Environmental Statistics

Week 5: More on Time Series — Temporal Correlation and Changepoints

Jafet Belmont and Craig Wilkie

Recap

- Last week, we introduced time series data and discussed methods for separating seasonality and trend.
- We also looked at smoothing techniques that allowed us to describe non-linear trends.
- This week, we will spend some time looking at models for autocorrelation.
- We will also look at ways of assessing changepoints in our time series.

Fitting Additive Models in R

Reminder: Additive Models

- Additive models allow us to incorporate smooth functions alongside linear terms.
- These models can be used extensively for environmental data where we have one or more non-linear trend.
- These models take the form

$$y_i = \alpha + \sum_{j=1}^{k} g_j(x_{ij}) + \epsilon_{ij}.$$

- Here $g_j()$ is a function for the jth explanatory variable and α is the overall mean.

Additive Models in R

- The R package `mgcv` was designed to allow extensions of generalised linear models (GLMs).
- Most relevant to this course is the ability to fit **generalised additive models** (GAMs).
- The *generalised* aspect means that we can also extend the standard additive model to situations where we have non-normal responses, but we will not focus on these in this course.

Fitting the Model

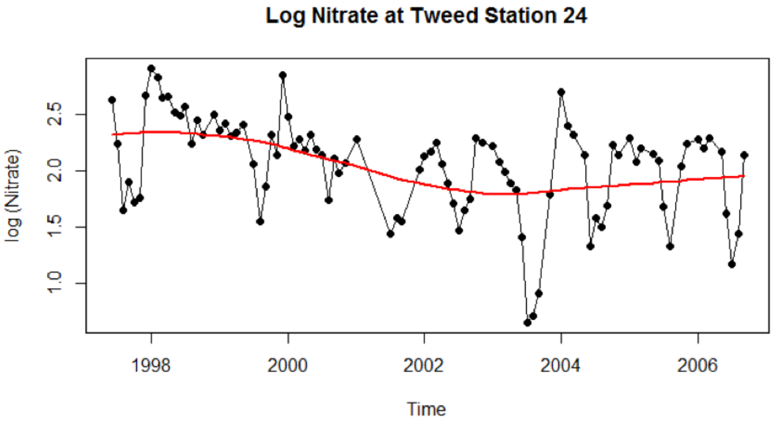
- We use the function `gam()` to fit our model. This works in a very similar manner to the `lm()` function.
- The smooth functions are represented by `s()`. These use the penalised splines approach described last week.
- Any linear terms can be included additively as normal.
- The model will take the form below, where you can include as many smooth or linear terms as you wish:

```
library(mgcv)

mod <- gam(response ~ s(smooth1) + s(smooth2) + linear)
```

Example: River nitrate level

- The nitrate levels in the River Tweed were measured monthly between 1997 and 2007.
- The red line is a simple LOWESS curve.



Fitting a model in R

```
> m1 <- gam(log_nitrate ~ s(Date))
> summary(m1)

Family: Gaussian
Link function: identity

Parametric coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   2.04454    0.03965    51.56     <2e-16

Approximate significance of smooth terms:
          edf   Ref.df     F    p-value
s(Date) 6.183   7.336   4.37   0.000272

R-sq.(adj) = 0.242   Deviance explained = 29.3%

GCV = 0.15847   Scale est. = 0.14623   n = 93
```

University of Glasgow

Interpreting the output

- We are mainly interested in the output related to smooth terms:

```
Approximate significance of smooth terms:
           edf  Ref.df     F    p-value
s(Date) 6.183   7.336   4.37   0.000272
```

- The p-value tells us the significance of the term, i.e., whether the smooth term is significantly different from a flat (horizontal) line.
- The p-value **doesn’t** tell us whether the smooth term is different from a linear term.
- The effective degrees of freedom (EDF) tells us how nonlinear the relationship is:
  - Higher EDF means a more nonlinear relationship.
  - An EDF of 1 indicates a linear relationship.

Interpreting the output

- We are mainly interested in the output related to smooth terms:

```
Approximate significance of smooth terms:
            edf   Ref.df      F     p-value
s(Date) 6.183    7.336    4.37    0.000272
```

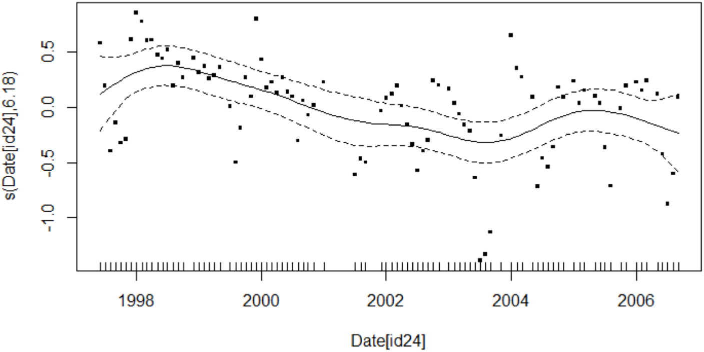
- In our example, the p-value is very small ($<< 0.05$), and therefore we have evidence that this smooth term is necessary in our model.
- The EDF for this term is 6.183, suggesting that this is far from linear and that a smooth term may be appropriate.
- We don't need to test for nonlinearity, since the model will penalise excess wiggliness, effectively fitting a linear term where appropriate.

Visualising our smooth function

- We can simply use the `plot` function to visualise our smooth function:

```
> plot(m1)
```

- Here we observe that we may have a bimodal shape, with peaks in 1999 and 2005.



Likelihood ratio test for significance

- We can also assess the significance of our smooth term using the `anova` function.
- We fit a simple linear regression and compare it to the additive model we have already fitted. The p-value confirms that the smooth term is necessary.

```
> m1 <- gam(log_nitrate ~ s(Date))
> m2 <- lm(log_nitrate ~ Date)

> anova(m2, m1)

Analysis of Variance Table
Model 1: log_nitrate ~ Date
Model 2: log_nitrate ~ s(Date)

     Res.Df    RSS    Df    SS      F   Pr(>F)
1    91.000 14.883
2    85.817 12.549 5.183 2.334 3.0794 0.01228
```

Autocorrelation

Autocorrelation

- We already know that environmental data are often measured over time, and that consecutive measurements are often related.
- This relationship between adjacent observations is known as **autocorrelation**.
- The term *autocorrelation* literally means *correlation with oneself*. Here, we can think of it as each point being correlated with ‘previous’ versions of itself.
- The strength of autocorrelation tends to be related to how far apart points are in time (known as **lag**). Points closer together have more in common than those further apart.

Why is autocorrelation a problem?

- Many statistical models rely on an assumption that our observations (more specifically our error terms) are independent.
- If we have correlation, then each observation ‘shares’ some information with other observations.
- This means that we have less independent information within our dataset and the *effective sample size* of the dataset will decrease.
- When we are calculate standard errors, confidence intervals etc., we are using the ‘wrong’ value of n.
- This can lead to us underestimate the variance and be overconfident in our results.

Measuring autocorrelation

- We can estimate the strength of temporal dependence using a sample **autocorrelation function (ACF)**.

- This function represents the autocorrelation of the data at a series of different lags in time.

- Assuming that we have a regularly spaced time series, we compute the sample ACF at lag k as

$$r(k) = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

(where $\bar{x}$ is the sample mean).

- We compute this for values from $k = 1, \ldots, K$, where K is some sensible maximum lag.

Measuring autocorrelation

- The ACF at lag 1, $r(1)$, is the correlation between the original data (lag 0) and the lag 1 data.
- $r(2)$ is the correlation between lag 0 and lag 2.

	A	B	C	D	E	F
1	DATE	lag0	lag1	lag2	lag3	lag4
2	Jan-95	278.44				
3	Feb-95	282.84	278.44			
4	Mar-95	289.15	282.84	278.44		
5	Apr-95	285.86	289.15	282.84	278.44	
6	May-95	282.03	285.86	289.15	282.84	278.44
7	Jun-95	278.55	282.03	285.86	289.15	282.84
8	Jul-95	276.52	278.55	282.03	285.86	289.15
9	Aug-95	275.33	276.52	278.55	282.03	285.86
10	Sep-95	274.24	275.33	276.52	278.55	282.03
11	Oct-95	274.14	274.24	275.33	276.52	278.55
12	Nov-95	274.90	274.14	274.24	275.33	276.52
13	Dec-95	276.33	274.90	274.14	274.24	275.33
14	Jan-96	277.37	276.33	274.90	274.14	274.24
15	Feb-96	279.66	277.37	276.33	274.90	274.14
16	Mar-96	284.81	279.66	277.37	276.33	274.90
17	Apr-96	285.63	284.81	279.66	277.37	276.33
18	May-96	280.81	285.63	284.81	279.66	277.37
19	Jun-96	278.70	280.81	285.63	284.81	279.66
20	Jul-96	275.57	278.70	280.81	285.63	284.81

Uncertainty for the ACF

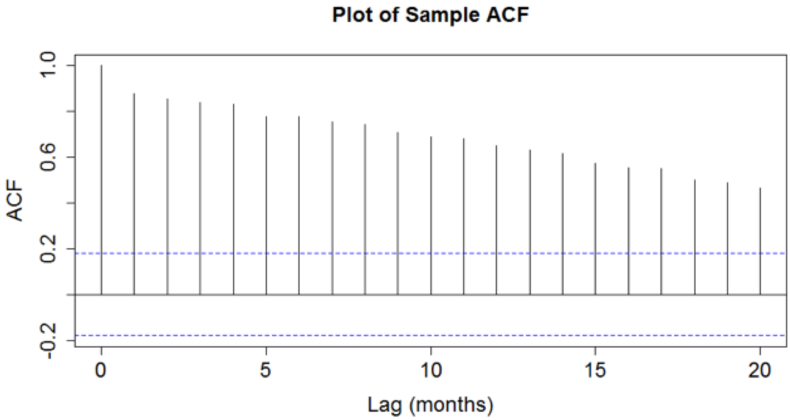
- Our sample ACF is an estimate of the overall ACF. So, we must consider uncertainty.
- Typically, we will compute a simple confidence interval around our point estimate at each lag as:

$$r(k) \pm 1.96\sqrt{\frac{1}{n}}$$

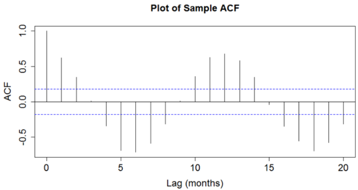
  where n is the number of observations in the time series.
- We plot the ACF using separate vertical lines for the size of the correlation at each lag, with dashed lines for the confidence intervals.
- If the lines lie within the confidence intervals, no autocorrelation is present.

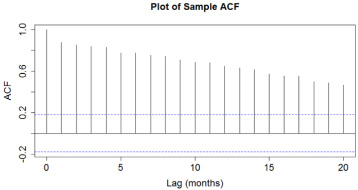
ACF Plots

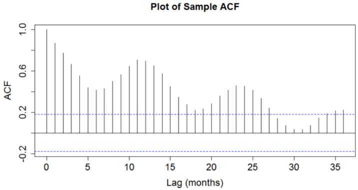
- Here, we have several lines outwith the confidence intervals. So, we have statistically significant evidence of autocorrelation in this dataset.



ACF Examples

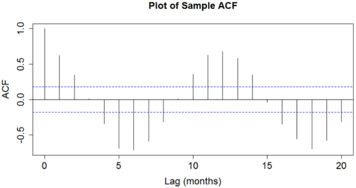


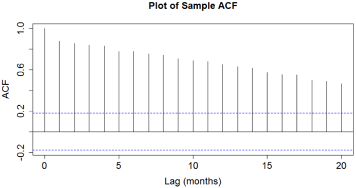


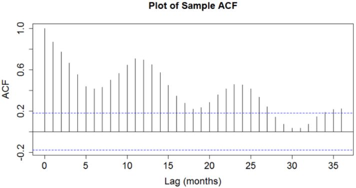


- Each of these three ACFs suggest that autocorrelation is present.

ACF Examples

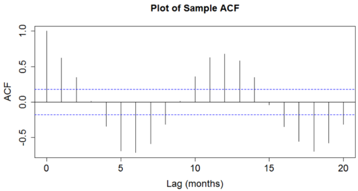


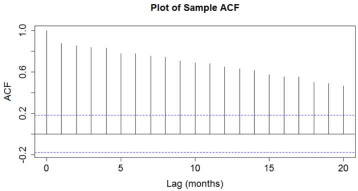


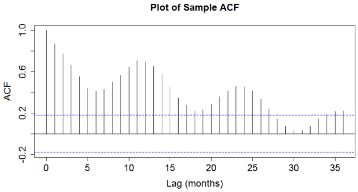


- Each of these three ACFs suggest that autocorrelation is present.
- *Top:* shows a repeating pattern — suggests seasonality.

ACF Examples

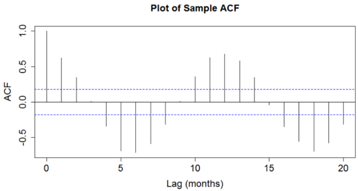


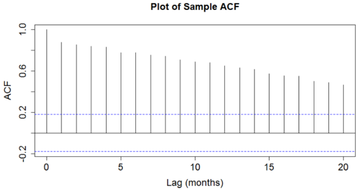


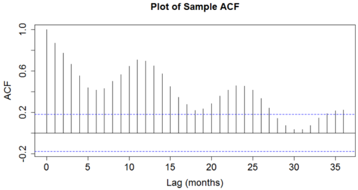


- Each of these three ACFs suggest that autocorrelation is present.
- *Top:* shows a repeating pattern — suggests seasonality.
- *Middle:* has a decreasing pattern — likely caused by trend.

ACF Examples

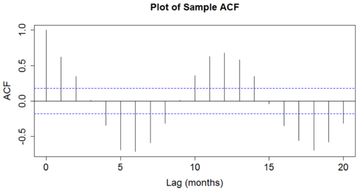


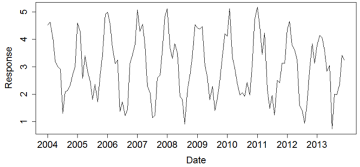


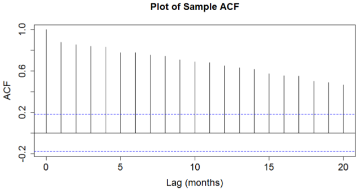


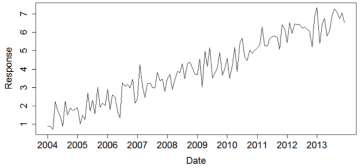
- Each of these three ACFs suggest that autocorrelation is present.
- *Top:* shows a repeating pattern — suggests seasonality.
- *Middle:* has a decreasing pattern — likely caused by trend.
- *Bottom:* has both patterns — probably seasonality AND trend.

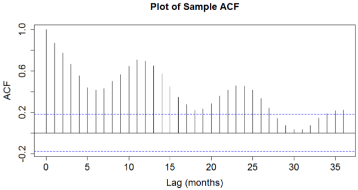
ACF Examples

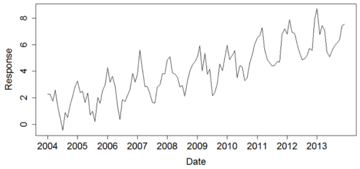












Dealing with autocorrelation

- If we have identified autocorrelation in our data, we have to find a way to account for it in our model.
- In some cases we may choose to simply treat it as a nuisance, and make adjustments to our standard errors to reflect the reduced effective sample size.
- The alternative is to explicitly account for the autocorrelation in our model.
  - For seasonal patterns, we may be able to eliminate it using methods discussed previously, such as harmonics.
  - For other types of autocorrelation, we may use approaches such as autoregressive integrated moving average (ARIMA).

Reminder: Assumptions

- Our examples look at autocorrelation in the data.
- Remember that we are assuming that the *errors* are independent.
- We must check for autocorrelation in the residuals *after* fitting a model.

Autocorrelation models

- **Autoregressive integrated moving average** (ARIMA) models are a general class of models that account for autocorrelation.

- These models combine aspects of two main classes of model: autoregressive (AR) and moving average (MA).

- Broadly speaking, AR(p) models assume that the current value is a function of the previous p observations.

- In contrast, MA(q) models assume that the current value can be computed by a linear regression on the q previous random error terms.

- These models are covered in more detail in the Time Series course, but will be addressed briefly here.

AR model

- An autoregressive (AR) model accounts for correlation by describing each value as a function of the previous values.
- The AR(p) process can be written as

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t.$$

- Here, ϕ_i is the 'autoregressive parameter' that measures the strength of the autocorrelation.
- $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ is simply random error, often referred to as noise.

MA model

- A moving average (MA) model accounts for correlation by describing each value as a function of the previous set of error terms.

- The MA(q) process can be written as

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i} + \epsilon_t.$$

- Here μ is the mean of the series and θ_i is the regression parameter associated with the ith lag.

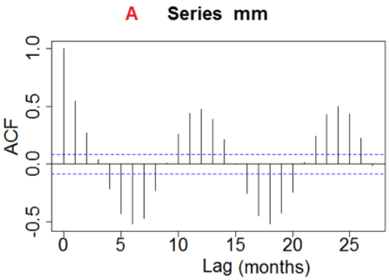
ARIMA

- The ARIMA is a combination of AR and MA processes.
- The I stands for *Integrated*, which relates to ‘differencing’, i.e. replacing a value with the difference between itself and the previous value.
- We write this model as ARIMA(p, d, q), where p is the order of the AR process, d is the degree of differencing and q is the order of the MA process.
- For example, ARIMA(1,0,0) would be equivalent to an AR(1) model and ARIMA(0,0,1) is an MA(1) model.

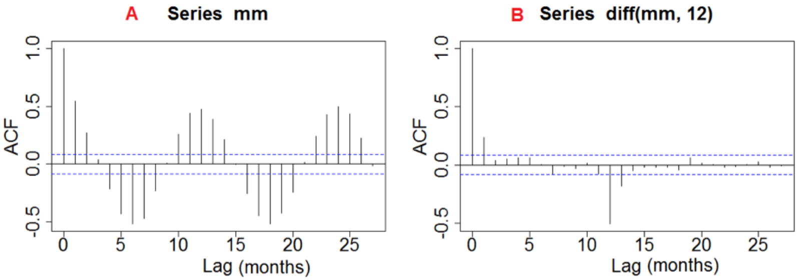
What model to fit?

- We can use the sample ACF to suggest the appropriate model to account for our autocorrelation.
- A smooth decay suggests that we have AR components.
- A less structured ACF might suggest that an MA is more appropriate.
- In practice, AR processes are less complex than MA processes and tend to be used more frequently as a result.

Examples

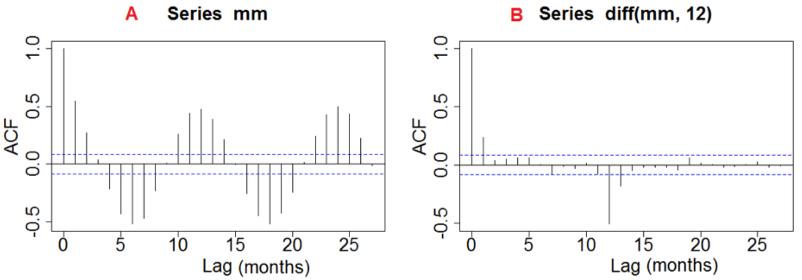

A
Series mm
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
20
25
Lag (months)
B
Series diff(mm, 12)
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
20
25
Lag (months)

Examples



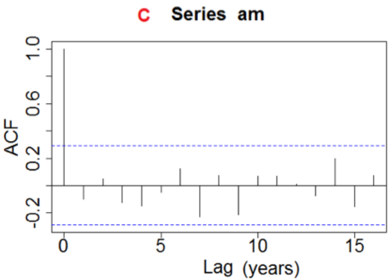
- Plot A has a clear seasonal pattern. Harmonics may be more appropriate than an ARIMA model.

Examples

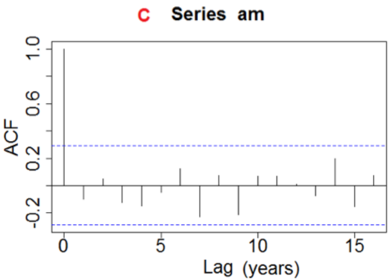


- Plot A has a clear seasonal pattern. Harmonics may be more appropriate than an ARIMA model.
- In Plot B, the value at lag 1 is outside the interval, and thus an AR(1) may be most suitable. (Note that we can probably ignore the spike at lag 12 as just random error.)

Examples

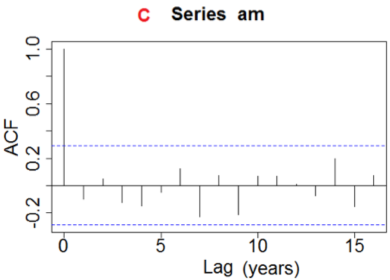


Examples



- Plot C does not appear to show any correlations outwith the error bars, and so we can conclude that there is no evidence of autocorrelation.

Examples



- Plot C does not appear to show any correlations outwith the error bars, and so we can conclude that there is no evidence of autocorrelation.
- Note that the bars are wider than in plots A and B. This is probably because we had less data available.

ARIMA in R

- We can use the `arima()` function in R to explore autocorrelation.
- We must first fit a linear model, and then extract the design matrix to use as an input to this function.
- For example, to fit an AR(1) model, we would use the following code:

```
trend.model0 <- lm(response ~ decimal.date)

X <- model.matrix(trend.model0)

trend.model1 <- arima(y, order = c(1, 0, 0), xreg = X, include.mean = FALSE)
```

ARIMA in R

- ARIMA methods are all based on regularly spaced data (measurements equally spaced in time).
- However, in some cases, we may have irregularly spaced data.
- If the data are roughly regular (just a small deviation here and there), we may be able to treat them as though they are regular.
- If we have missing data, we may be able to impute or interpolate without too many issues.
- In cases where we have completely irregular data, we may need to use more complex statistical methods (which will not be covered in this course).

Changepoints

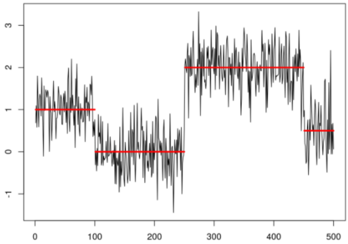
Changepoints

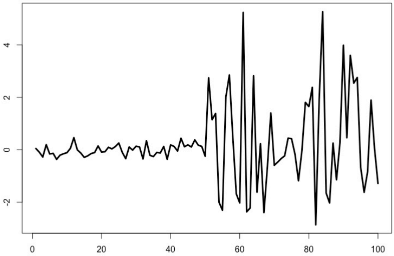
- One of the main reasons that we analyse environmental data is to detect changes.
- Sometimes, these changes occur organically, either as the result of some natural environmental process, or some non-deliberate human action.
- In some other occasions, these changes occur by design, as the result of a deliberate and controlled human action (e.g. policy changes).
- Regardless of the reason for the change, we want to understand more about when it happened and the extent of the change.

Changepoints

- In statistics, a **changepoint** is a point in time after which some or all of the model parameters might change.
- Most commonly, this is a change in mean or variance, but it could also be a change in some other feature of the data.
- We may not always know exactly when the changepoint occurs, or whether we have a changepoint at all.
- In some cases, we may have more than one changepoint.

Examples

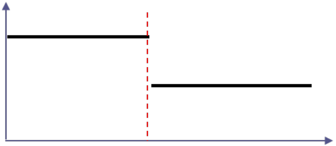




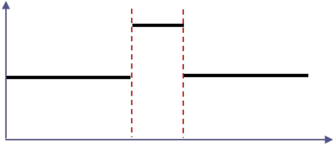
- Reasons for changepoints might include:
  - Environmental events, e.g. flooding, volcanic eruption.
  - Policy, e.g. low emissions zones, water quality regulations.
  - Changes to measuring equipment.

Some simple examples of changepoints include:

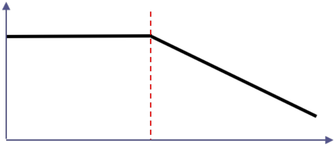
- A shift up (or down) of the mean.



- A short-term change in the mean.

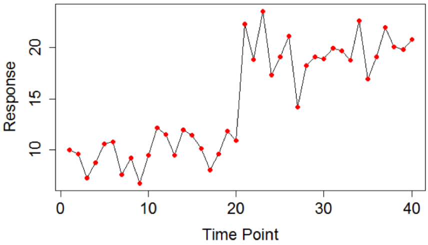


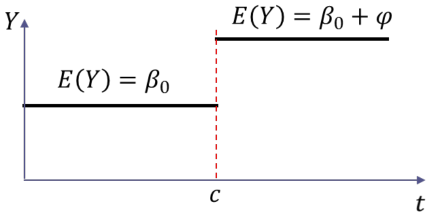
- A change in a model parameter, e.g. slope.



Known changepoint — mean shift

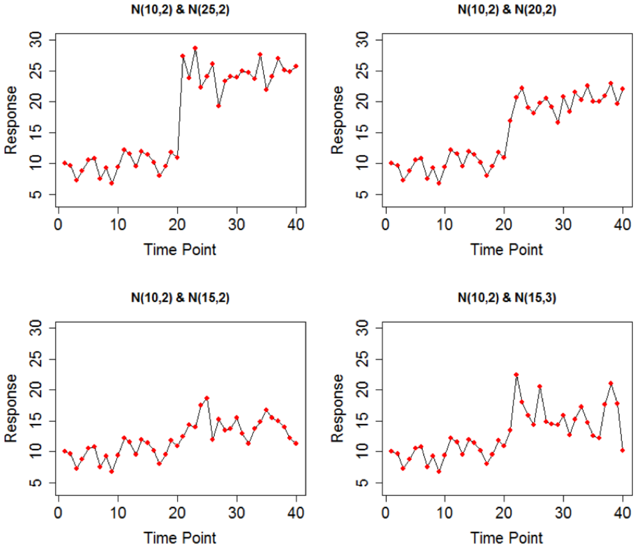
- Consider a series with two different mean levels.
- The first 20 observations come from N$(10, 1)$.
- The next 20 observations come from N$(15, 1)$.
- Our ability to detect this change depends on the size of the change and the variability in the data.





Known changepoint — mean shift

It can be difficult to distinguish changepoints from trend



Known changepoint — mean shift

- We have a series of data Y_i collected at a set of timepoints t_i with $i = 1, \ldots, n$.

- If our known changepoint is at time c, then we can construct an indicator function

$$\mathcal{I}_{t_i} = \begin{cases} 0 & \text{if } t_i < c \\ 1 & \text{if } t_i \geq c \end{cases}$$

- This can then be included as a parameter in our regression model

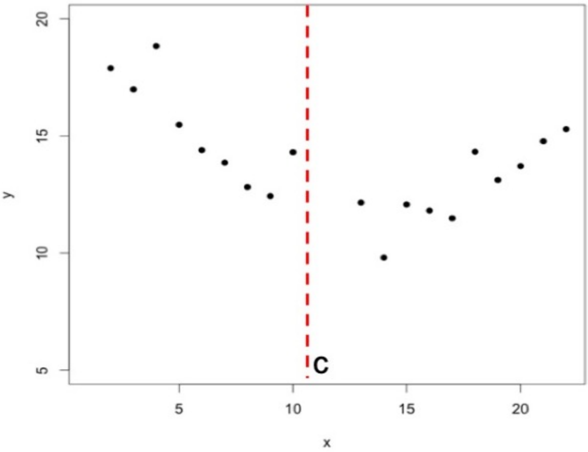
$$Y_i = \beta_0 + \varphi \mathcal{I}_{t_i} + \epsilon_i$$

Known changepoint — mean shift

- Here, φ, the coefficient of the indicator function, can be described as the **intervention effect**.
- It controls the size of the mean shift in our model. We have
  - $E(Y_i) = \beta_0$ before the changepoint
  - $E(Y_i) = \beta_0 + \varphi$ after the changepoint.
- If this parameter is significant in our model, that implies that we have a significant change in mean at timepoint *c*.

Known changepoint — change in slope

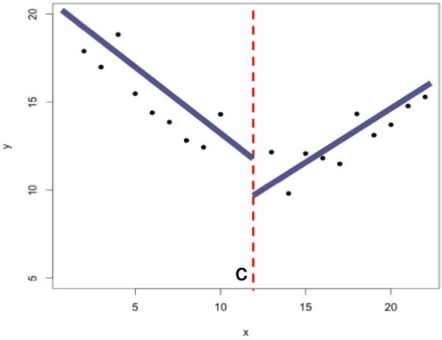
- We also need to consider examples where we observe a **change in slope** at a known timepoint.



Known changepoint — change in slope

- It would be possible to fit two separate regressions

$$Y_i = \alpha_1 + \beta_1 x_i + \epsilon_i \quad \text{for } x < c$$
$$Y_i = \alpha_2 + \beta_2 x_i + \epsilon_i \quad \text{for } x \geq c$$



- However, this seems quite simplistic, and it would be better to have a single continuous model.

Piecewise regression

- We want our regression to be continuous at c such that we have:

$$\alpha_1 + \beta_1 c = \alpha_2 + \beta_2 c$$

- This can be rewritten in terms of a single model parameter as:

$$\alpha_2 = \alpha_1 + c(\beta_1 - \beta_2)$$

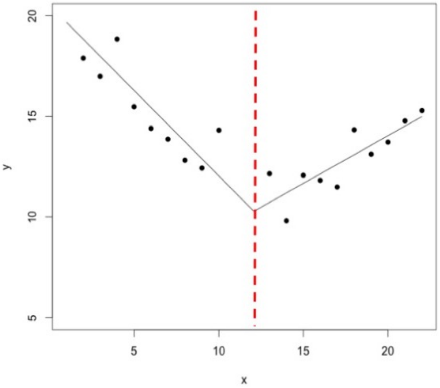
- We can thus update our equations to the following. This is known as **piecewise regression** (or segmented regression)

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 x_i + \epsilon_i && \text{for } x < c \\ Y_i &= \alpha_1 + (\beta_1 - \beta_2)c + \beta_2 x_i + \epsilon_i && \text{for } x \geq c \end{aligned}$$

(Note that this could be expressed as a single model using our indicator function.)

Known changepoint — change in slope

- The two linear parts of our model now meet at c.
- Note that our piecewise model is more efficient than two separate regressions, since it uses one fewer parameter (no α_2).

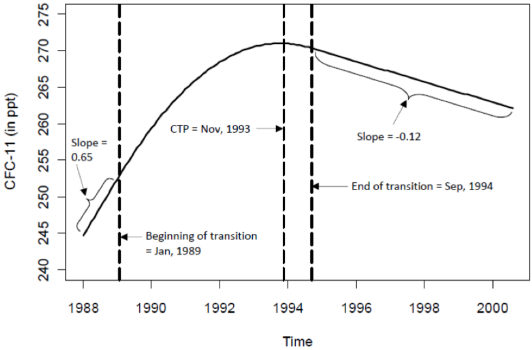


Known changepoint — more complex patterns

- In many cases, we may have more complex changes to our trend.
- There are a variety of more advanced models for known changepoints, but these are all based on the same underlying principles.
- For example, the **bent cable** model allows for an extended ‘transition phase’ between the two slopes, often represented by a smooth curve.
- This can often be more realistic than a sharp change in slope.

Example: Bent cable for CFC data

- Chlorofluorocarbons (CFCs) are pollutants which were often used in aerosols.
- Their use was phased out in the 1990s as a result of environmental policy. We can see this ‘phasing out’ period represented in the model.

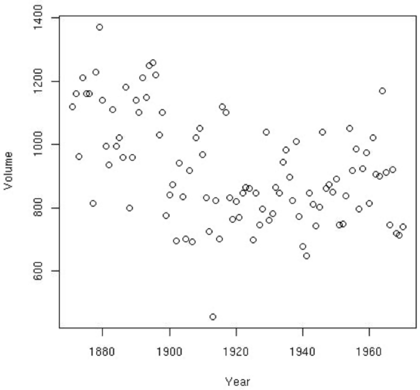


Unknown changepoint

- It can be more challenging to fit a changepoint model when you don't clearly know exactly when the change occurred.
- We could try to estimate it visually by looking at a plot, but it may be more appropriate to use statistical modelling.
- One of the most popular methods is an iterative approach, which searches across the entire range of our data for possible changepoints.
- This approach compares a series of piecewise models to a standard linear regression, and highlights whether any changepoints exist, and if so, how many.

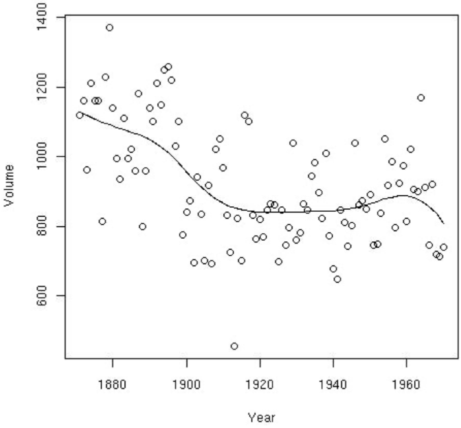
Example: River Nile

- We have historic data on the levels of the River Nile around the city of Aswan, Egypt.
- Is there any evidence of a change in water volume? If so, when did it occur?



Example: River Nile

- We can examine the data by fitting a LOWESS curve.
- There does appear to be a change around 1900. However, we need to explore this further via a model.



Fitting in R

- We use the `segmented()` function in R (in the package also called `segmented`) to fit an unknown changepoint model.

```
out.lm <- lm(Volume ~ Year)
mod <- segmented(out.lm, seg.Z = ~Year, psi = 1900)
```

- First, fit a standard regression using `lm()`.
- We then pass the linear model into our `segmented()` function along with an initial estimate of the changepoint.
- This initial estimate (`psi = 1900`) is used as a starting point for our iterative algorithm.

Example: River Nile

```
Estimated Break-Point(s):
psi1.x
  1913

slope(mod)
$x
          Est.  St.Err. t value
slope1 -8.1820   1.759  -4.650
slope2  0.7458   1.084   0.688
```

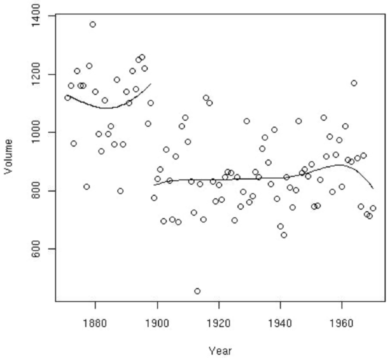
- The final model output suggests that the changepoint occurred in 1913.
- Prior to 1913, the volume was decreasing by 8.18 units per year. Afterwards, it was increasing by 0.75 units per year.

Example: River Nile

- The Aswan Low Dam was constructed between 1899–1902, massively impacting river levels in the area.
- Therefore, it is more sensible to fit a model that introduces a mean shift, rather than a change of slope.
- Subject matter expertise is key!

Example: River Nile

- In this case, given there is a clear reason why the time series will change either side of the dam's construction, we need to fit two separate models.
- The plot below shows two separate penalised spline models for the 'before' and 'after' periods.



Summary points

Summary points

- This relationship between adjacent observations in a time series is known as **autocorrelation**.

- We can estimate the strength of temporal dependence using a sample **autocorrelation function (ACF)**, defined for lag k (assuming a regularly spaced time series) as

$$r(k) = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

Summary points

- **Autoregressive integrated moving average** (ARIMA) models are a general class of models which account for autocorrelation.
- AR(p) models assume that the current value is a function of the previous p observations.
- MA(q) models assume that the current value can be computed by a linear regression on the q previous random error terms.
- The AR(p) process can be written as

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t.$$

- The MA(q) process can be written as

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t.$$

Summary points

- A **changepoint** is a point in time after which some or all of the model parameters might change.

- Some simple examples of changepoints include:
  - A shift up (or down) of the mean.
  - A short-term change in the mean.
  - A change in a model parameter, e.g. slope.

- We can model such data using **piecewise regression**, or the **bent cable** model.