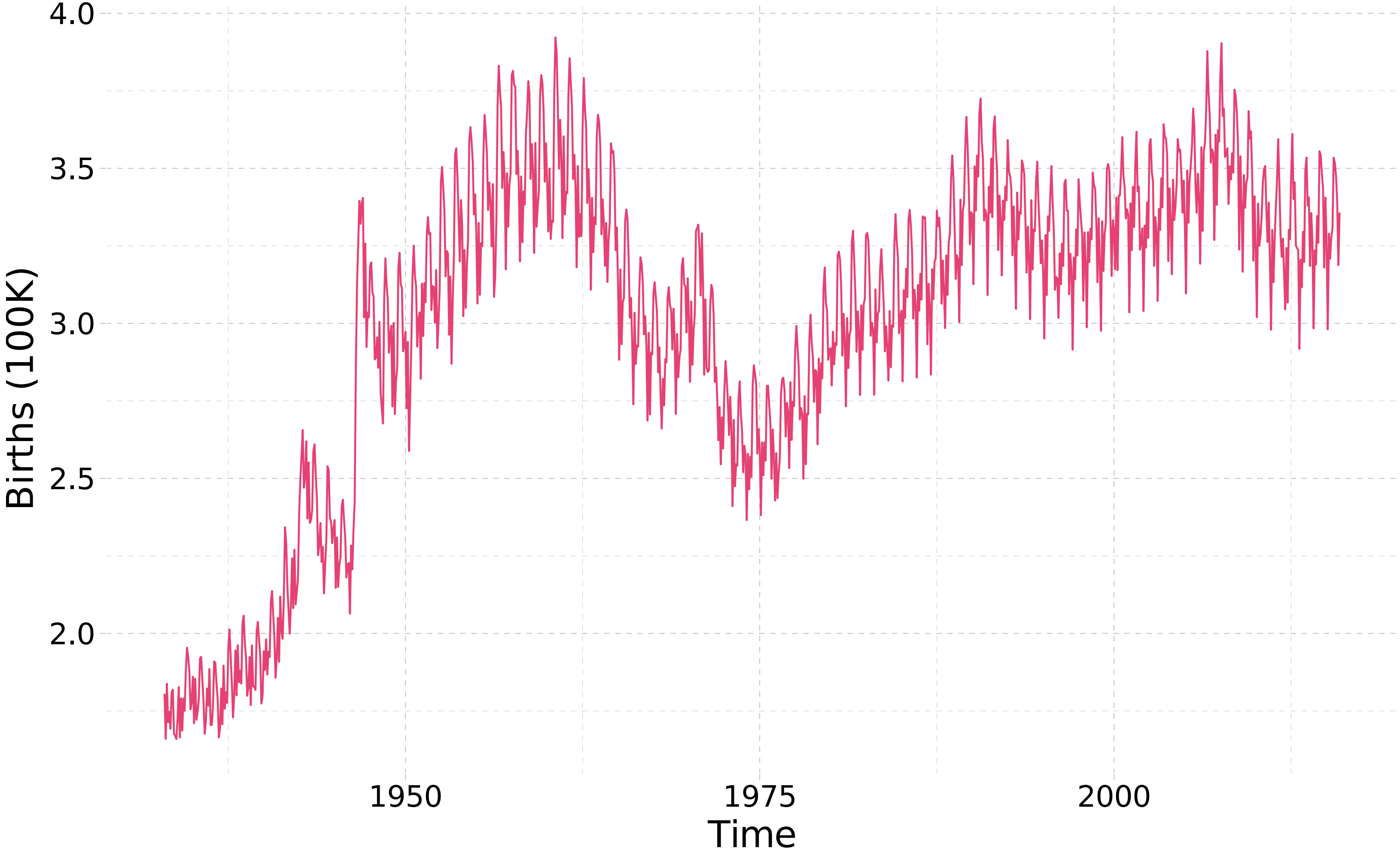
- Sampled *across* a population (*e.g.*, people, counties, countries).
- Sampled at *one moment* in time (*e.g.*, Jan. 1, 2015).
- We had n *individuals*, each indexed i in $\{1, \ldots, n\}$.

Today, we focus on a different type of data: **time-series data**.

- Sampled within one unit/individual (*e.g.*, Oregon).
- Observe multiple times for the same unit (*e.g.*, Oregon: 1990–2020).
- We have T time periods, each indexed t in $\{1, \ldots, T\}$.

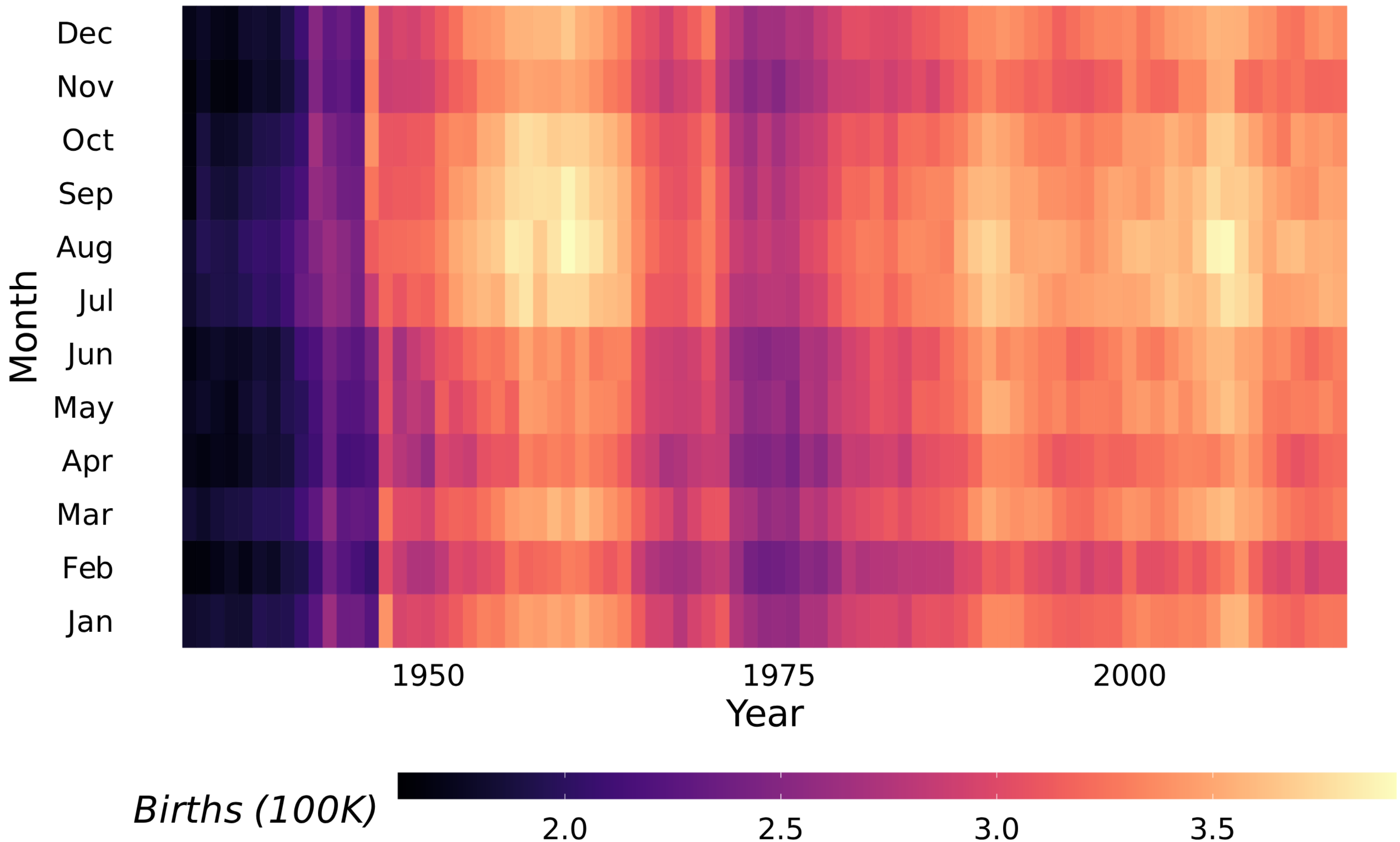
Time series data: Example

US monthly births, 1933–2015: Classic time-series graph



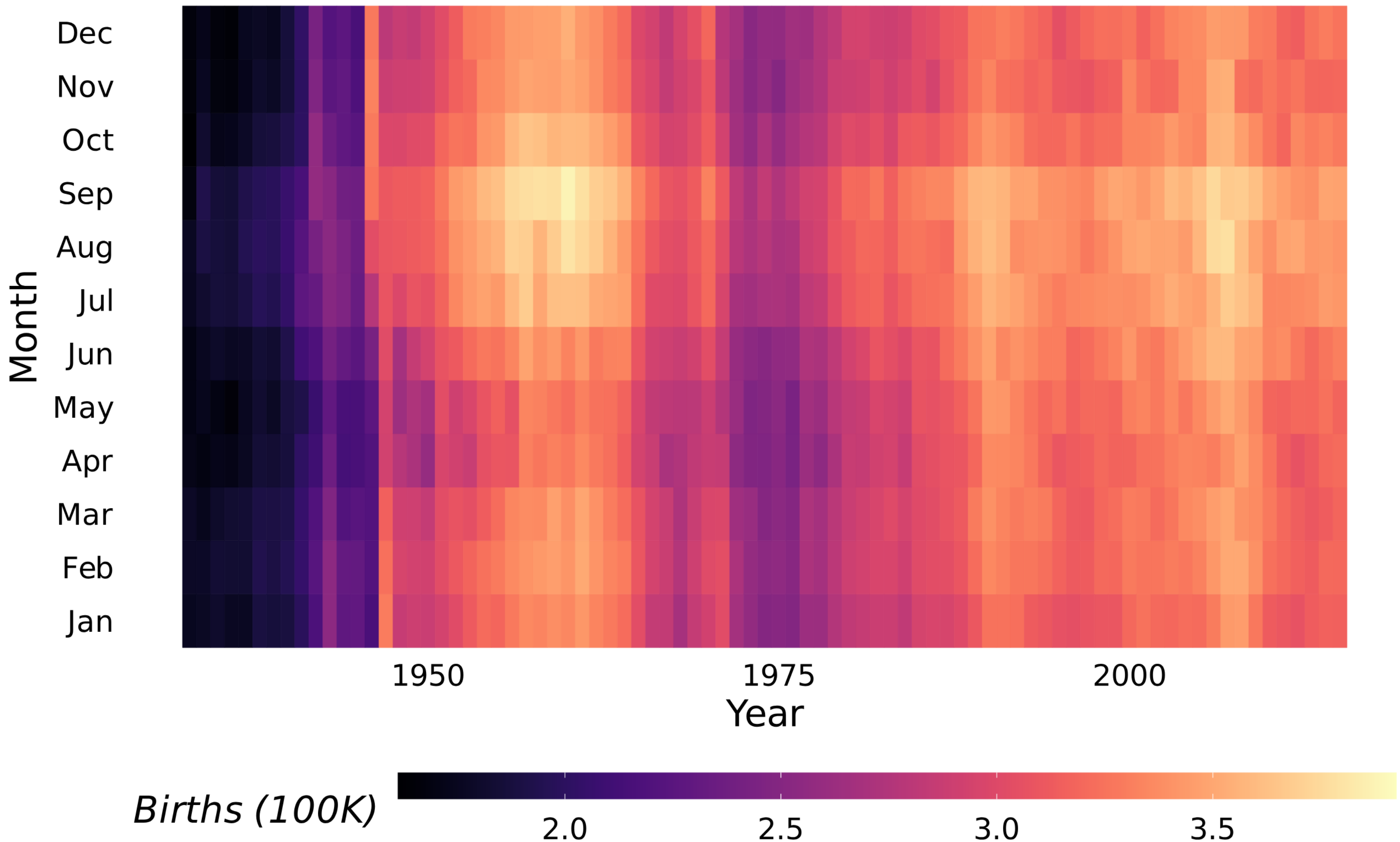
Time series data: Example

US monthly births, 1933–2015: Newfangled time-series graph



Time series data: Example

US monthly births per 30 days, 1933–2015: Newfangled time-series graph



You already have (many of) the tools

- Time series data open some **new questions and new challenges** for statistical analysis

You already have (many of) the tools

- Time series data open some **new questions and new challenges** for statistical analysis
- But you **already have many of the tools** you need!

You already have (many of) the tools

- Time series data open some **new questions and new challenges** for statistical analysis

- But you **already have many of the tools** you need!

- E.g., recall:

$$Ozone_t = \beta_0 + \beta_1 Temp_t + \varepsilon_t$$

You already have (many of) the tools

- Time series data open some **new questions and new challenges** for statistical analysis

- But you **already have many of the tools** you need!

- E.g., recall:

$$Ozone_t = \beta_0 + \beta_1 Temp_t + \varepsilon_t$$

- Description of `airquality` data:

> Daily air quality measurements in New York, May to September 1973.

You already have (many of) the tools

- Time series data open some **new questions and new challenges** for statistical analysis

- But you **already have many of the tools** you need!

- E.g., recall:

$$Ozone_t = \beta_0 + \beta_1 Temp_t + \varepsilon_t$$

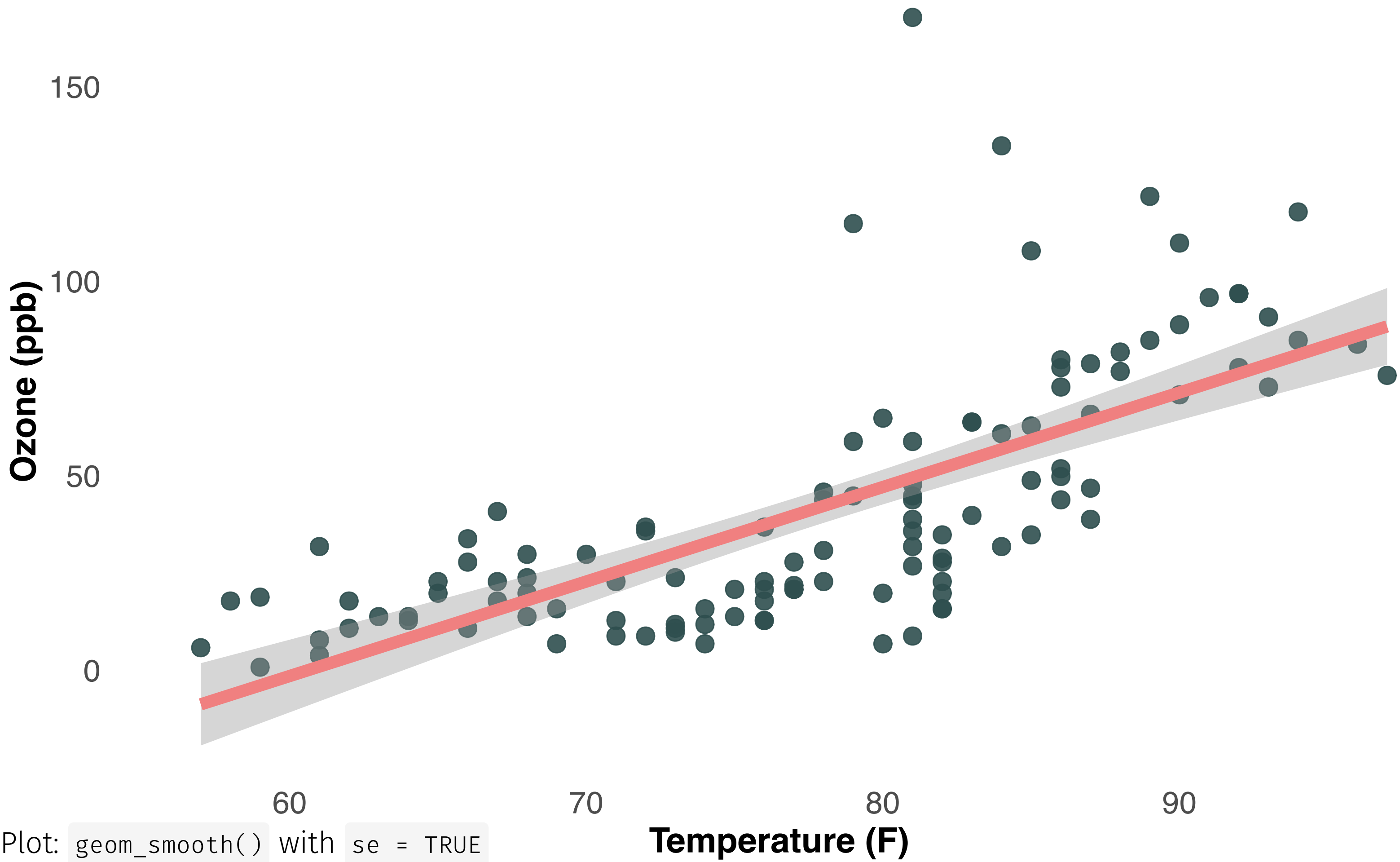
- Description of `airquality` data:

> Daily air quality measurements in New York, May to September 1973.

- These are **time series data** and we already ran an OLS regression with them!

You already have (many of) the tools

$$Ozone_t = \beta_0 + \beta_1 Temp_t + \varepsilon_t$$



Plot: `geom_smooth()` with `se = TRUE`

You already have (many of) the tools

Let *date* indicate the date, ranging from May, 1 to September 31, 1973.

You already have (many of) the tools

Let *date* indicate the date, ranging from May, 1 to September 31, 1973.

We can also estimate:

$$Ozone_t = \beta_0 + \beta_1 date_t + \varepsilon_t$$

You already have (many of) the tools

Let *date* indicate the date, ranging from May, 1 to September 31, 1973.

We can also estimate:

$$Ozone_t = \beta_0 + \beta_1 date_t + \varepsilon_t$$

```
airqts = airquality %>% mutate(date = make_datetime(1973, Month,Day))
head(airqts)
#>   Ozone Solar.R Wind Temp Month Day       date
#> 1    41     190  7.4   67     5   1 1973-05-01
#> 2    36     118  8.0   72     5   2 1973-05-02
#> 3    12     149 12.6   74     5   3 1973-05-03
#> 4    18     313 11.5   62     5   4 1973-05-04
#> 5    NA      NA 14.3   56     5   5 1973-05-05
#> 6    28      NA 14.9   66     5   6 1973-05-06
```

You already have (many of) the tools

Let *date* indicate the date, ranging from May, 1 to September 31, 1973.

We can also estimate:

$$Ozone_t = \beta_0 + \beta_1 date_t + \varepsilon_t$$

```
airqts = airquality %>% mutate(date = make_datetime(1973, Month,Day))
head(airqts)
#>   Ozone Solar.R Wind Temp Month Day       date
#> 1    41     190  7.4   67     5   1 1973-05-01
#> 2    36     118  8.0   72     5   2 1973-05-02
#> 3    12     149 12.6   74     5   3 1973-05-03
#> 4    18     313 11.5   62     5   4 1973-05-04
#> 5    NA      NA 14.3   56     5   5 1973-05-05
#> 6    28      NA 14.9   66     5   6 1973-05-06
```

- Regression of *Ozone* on *date* estimates a **linear trend** in ozone
- Tip: `make_datetime()` from the `lubridate` package (handy for dates and times)

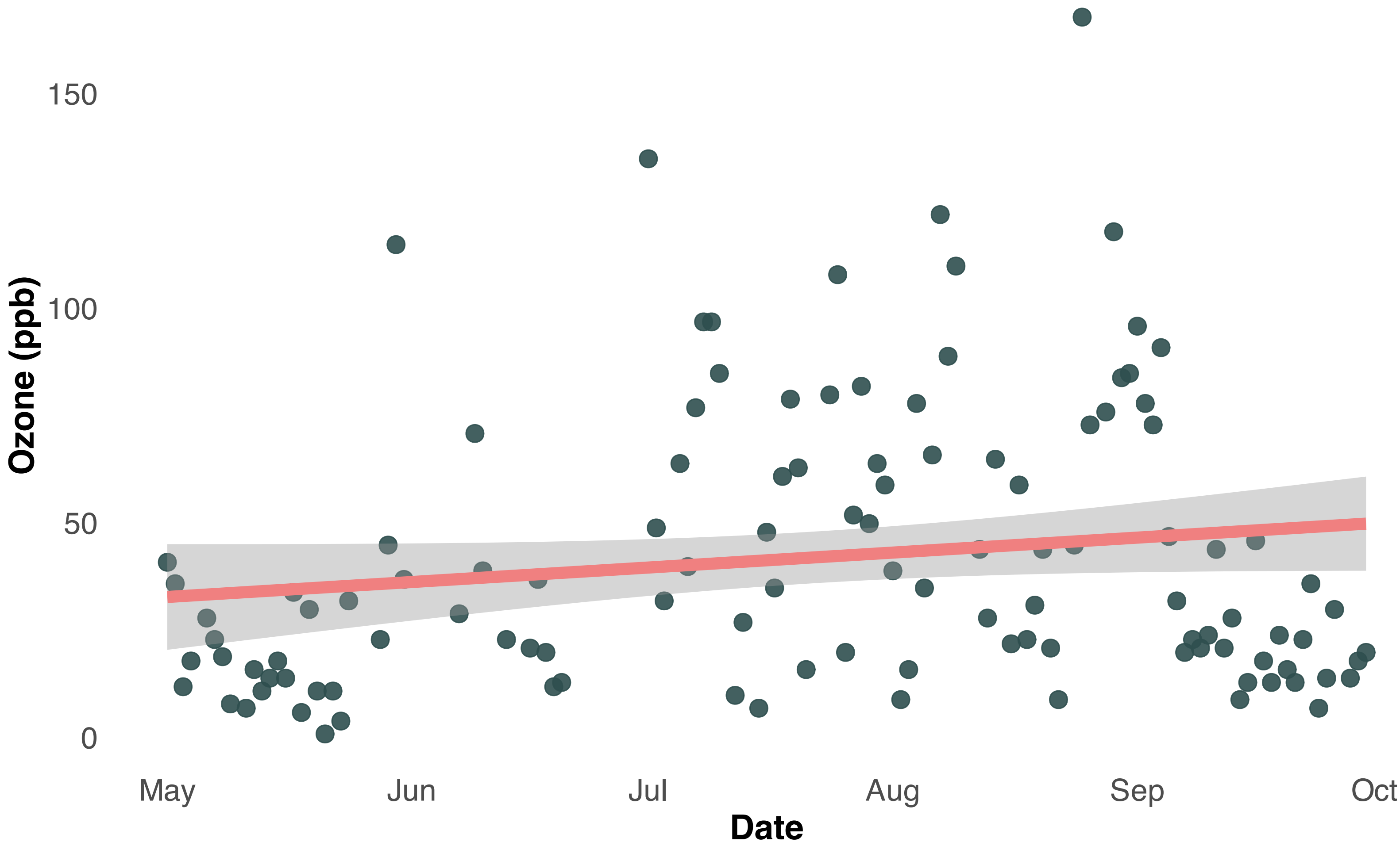
You already have (many of) the tools

$$Ozone_t = \beta_0 + \beta_1 date_t + \varepsilon_t$$

```
summary(lm(Ozone ~ date, data = airqts))
#> 
#> Call:
#> lm(formula = Ozone ~ date, data = airqts)
#> 
#> Residuals:
#>    Min     1Q Median     3Q    Max 
#> -42.32 -24.58  -8.39  20.46 122.05 
#> 
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)  
#> (Intercept) -1.04e+02   8.59e+01   -1.21    0.230  
#> date         1.30e-06   7.65e-07    1.70    0.092 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 32.7 on 114 degrees of freedom
#>   (37 observations deleted due to missingness)
#> Multiple R-squared:  0.0247,	Adjusted R-squared:  0.0162 
#> F-statistic: 2.89 on 1 and 114 DF,  p-value: 0.092
```

You already have (many of) the tools

$$Ozone_t = \beta_0 + \beta_1 date_t + \varepsilon_t$$



You already have (many of) the tools

- Many of the summary statistics, regression, and hypothesis testing tools apply to time series data without much adjustment

You already have (many of) the tools

- Many of the summary statistics, regression, and hypothesis testing tools apply to time series data without much adjustment
- But there are some new **features** we want to explore:
    - Does my data have exhibit **trending behavior**?
    - Is there **seasonality**?
    - Is my data **cyclical**?

You already have (many of) the tools

- Many of the summary statistics, regression, and hypothesis testing tools apply to time series data without much adjustment
- But there are some new **features** we want to explore:
  - Does my data have exhibit **trending behavior**?
  - Is there **seasonality**?
  - Is my data **cyclical**?
- And some new **challenges** to overcome:
  - Additional **assumptions** needed in OLS
  - Threat to existing assumptions: Are our error terms **independent**? Is **exogeneity** harder now?

Decomposition

Time series components

Seasonality

A repeated pattern over known and equal periods (e.g., month; quarter, decade)

Time series components

Seasonality

A repeated pattern over known and equal periods (e.g., month; quarter, decade)

Cyclicality

A broader cyclical trend with unknown and/or unequal periods (e.g., business cycle, ENSO)

Time series components

Seasonality

A repeated pattern over known and equal periods (e.g., month; quarter, decade)

Cyclicality

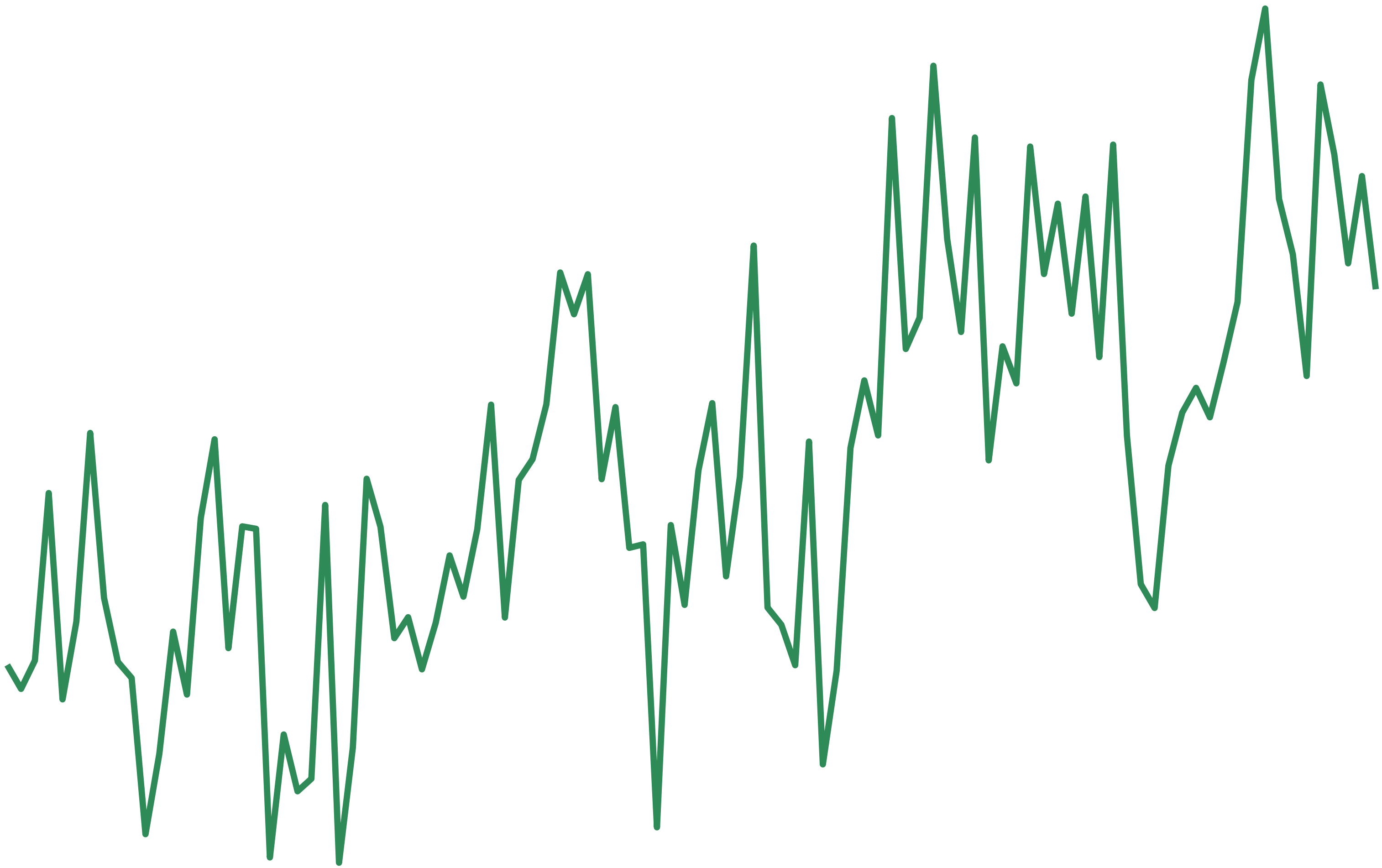
A broader cyclical trend with unknown and/or unequal periods (e.g., business cycle, ENSO)

Trends

Long-term increase or decrease in the data (not necessarily linear!)

Time series components

Often, seasonality, cyclicality and trends occur all at the same time:



Time series components

For many time series,* we can decompose the data into:

$$y_t = S_t + T_t + R_t$$

where S_t is a **seasonal** component, T_t is the cycle *and* trend components, and R_t is the remainder.

Time series components

For many time series,[*] we can decompose the data into:

$$y_t = S_t + T_t + R_t$$

where S_t is a **seasonal** component, T_t is the cycle *and* trend components, and R_t is the remainder.

Decomposition allows us to isolate each component of the time series visually and quantitatively.

[*]: This decomposition is "additive", which works for many time series. See Hyndman for details on more complex "multiplicative" decomposition.

Decomposition: Moving averages

A key tool in "decomposing" a time series into its component parts is computing a **moving average**

Decomposition: Moving averages

A key tool in "decomposing" a time series into its component parts is computing a **moving average**

A moving average of order *m* is computed as:

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}$$

where $m = 2k + 1$.

Decomposition: Moving averages

A key tool in "decomposing" a time series into its component parts is computing a **moving average**

A moving average of order *m* is computed as:

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}$$

where $m = 2k + 1$.

The moving average gives you an estimate of the irregular trend-cycle component T at time *t* by averaging values of the time series within *k* periods of *t*

Moving average example

Computing an $m = 5$ moving average over the data plotted on the last slide:

```
df = as.data.frame(cbind(x, y)) # these are the data we plotted above
df = df %>% mutate(ma = slider::slide_dbl(y, mean,
                   .before = 2, .after = 2, .complete = TRUE))
```

Moving average example

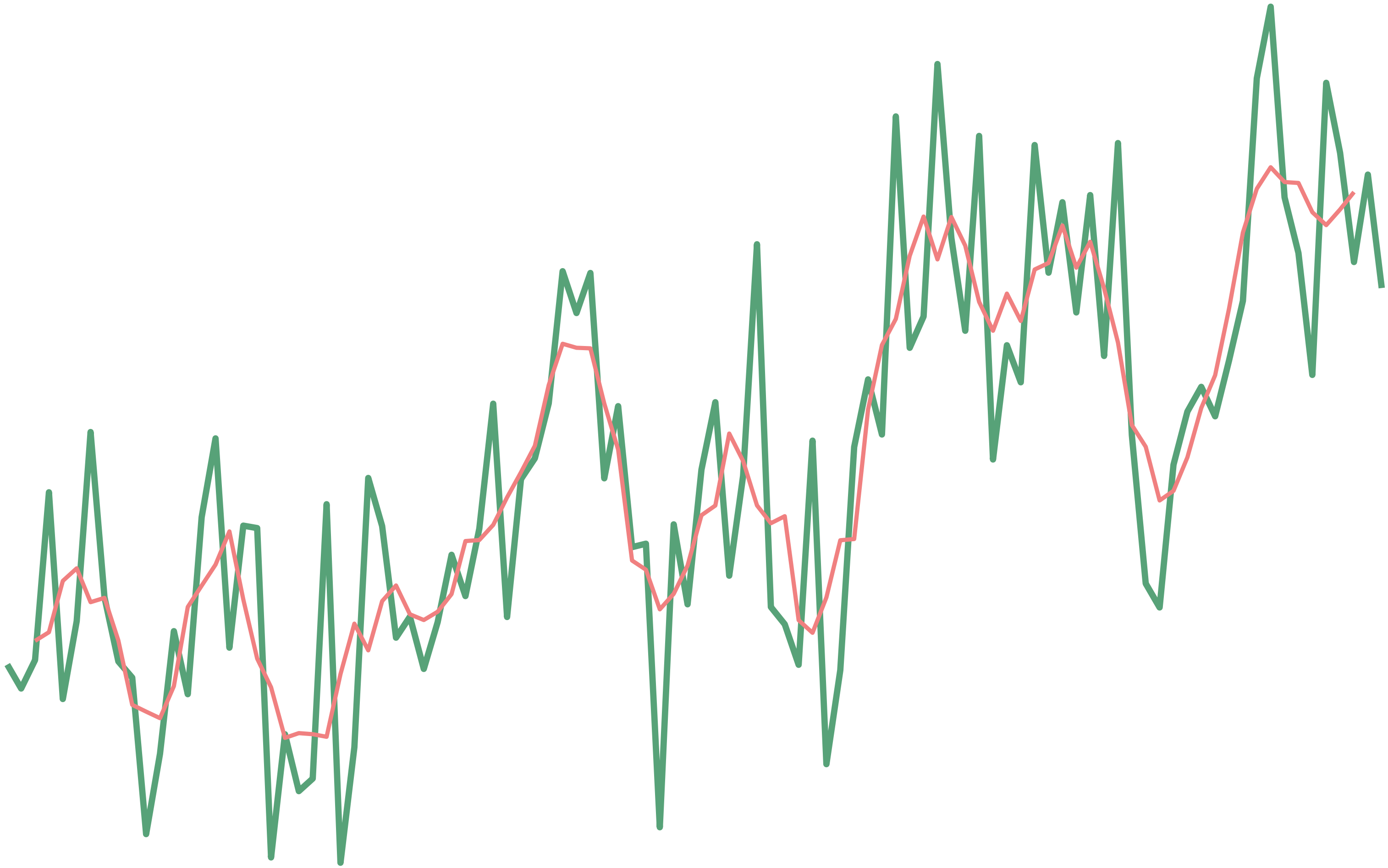
Computing an $m = 5$ moving average over the data plotted on the last slide:

```
df = as.data.frame(cbind(x, y)) # these are the data we plotted above
df = df %>% mutate(ma = slider::slide_dbl(y, mean,
                   .before = 2, .after = 2, .complete = TRUE))
```

- Helpful package: `slider` (there are others too!)
- Option `.complete=TRUE` ensures only moving windows with complete data are computed

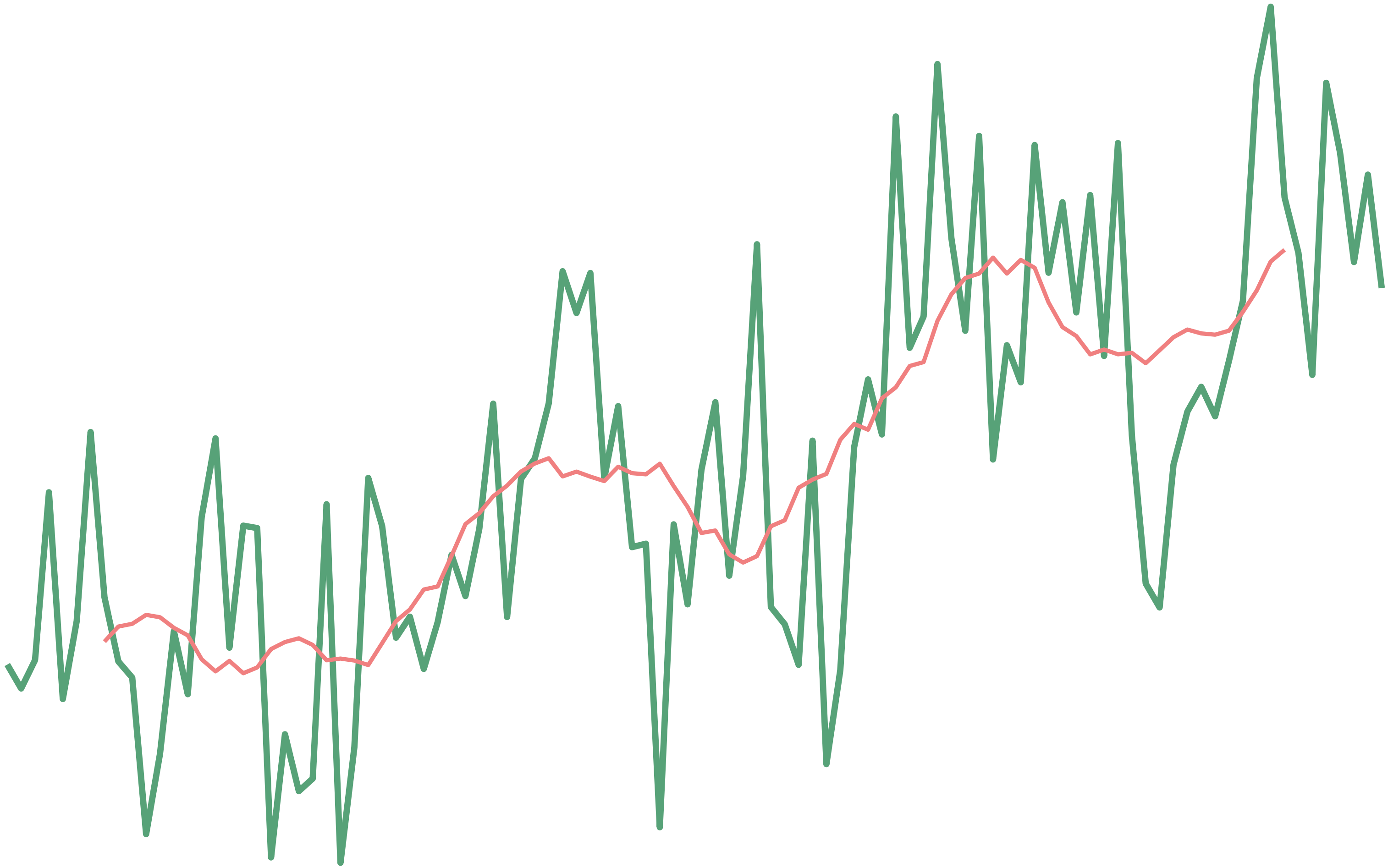
Moving average example

Computing an $m = 5$ moving average:



Moving average example

Computing an $m = 15$ moving average:



Classical decomposition

Step 1: estimate a moving average

Estimate an m-moving average to compute $\hat{T}_t$

Classical decomposition

Step 1: estimate a moving average

Estimate an m-moving average to compute $\hat{T}_t$

Step 2: calculate the de-trended series

De-trended series $= y_t - \hat{T}_t$

Classical decomposition

Step 1: estimate a moving average

Estimate an m-moving average to compute $\hat{T}_t$

Step 2: calculate the de-trended series

De-trended series $= y_t - \hat{T}_t$

Step 3: calculate seasonality

Simple average over de-trended series for each season s

Classical decomposition

Step 1: estimate a moving average

Estimate an m-moving average to compute $\hat{T}_t$

Step 2: calculate the de-trended series

De-trended series $= y_t - \hat{T}_t$

Step 3: calculate seasonality

Simple average over de-trended series for each season s

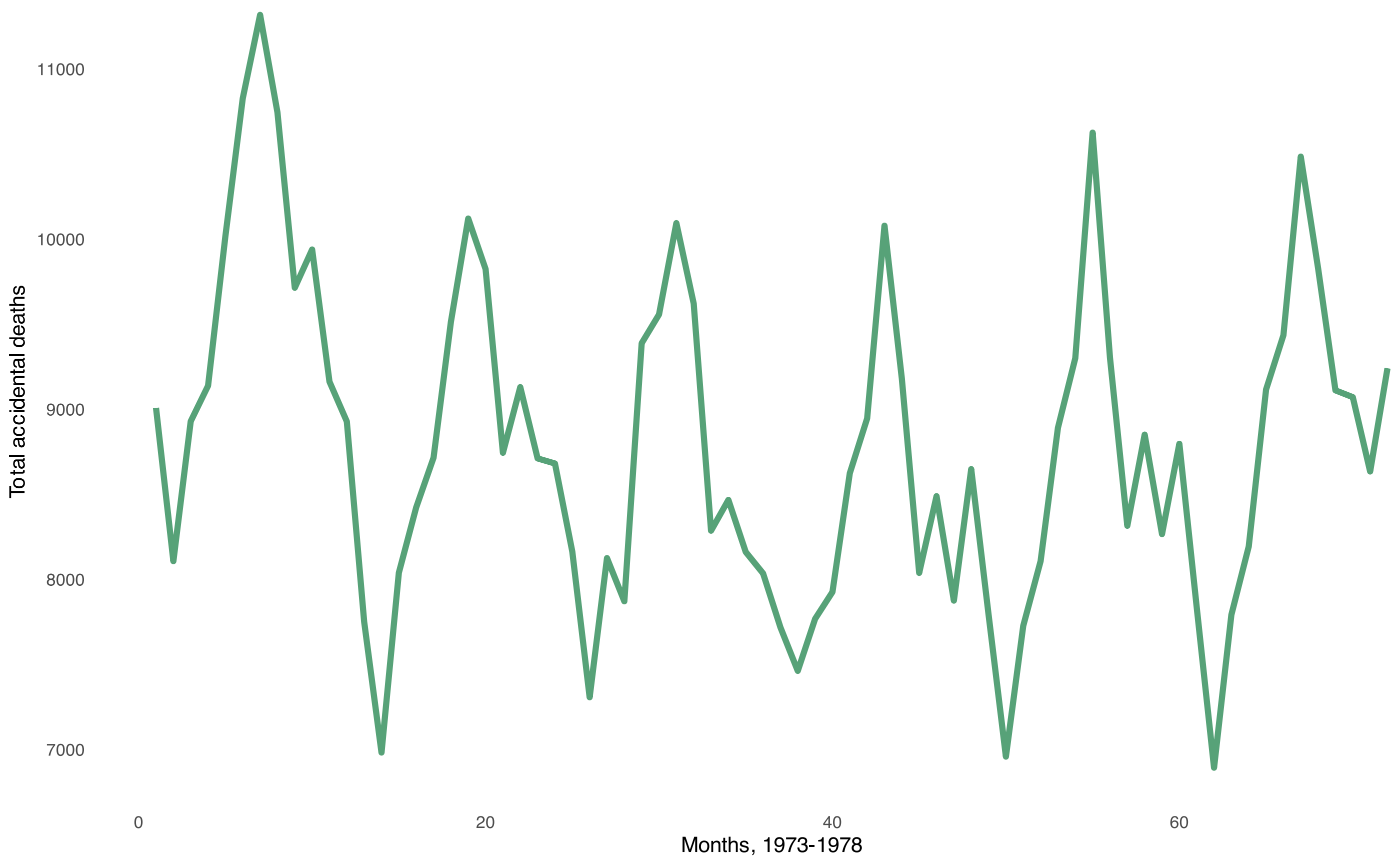
Step 4: remainder

Whatever is left over

Classical decomposition

Consider a time series of monthly totals of accidental deaths in the USA:

```
df = USAccDeaths
```



Classical decomposition

Let's decompose the accidental deaths time series.

You can do this by hand, or...

Classical decomposition

Let's decompose the accidental deaths time series.

You can do this by hand, or...

```
decomp = as_tsibble(USAccDeaths) %>%
  model(
    classical_decomposition(value, type = "additive")
  ) %>%
  components()
head(decomp)
#> # A dable: 6 x 7 [1M]
#> # Key:      .model [1]
#> # :         value = trend + seasonal + random
#>   .model                         index value trend seasonal random season_adj
#>   <chr>                          <mth> <dbl> <dbl>    <dbl>  <dbl>         <c
#> 1 "classical_decomposition(v… 1973 Jan  9007    NA    -806.     NA          98
#> 2 "classical_decomposition(v… 1973 Feb  8106    NA   -1523.     NA          96
#> 3 "classical_decomposition(v… 1973 Mar  8928    NA    -741.     NA          96
#> 4 "classical_decomposition(v… 1973 Apr  9137    NA    -515.     NA          96
#> 5 "classical_decomposition(v… 1973 May 10017    NA     340.     NA          96
#> 6 "classical_decomposition(v… 1973 Jun 10826    NA     745.     NA          10(
```

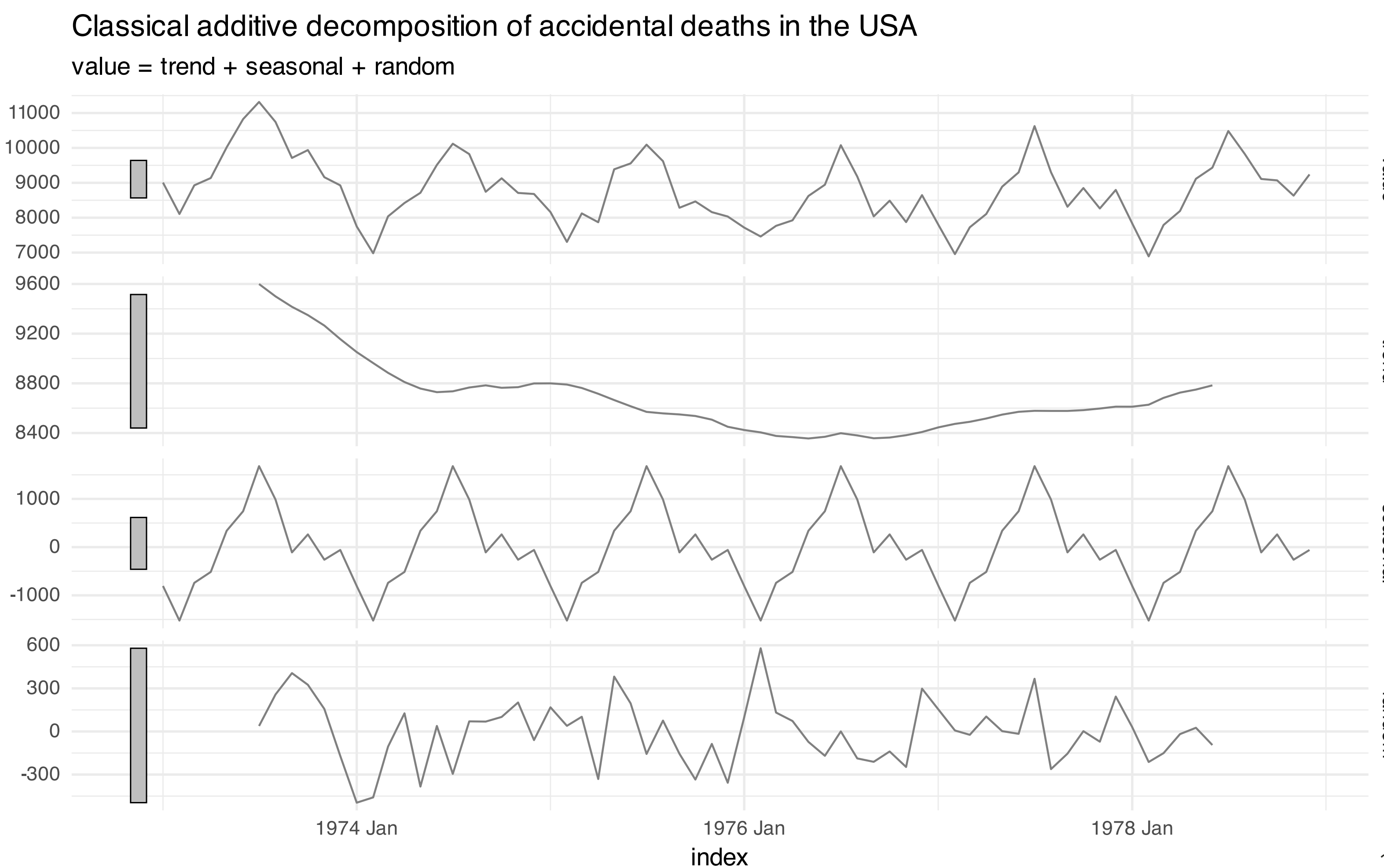
Classical decomposition

You can do this by hand, or...

```
as_tsibble(USAccDeaths) %>%
  model(
    classical_decomposition(value, type = "additive")
  ) %>%
  components() %>%
  autoplot() +
  labs(title = "Classical additive decomposition of accidental deaths in the USA
```

Classical decomposition

You can do this by hand, or...



Decomposition

- As outlined in Hyndman & Athanasopoulos, **classical decomposition has some drawbacks**:
  - Assumes the seasonal component is fixed over time
  - Loses data at the start and end (due to moving average)
  - Can be sensitive to outliers/short-run anomalous behavior

Decomposition

- As outlined in Hyndman & Athanasopoulos, **classical decomposition has some drawbacks**:
    - Assumes the seasonal component is fixed over time
    - Loses data at the start and end (due to moving average)
    - Can be sensitive to outliers/short-run anomalous behavior
- **Seasonal and Trend Decomposition using Loess** (STL)
    - Flexible and versatile method
    - Seasonal component can change over time
    - Robust to outliers
    - use `STL()` in place of `classical_decomposition()`

Why decompose a time series?

1. To **better understand** your data
    - Do summers tend to have higher crime?
    - Is there an positive trend in ocean temperatures?
    - Does deforestation follow business cycles?

Why decompose a time series?

1. To **better understand** your data

    - Do summers tend to have higher crime?
    - Is there an positive trend in ocean temperatures?
    - Does deforestation follow business cycles?

2. To aid in **forecasting**

    - You can forecast using estimated seasonality and trend-cycles
    - Details are not covered in this class, see Hyndman & Athanasopoulos for an overview and implementation in `R`

Autocorrelation

Autocorrelation

Many time series data are **autocorrelated**, meaning past values are correlated with future values (note: also called **serial correlation**)

Autocorrelation

Many time series data are **autocorrelated**, meaning past values are correlated with future values (note: also called **serial correlation**)

That is, y_t may be correlated with y_{t-1}, y_{t-2}, y_{t-12}, etc.

Autocorrelation

Many time series data are **autocorrelated**, meaning past values are correlated with future values (note: also called **serial correlation**)

That is, y_t may be correlated with y_{t-1}, y_{t-2}, y_{t-12}, etc.

This matters both for interpreting OLS output (in a few slides), and for understanding our data (helpful for identifying any seasonality).

Autocorrelation

For example:

- Today's temperature is **positively** correlated with yesterday's temperature: $cor(y_t, y_{t-1}) > 0$

Autocorrelation

For example:

- Today's temperature is **positively** correlated with yesterday's temperature: $cor(y_t, y_{t-1}) > 0$
- Today's temperature is **negatively** correlated with temperatures 6 months ago: $cor(y_t, y_{t-182}) < 0$

Autocorrelation

For example:

- Today's temperature is **positively** correlated with yesterday's temperature: $cor(y_t, y_{t-1}) > 0$
- Today's temperature is **negatively** correlated with temperatures 6 months ago: $cor(y_t, y_{t-182}) < 0$
- Today's temperature may have **no correlation** with temperatures 7 days ago: $cor(y_t, y_{t-7}) = 0$

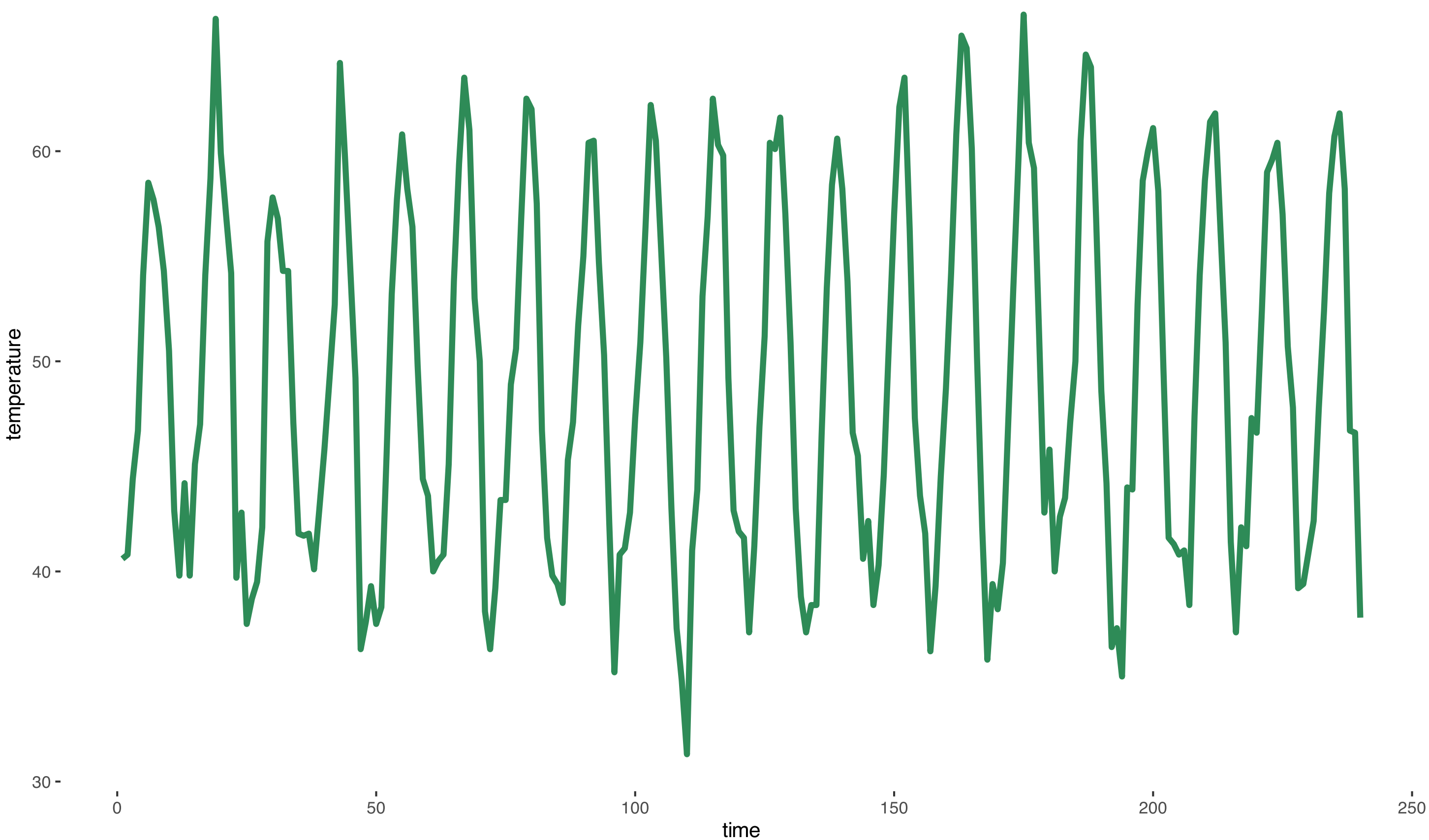
Autocorrelation

We can describe autocorrelation using an **autocorrelation function** or ACF.

Autocorrelation

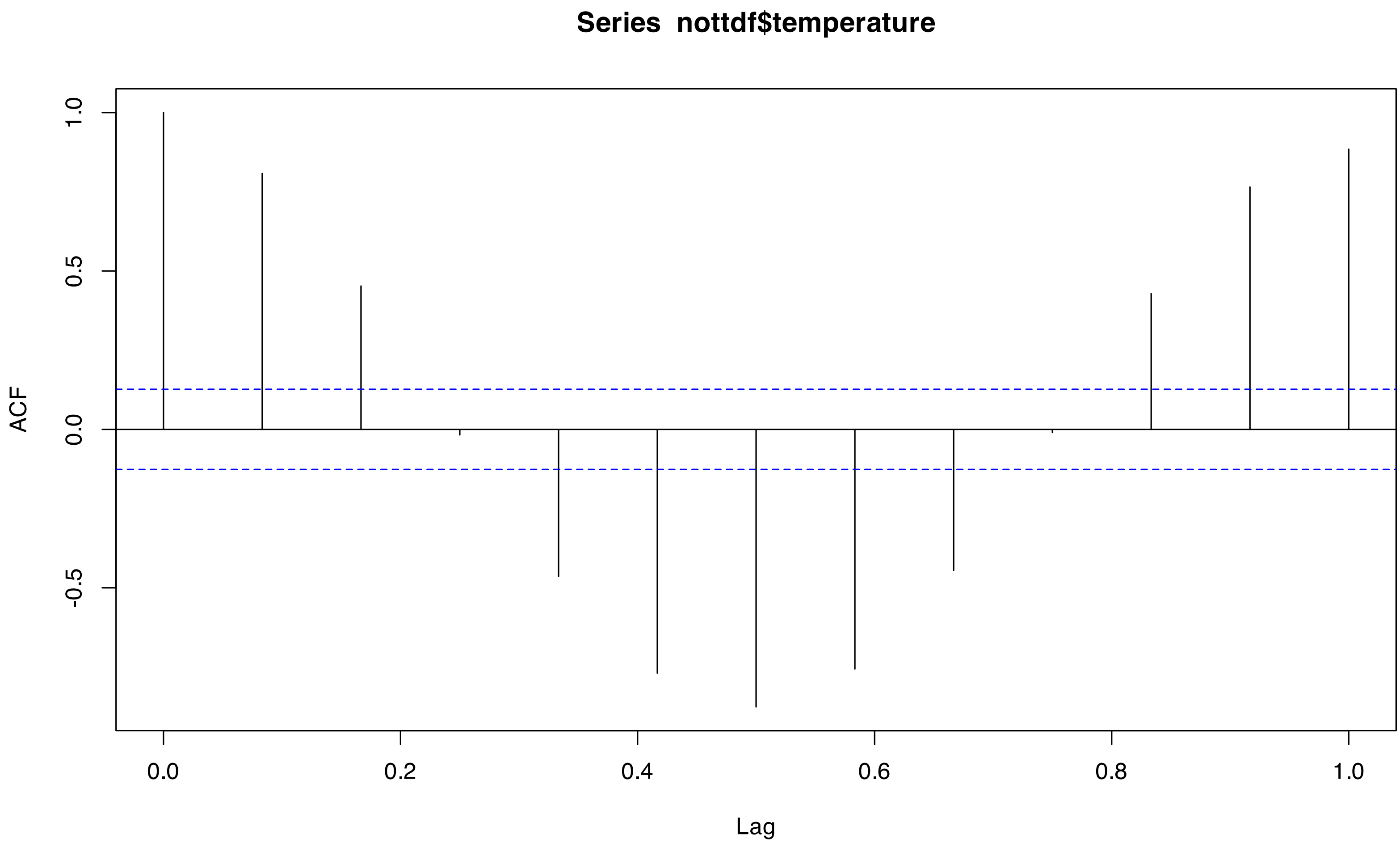
We can describe autocorrelation using an **autocorrelation function** or ACF.

Consider a **monthly** temperature time series for Nottingham Castle



Autocorrelation Function (ACF)

```
acf(nottdf$temperature, lag.max=12)
```



Autocorrelation Function (ACF)

> `acf()` plots an ACF for you!

- The height of each line indicates the correlation between temperature today and temperature l days ago
- Confidence intervals are shown in blue by default -- indicate if $cor(y_t, y_{t-l})$ is statistically distinguishable from zero (or not)
- Helps to identify periodicity of seasonality

Autocorrelation Function (ACF)

> `acf()` plots an ACF for you!

- The height of each line indicates the correlation between temperature today and temperature l days ago
- Confidence intervals are shown in blue by default -- indicate if $cor(y_t, y_{t-l})$ is statistically distinguishable from zero (or not)
- Helps to identify periodicity of seasonality

> Definition: **white noise** is a random time series in which there is no correlation across time periods (rare in the real world!). Here, the ACF would look noisy and correlations would largely fall within the blue confidence interval.

Time series and OLS

Intro to time series and OLS

Our model now looks something like

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + u_t$$

Intro to time series and OLS

Our model now looks something like

$$\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + u_t$$

or perhaps

$$\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + \beta_3 \textbf{passage}_{t-1} + u_t$$

Intro to time series and OLS

Our model now looks something like

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + u_t$$

or perhaps

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_3 \mathbf{passage}_{t-1} + u_t$$

maybe even

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_3 \mathbf{passage}_{t-1} + \beta_4 \mathbf{salmon}_{t-1} + u_t$$

Intro to time series and OLS

Our model now looks something like

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + u_t$$

or perhaps

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_3 \mathbf{passage}_{t-1} + u_t$$

maybe even

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_3 \mathbf{passage}_{t-1} + \beta_4 \mathbf{salmon}_{t-1} + u_t$$

where $t-1$ denotes the time period prior to t (*lagged* stream passage or salmon returns).

Time-series models

Updated OLS assumptions

1. **New: Weakly persistent outcomes**—essentially, x_{t+k} in the distant period $t+k$ is weakly correlated with period x_t (when k is "big").

2. y_t is a **linear function** of its parameters and disturbance.

3. There is **some variation** in our explanatory variables

4. **Harder to satisfy:** The u_t have conditional mean of zero (**exogeneity**), $\boldsymbol{E}[u_t|X] = 0$.

5. **Harder to satisfy:** The u_t are **normally distributed** and **homoskedastic** with **zero correlation** between u_t and u_s, *i.e.*, $u_t \overset{\text{iid}}{\sim} N\left(0,\, \sigma^2\right)$, $\text{Var}(u_t|X) = \text{Var}(u_t) = \sigma^2$, and $\text{Cor}(u_t,\, u_s|X) = 0$.

Time-series models

Model options

Time-series modeling boils down to two classes of models.

1. **Static models:** Do not allow for persistent effects.

2. **Dynamic models:** Allow for persistent effects.

Time-series models

Model options

Time-series modeling boils down to two classes of models.

1. **Static models:** Do not allow for persistent effects.

2. **Dynamic models:** Allow for persistent effects.

    - Models with **lagged explanatory** variables

    - **Autoregressive, distributed-lag** (ADL) models

Model options

Option 1: **Static models**

Static models assume the outcome depends upon only the current period.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + u_t$$

Model options

Option 1: Static models

Static models assume the outcome depends upon only the current period.

$$\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + u_t$$

Here, we must believe that stream passage **immediately** affects the number of salmon returns and does not affect on the numbers of returns in the future.

Model options

Option 1: Static models

Static models assume the outcome depends upon only the current period.

$$\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + u_t$$

Here, we must believe that stream passage **immediately** affects the number of salmon returns and does not affect on the numbers of returns in the future.

We also need to believe current salmon returns do not depend upon previous stream passage conditions.

Model options

Option 1: Static models

Static models assume the outcome depends upon only the current period.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + u_t$$

Here, we must believe that stream passage **immediately** affects the number of salmon returns and does not affect on the numbers of returns in the future.

We also need to believe current salmon returns do not depend upon previous stream passage conditions.

Can be a very restrictive way to consider time-series data.

Model options

Option 2: Dynamic models

Dynamic models allow the outcome to depend upon other periods.

Model options

Option 2a: Dynamic models with lagged explanatory variables

These models allow the outcome to depend upon the explanatory variable(s) in other periods.

$$\begin{aligned}\text{salmon}_t = &\beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \\ &\beta_3 \text{passage}_{t-2} + \beta_4 \text{passage}_{t-3} + u_t\end{aligned}$$

Model options

Option 2a: Dynamic models with lagged explanatory variables

These models allow the outcome to depend upon the explanatory variable(s) in other periods.

$$\begin{aligned}\text{salmon}_t =& \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \\ & \beta_3 \text{passage}_{t-2} + \beta_4 \text{passage}_{t-3} + u_t\end{aligned}$$

Here, passage **immediately** affects the number of salmon returns *and* affects **future** numbers of returns.

Model options

Option 2a: Dynamic models with lagged explanatory variables

These models allow the outcome to depend upon the explanatory variable(s) in other periods.

$$\begin{aligned}\textbf{salmon}_t =& \beta_0 + \beta_1 \textbf{passage}_t + \beta_2 \textbf{passage}_{t-1} + \\ & \beta_3 \textbf{passage}_{t-2} + \beta_4 \textbf{passage}_{t-3} + u_t\end{aligned}$$

Here, passage **immediately** affects the number of salmon returns *and* affects **future** numbers of returns.

In other words: salmon returns today depend today's stream passage conditions and *lags* of passage—*e.g.*, last year's passage, the year before last, etc...

Model options

Option 2a: Dynamic models with lagged explanatory variables

These models allow the outcome to depend upon the explanatory variable(s) in other periods.

$$\begin{aligned}\text{salmon}_t =& \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \\ & \beta_3 \text{passage}_{t-2} + \beta_4 \text{passage}_{t-3} + u_t\end{aligned}$$

Here, passage **immediately** affects the number of salmon returns *and* affects **future** numbers of returns.

In other words: salmon returns today depend today's stream passage conditions and *lags* of passage—*e.g.*, last year's passage, the year before last, etc...

Estimate *total* effects by summing lags' coefficients, *e.g.*, $\beta_1 + \beta_2 + \beta_3 + \beta_4$.

Model options

Option 2a: Dynamic models with lagged explanatory variables

These models allow the outcome to depend upon the explanatory variable(s) in other periods.

$$\begin{aligned}\textbf{salmon}_t =& \beta_0 + \beta_1 \textbf{passage}_t + \beta_2 \textbf{passage}_{t-1} + \\ & \beta_3 \textbf{passage}_{t-2} + \beta_4 \textbf{passage}_{t-3} + u_t\end{aligned}$$

Here, passage **immediately** affects the number of salmon returns *and* affects **future** numbers of returns.

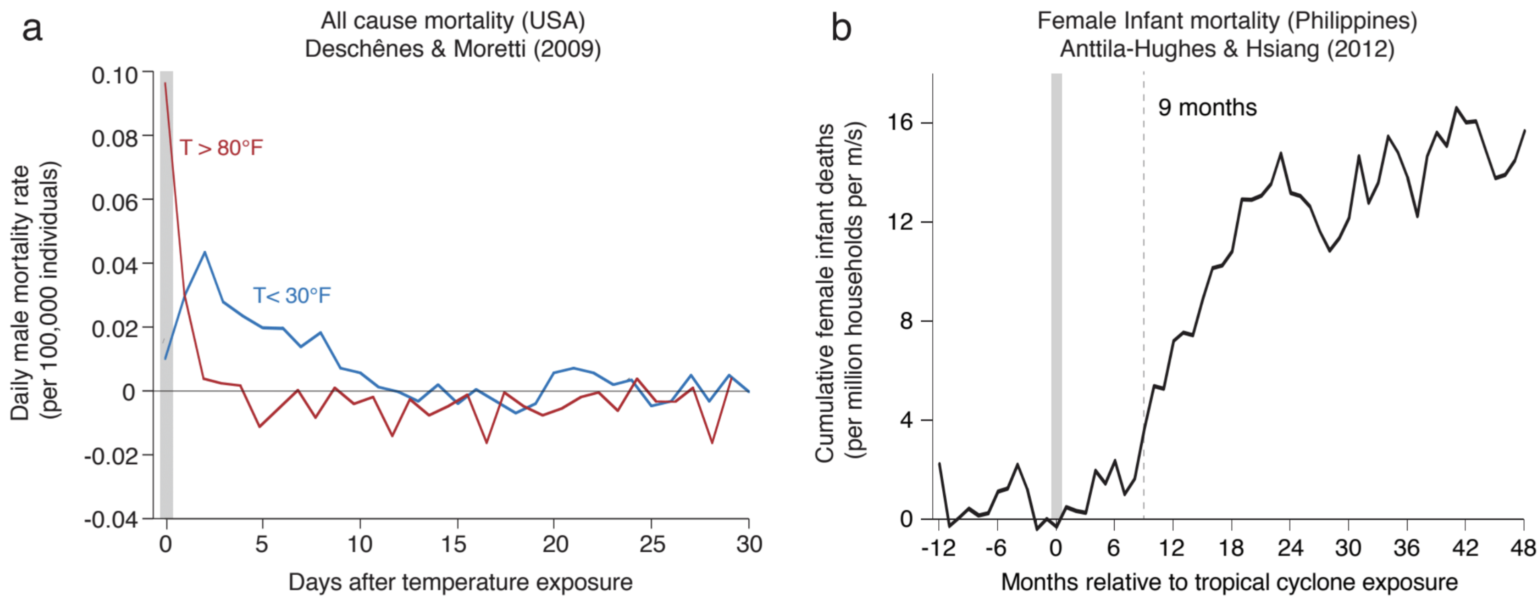
In other words: salmon returns today depend today's stream passage conditions and *lags* of passage—*e.g.*, last year's passage, the year before last, etc...

Estimate *total* effects by summing lags' coefficients, *e.g.*, $\beta_1 + \beta_2 + \beta_3 + \beta_4$.

Note: We still assume current salmon returns don't affect future returns.

Model options

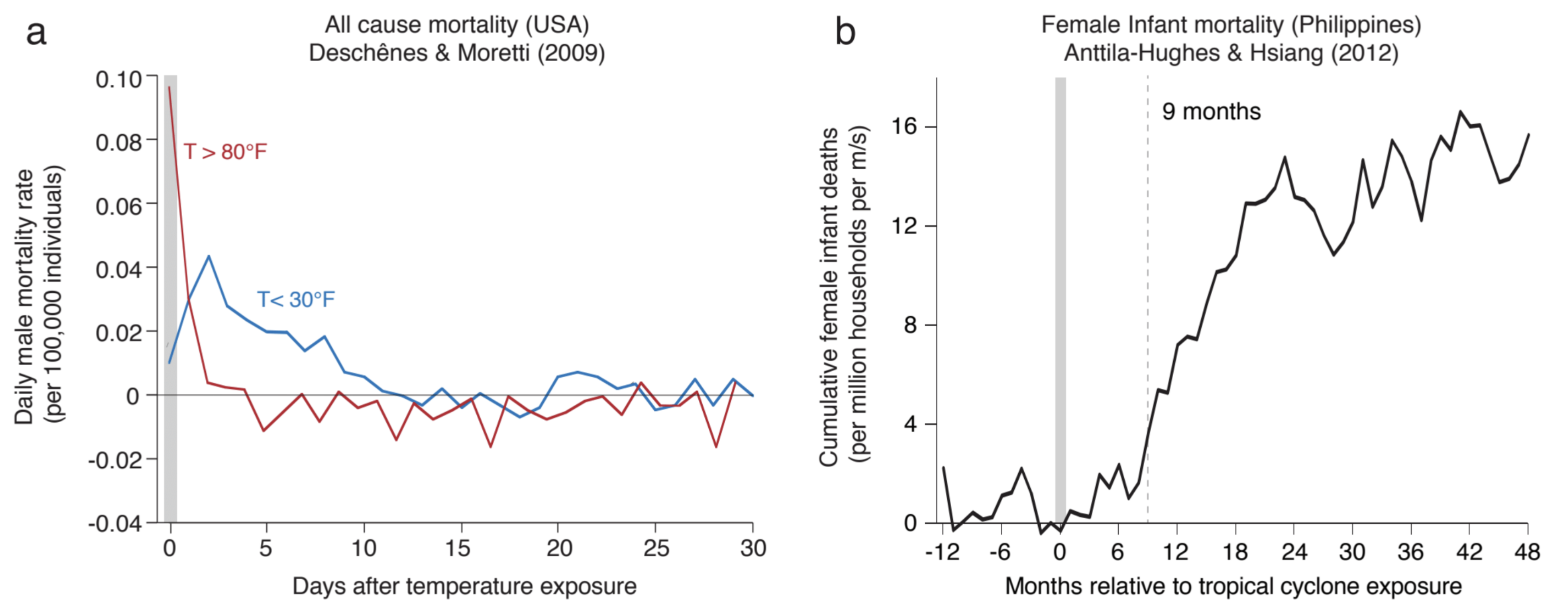
Lagged explanatory variables in empirical research:



- **Left:** coefficients on lagged temperature variables
- **Right:** sum of coefficients (cumulative effect) on cyclone intensity

Model options

Lagged explanatory variables in empirical research:



- **Left:** coefficients on lagged temperature variables
- **Right:** sum of coefficients (cumulative effect) on cyclone intensity

> Q: Can you think of other examples of lagged effects?

Model options

Option 2b: Autoregressive distributed-lag (ADL) models

These models allow the outcome to depend upon the explanatory variable(s) and/or the outcome variable in prior periods.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_3 \text{salmon}_{t-1} + u_t$$

Model options

Option 2b: Autoregressive distributed-lag (ADL) models

These models allow the outcome to depend upon the explanatory variable(s) and/or the outcome variable in prior periods.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_3 \text{salmon}_{t-1} + u_t$$

Here, current passage affects **current** salmon and **future** salmon.

Model options

Option 2b: Autoregressive distributed-lag (ADL) models

These models allow the outcome to depend upon the explanatory variable(s) and/or the outcome variable in prior periods.

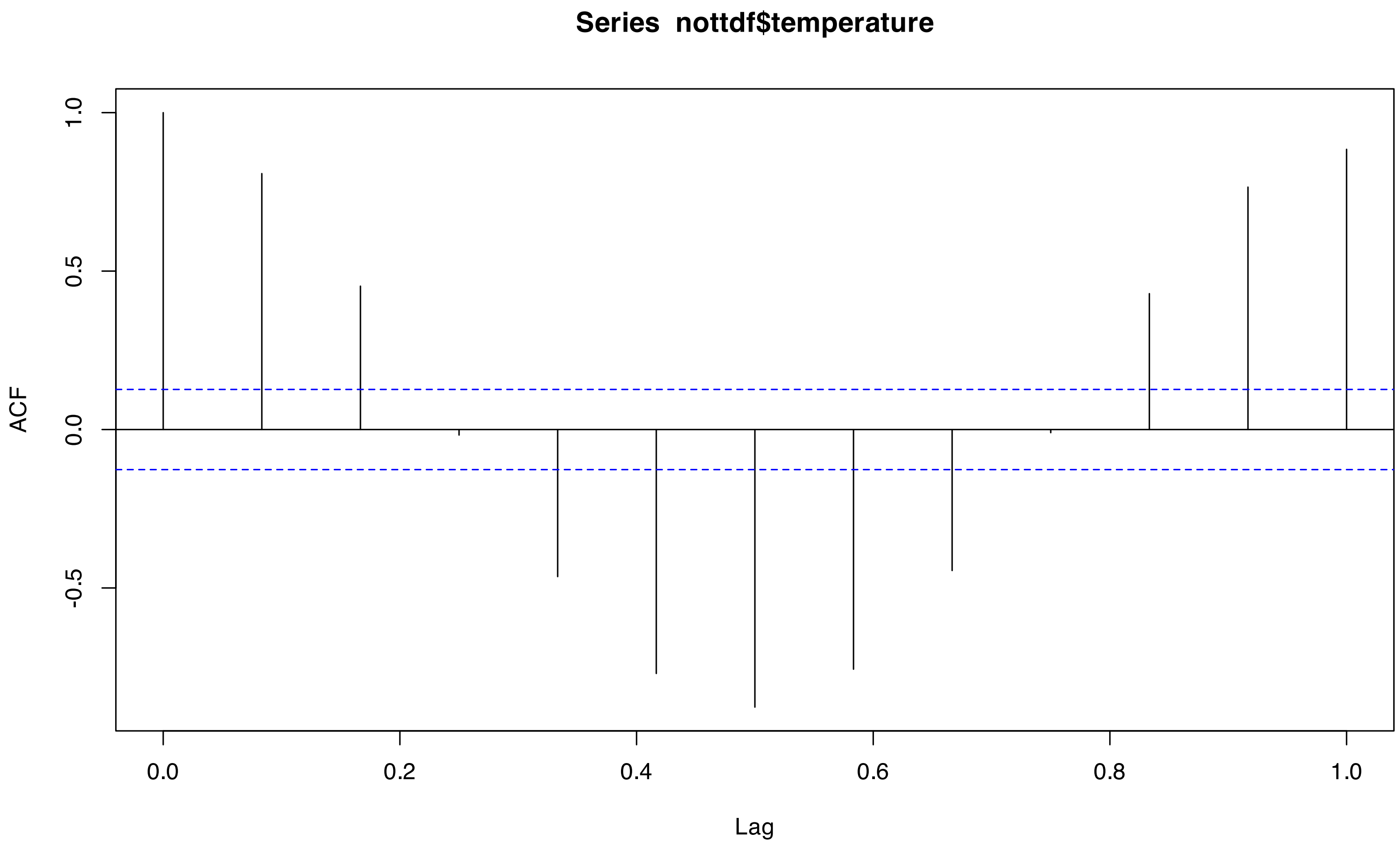
$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_3 \text{salmon}_{t-1} + u_t$$

Here, current passage affects **current** salmon and **future** salmon.

In addition, current salmon returns affect future salmon returns—we're allowing lags of the outcome variable.

Do you need an ADL?

Hint: Autocorrelation Function (ACF)



Autoregressive distributed-lag models

Numbers of lags

ADL models are often specified as $\mathrm{ADL}(p, q)$, where

- p is the (maximum) number of **lags** for the outcome variable.
- q is the (maximum) number of **lags** for explanatory variables.

Autoregressive distributed-lag models

Numbers of lags

ADL models are often specified as $\text{ADL}(p, q)$, where

- p is the (maximum) number of **lags** for the outcome variable.
- q is the (maximum) number of **lags** for explanatory variables.

Example: $\text{ADL}(1, 0)$

Autoregressive distributed-lag models

Numbers of lags

ADL models are often specified as $\text{ADL}(p, q)$, where

- p is the (maximum) number of **lags** for the outcome variable.
- q is the (maximum) number of **lags** for explanatory variables.

Example: $\text{ADL}(1, 0)$

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

Autoregressive distributed-lag models

Numbers of lags

ADL models are often specified as $\mathbf{ADL}(p, q)$, where

- p is the (maximum) number of **lags** for the outcome variable.
- q is the (maximum) number of **lags** for explanatory variables.

Example: $\mathbf{ADL}(1, 0)$

$$\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + \beta_2 \textbf{salmon}_{t-1} + u_t$$

Example: $\mathbf{ADL}(2, 2)$

Autoregressive distributed-lag models

Numbers of lags

ADL models are often specified as $\text{ADL}(p, q)$, where

- p is the (maximum) number of **lags** for the outcome variable.
- q is the (maximum) number of **lags** for explanatory variables.

Example: $\text{ADL}(1, 0)$

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

Example: $\text{ADL}(2, 2)$

$$\begin{aligned} \text{salmon}_t = &\beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_3 \text{passage}_{t-2} \\ &+ \beta_4 \text{salmon}_{t-1} + \beta_5 \text{salmon}_{t-2} + u_t \end{aligned}$$

Autoregressive distributed-lag models

Complexity

Due to their lags, ADL models actually estimate even more complex relationships than you might first guess.

Autoregressive distributed-lag models

Complexity

Due to their lags, ADL models actually estimate even more complex relationships than you might first guess.

Consider ADL(1, 0): $\textbf{salmon}_t = \beta_0 + \beta_1 \textbf{passage}_t + \beta_2 \textbf{salmon}_{t-1} + u_t$

Autoregressive distributed-lag models

Complexity

Due to their lags, ADL models actually estimate even more complex relationships than you might first guess.

Consider ADL(1, 0): $\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_2 \mathbf{salmon}_{t-1} + u_t$

Write out the model for period $t - 1$:

$$\mathbf{salmon}_{t-1} = \beta_0 + \beta_1 \mathbf{passage}_{t-1} + \beta_2 \mathbf{salmon}_{t-2} + u_{t-1}$$

Autoregressive distributed-lag models

Complexity

Due to their lags, ADL models actually estimate even more complex relationships than you might first guess.

Consider ADL(1, 0): $\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$

Write out the model for period $t-1$:

$$\text{salmon}_{t-1} = \beta_0 + \beta_1 \text{passage}_{t-1} + \beta_2 \text{salmon}_{t-2} + u_{t-1}$$

which we can substitute in for salmon_{t-1} in the first equation, *i.e.*,

$$\begin{aligned} \text{salmon}_t =& \beta_0 + \beta_1 \text{passage}_t + \\ & \beta_2 \underbrace{\left(\beta_0 + \beta_1 \text{passage}_{t-1} + \beta_2 \text{salmon}_{t-2} + u_{t-1} \right)}_{\text{salmon}_{t-1}} + u_t \end{aligned}$$

Complexity

Continuing...

$$
\begin{aligned}
\text{salmon}_t =& \beta_0 + \beta_1 \text{passage}_t + \\
& \beta_2 \underbrace{\left(\beta_0 + \beta_1 \text{passage}_{t-1} + \beta_2 \text{salmon}_{t-2} + u_{t-1}\right)}_{\text{salmon}_{t-1}} + u_t \\
=& \beta_0 \left(1 + \beta_2\right) + \beta_1 \text{passage}_t + \beta_1 \beta_2 \text{passage}_{t-1} + \\
& \beta_2^2 \text{salmon}_{t-2} + u_t + \beta_2 u_{t-1}
\end{aligned}
$$

Complexity

Continuing...

$$\begin{aligned}\mathbf{salmon}_t =& \beta_0 + \beta_1 \mathbf{passage}_t + \\ & \beta_2 \underbrace{\left(\beta_0 + \beta_1 \mathbf{passage}_{t-1} + \beta_2 \mathbf{salmon}_{t-2} + u_{t-1}\right)}_{\mathbf{salmon}_{t-1}} + u_t \\ =& \beta_0 \left(1 + \beta_2\right) + \beta_1 \mathbf{passage}_t + \beta_1 \beta_2 \mathbf{passage}_{t-1} + \\ & \beta_2^2 \mathbf{salmon}_{t-2} + u_t + \beta_2 u_{t-1}\end{aligned}$$

We could then substitute in the equation for $\mathbf{salmon}_{t-2}$, $\mathbf{salmon}_{t-3}$, ...

Complexity

Eventually we arrive at

$$\begin{aligned}
\text{salmon}_t =& \beta_0 \left(1 + \beta_2 + \beta_2^2 + \beta_2^3 + \cdots\right) + \\
& \beta_1 \left(\text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_2^2 \text{passage}_{t-2} + \cdots\right) + \\
& u_t + \beta_2 u_{t-1} + \beta_2^2 u_{t-2} + \cdots
\end{aligned}$$

Complexity

Eventually we arrive at

$$\begin{aligned}\text{salmon}_t = &\beta_0 \left(1 + \beta_2 + \beta_2^2 + \beta_2^3 + \cdots\right) + \\ &\beta_1 \left(\text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_2^2 \text{passage}_{t-2} + \cdots\right) + \\ &u_t + \beta_2 u_{t-1} + \beta_2^2 u_{t-2} + \cdots\end{aligned}$$

The point?

Complexity

Eventually we arrive at

$$\begin{aligned}\text{salmon}_t =& \beta_0 \left(1 + \beta_2 + \beta_2^2 + \beta_2^3 + \cdots\right) + \\ & \beta_1 \left(\text{passage}_t + \beta_2 \text{passage}_{t-1} + \beta_2^2 \text{passage}_{t-2} + \cdots\right) + \\ & u_t + \beta_2 u_{t-1} + \beta_2^2 u_{t-2} + \cdots\end{aligned}$$

The point?

By including just **one lag of the dependent variable**—as in a ADL(1, 0)—we implicitly include *many lags* of the explanatory variables and disturbances.†

† These lags enter into the equation in a very specific way—not the most flexible specification.

Time-series models

Updated OLS assumptions

1. **New: Weakly persistent outcomes**—essentially, x_{t+k} in the distant period $t+k$ is weakly correlated with period x_t (when k is "big").

2. y_t is a **linear function** of its parameters and disturbance.

3. There is **some variation** in our explanatory variables

4. **Harder to satisfy:** The u_t have conditional mean of zero (**exogeneity**), $\boldsymbol{E}[u_t|X]=0$.

5. **Harder to satisfy:** The u_t are **normally distributed** and **homoskedastic** with **zero correlation** between u_t and u_s, *i.e.*, $u_t \overset{\text{iid}}{\sim} N\left(0,\, \sigma^2\right)$, $\mathrm{Var}(u_t|X)=\mathrm{Var}(u_t)=\sigma^2$, and $\mathrm{Cor}(u_t,\, u_s|X)=0$.

Unbiased coefficients

As before, the unbiased-ness of OLS is going to depend upon our exogeneity assumption, *i.e.*, $\boldsymbol{E}[u_t|X] = \mathbf{0}$.

Unbiased coefficients

As before, the unbiased-ness of OLS is going to depend upon our exogeneity assumption, *i.e.*, $\boldsymbol{E}[u_t|X] = \mathbf{0}$.

We can split this assumption into two parts.

Unbiased coefficients

As before, the unbiased-ness of OLS is going to depend upon our exogeneity assumption, *i.e.*, $\boldsymbol{E}[u_t|X] = 0$.

We can split this assumption into two parts.

1. The disturbance u_t is independent of the explanatory variables in the **same period** (*i.e.*, X_t).

Unbiased coefficients

As before, the unbiased-ness of OLS is going to depend upon our exogeneity assumption, *i.e.*, $\boldsymbol{E}[u_t|X] = 0$.

We can split this assumption into two parts.

1. The disturbance u_t is independent of the explanatory variables in the **same period** (*i.e.*, X_t).

2. The disturbance u_t is independent of the explanatory variables in the **other periods** (*i.e.*, X_s for $s \neq t$).

Unbiased coefficients

As before, the unbiased-ness of OLS is going to depend upon our exogeneity assumption, *i.e.*, $\boldsymbol{E}[u_t|X] = 0$.

We can split this assumption into two parts.

1. The disturbance u_t is independent of the explanatory variables in the **same period** (*i.e.*, X_t).

2. The disturbance u_t is independent of the explanatory variables in the **other periods** (*i.e.*, X_s for $s \neq t$).

We need both of these parts to be true for OLS to be unbiased.

Unbiased coefficients

We need both parts of our exogeneity assumption for OLS to be unbiased:

I.e., to guarantee the numerator equals zero, we need $\boldsymbol{E}[u_t|X] = \mathbf{0}$—for both $\boldsymbol{E}[u_t|X_t] = 0$ *and* $\boldsymbol{E}[u_t|X_s] = 0$ $(s \neq t)$.

Unbiased coefficients

We need both parts of our exogeneity assumption for OLS to be unbiased:

I.e., to guarantee the numerator equals zero, we need $\boldsymbol{E}[u_t|X] = 0$—for both $\boldsymbol{E}[u_t|X_t] = 0$ *and* $\boldsymbol{E}[u_t|X_s] = 0$ $(s \neq t)$.

The second part of our exogeneity assumption—requiring that u_t is independent of all regressors in other periods—fails with dynamic models with lagged outcome variables.

Unbiased coefficients

We need both parts of our exogeneity assumption for OLS to be unbiased:

I.e., to guarantee the numerator equals zero, we need $\boldsymbol{E}[u_t|X] = 0$—for both $\boldsymbol{E}[u_t|X_t] = 0$ *and* $\boldsymbol{E}[u_t|X_s] = 0$ $(s \neq t)$.

The second part of our exogeneity assumption—requiring that u_t is independent of all regressors in other periods—fails with dynamic models with lagged outcome variables.

Thus, **OLS is biased for dynamic models with lagged outcome variables**.

Unbiased coefficients

To see why dynamic models with lagged outcome variables violate our exogeneity assumption, consider two periods of our simple ADL(1, 0) model.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t \tag{1}$$

$$\text{salmon}_{t+1} = \beta_0 + \beta_1 \text{passage}_{t+1} + \beta_2 \text{salmon}_t + u_{t+1} \tag{2}$$

Unbiased coefficients

To see why dynamic models with lagged outcome variables violate our exogeneity assumption, consider two periods of our simple ADL(1, 0) model.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t \tag{1}$$

$$\text{salmon}_{t+1} = \beta_0 + \beta_1 \text{passage}_{t+1} + \beta_2 \text{salmon}_t + u_{t+1} \tag{2}$$

In (1), u_t clearly correlates with salmon_t.

Unbiased coefficients

To see why dynamic models with lagged outcome variables violate our exogeneity assumption, consider two periods of our simple ADL(1, 0) model.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t \tag{1}$$

$$\text{salmon}_{t+1} = \beta_0 + \beta_1 \text{passage}_{t+1} + \beta_2 \text{salmon}_t + u_{t+1} \tag{2}$$

In (1), u_t clearly correlates with salmon_t.

However, salmon_t is a regressor in (2) (lagged dependent variable).

Unbiased coefficients

To see why dynamic models with lagged outcome variables violate our exogeneity assumption, consider two periods of our simple ADL(1, 0) model.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t \tag{1}$$

$$\text{salmon}_{t+1} = \beta_0 + \beta_1 \text{passage}_{t+1} + \beta_2 \text{salmon}_t + u_{t+1} \tag{2}$$

In (1), u_t clearly correlates with salmon_t.

However, salmon_t is a regressor in (2) (lagged dependent variable).

$\therefore$ The disturbance in t (u_t) correlates with a regressor in $t+1$ (salmon_t).

Unbiased coefficients

To see why dynamic models with lagged outcome variables violate our exogeneity assumption, consider two periods of our simple ADL(1, 0) model.

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t \tag{1}$$

$$\text{salmon}_{t+1} = \beta_0 + \beta_1 \text{passage}_{t+1} + \beta_2 \text{salmon}_t + u_{t+1} \tag{2}$$

In (1), u_t clearly correlates with salmon_t.

However, salmon_t is a regressor in (2) (lagged dependent variable).

$\therefore$ The disturbance in t (u_t) correlates with a regressor in $t+1$ (salmon_t).

This correlation violates the second part of our exogeneity requirement.

Unbiased coefficients

All is not lost.

Unbiased coefficients

All is not lost.

If we have **contemporaneous exogeneity**, OLS is what we call **consistent**: as $T \to \infty$, $\hat{\beta} \to \beta$ (you need a lot of data!)

Contemporaneous exogeneity: each disturbance is uncorrelated with the explanatory variables in the same period, *i.e.*,

$$\boldsymbol{E}[u_t|X_t] = 0$$

Unbiased coefficients

All is not lost.

If we have **contemporaneous exogeneity**, OLS is what we call **consistent**: as $T \to \infty$, $\hat{\beta} \to \beta$ (you need a lot of data!)

Contemporaneous exogeneity: each disturbance is uncorrelated with the explanatory variables in the same period, *i.e.*,

$$\boldsymbol{E}[u_t|X_t] = 0$$

With contemporaneous exogeneity, OLS estimates for the coefficients in a time series model are **consistent** (whew)

Autocorrelation in the error term

The time series version of our assumption about OLS errors includes the following:

> There must be **zero correlation** between u_t and u_s, *i.e.*, $\text{Cor}(u_t,\, u_s|X) = 0$.

Autocorrelation in the error term

The time series version of our assumption about OLS errors includes the following:

> There must be **zero correlation** between u_t and u_s, *i.e.*, $\text{Cor}(u_t,\, u_s | X) = 0$.

When might this fail?

- Anytime you have unobserved variables that correlate over time and influence the outcome

Autocorrelation in the error term

The time series version of our assumption about OLS errors includes the following:

> There must be **zero correlation** between u_t and u_s, *i.e.*, $\text{Cor}(u_t,\ u_s|X) = 0$.

When might this fail?

- Anytime you have unobserved variables that correlate over time and influence the outcome

Are we worried? In a static model or with lagged explanatory variables:

- OLS is **inefficient**, i.e., no longer the lowest variance unbiased estimator
- That is, your standard errors are no longer correct
- However, violating this assumption does not introduce bias (whew!)

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_2 \mathbf{salmon}_{t-1} + u_t$$

where

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

Autocorrelation

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

where

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

Problem:

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\mathbf{salmon}_t = \beta_0 + \beta_1 \mathbf{passage}_t + \beta_2 \mathbf{salmon}_{t-1} + u_t$$

where

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

Problem: Both $\mathbf{salmon}_{t-1}$ (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

where

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

Problem: Both salmon_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?

Autocorrelation

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

where

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

Problem: Both salmon_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?

A: It violates contemporaneous exogeneity

Autocorrelation

OLS and lagged outcome variables

Consider a model with one lag of the outcome variable—ADL(1, 0)—model with AR(1) disturbances

$$\text{salmon}_t = \beta_0 + \beta_1 \text{passage}_t + \beta_2 \text{salmon}_{t-1} + u_t$$

where

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

Problem: Both salmon_{t-1} (a regressor in the model for time t) and u_t (the disturbance for time t) depend upon u_{t-1}. *I.e.*, a regressor is correlated with its contemporaneous disturbance.

Q: Why is this a problem?

A: It violates contemporaneous exogeneity, *i.e.*, $\text{Cov}(x_t,\, u_t) \neq 0$.

Testing for serial/autocorrelation

- Fortunately, it's **easy to test for autocorrelation** to evaluate whether your model is biased (lagged dependent variable) and/or inefficient (lagged explanatory variables)

Testing for serial/autocorrelation

- Fortunately, it's **easy to test for autocorrelation** to evaluate whether your model is biased (lagged dependent variable) and/or inefficient (lagged explanatory variables)

- Basic idea:

    - Run OLS using your preferred specification
    - Recover residuals $e_t = y_t - \hat{y}_t$
    - Test whether $\hat{\theta}$ is statistically distinguishable from zero in

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots$$

    - Implement in `R` with: `dwtest()`, `bgtest()`

Testing for serial/autocorrelation

- Fortunately, it's **easy to test for autocorrelation** to evaluate whether your model is biased (lagged dependent variable) and/or inefficient (lagged explanatory variables)

- Basic idea:

  - Run OLS using your preferred specification
  - Recover residuals $e_t = y_t - \hat{y}_t$
  - Test whether $\hat{\theta}$ is statistically distinguishable from zero in

$$e_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots$$

  - Implement in `R` with: `dwtest()`, `bgtest()`

- Autocorrelation may arise because your model is **misspecified**. Consider adding additional lags and/or explanatory variables if errors are correlated

Summary: Time series and OLS

- Our model now has t subscripts for **time periods**.
- **Dynamic models** allow **lags** of explanatory and/or outcome variables.
- We changed our **exogeneity** assumption to **contemporaneous exogeneity**, *i.e.*, $\boldsymbol{E}[u_t|X_t] = 0$
- Including **lags of outcome variables** can lead to **biased coefficient estimates** from OLS (but fortunately they are still **consistent**)
- **Lagged explanatory variables** make **OLS inefficient** (i.e., mess up our standard errors)
- **Autocorrelation in the error + lagged dependent variables** make **OLS biased**. Watch out! Test for serial/autocorrelation, check for misspecification of your model.

Slides created via the R package **xaringan**.

Some slide components were borrowed from Ed Rubin and Allison Horst.