Temporal Correlation and Changepoints

We begin this week by finishing our notes on additive models in R, and then move onto temporal correlation and changepoints.

1 Additive Models in R

1.1 Fitting Additive Models in R

We can fit GAMs in R using the package `mgcv`, which was designed to allow extensions of generalised linear models (GLMs). 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.

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 Tweed nitrate level

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

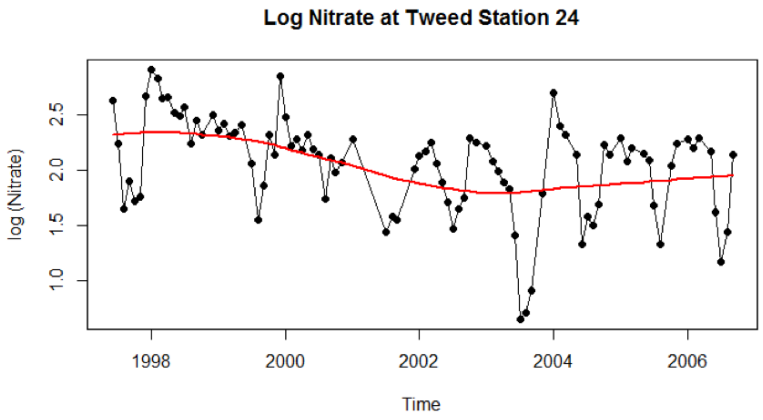


Figure 1: Log(Nitrate) by time for River Tweed. Black points are data. Red line is LOWESS curve.

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

```
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
```

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.

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 can 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:

```r
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
```

The p-value confirms that the smooth term is necessary instead of a linear term. (Note that we don't really need to test for nonlinearity, since the model should penalise excess wiggliness, effectively fitting a linear term where appropriate.)

1.2 Visualising additive models in R

Unlike linear terms, we can't simply report parameter estimates for smooth terms in additive models. We can instead simply use the `plot` function to visualise our smooth functions.

Example: River Tweed nitrate level (continued)

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

```
plot(m1)
```

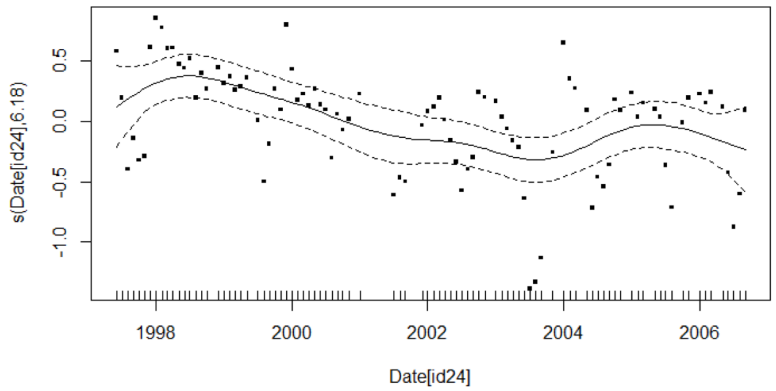


Figure 2: Plot of fitted smooth of time for River Tweed log(Nitrate) data. The black points are partial residuals, the black line is the estimated smooth, and the dashed lines are the 95% interval bounds for the estimated smooth.

In reality, we might spot that there is a seasonal pattern evident in the plot, which our smooth does not capture. We could capture this by either:

- fitting a smooth to the months plus a smooth to years (to capture the seasonal pattern plus long term smooth trend), as we saw at the end of last week's notes, or
- increasing the basis dimension for our smooth, using the `k` argument of the `s()` function. (See the help file of the `s` function in R for details, by running `?mgcv::s`.)

2 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.

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 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 underestimating the variance and being overconfident in our results.

2.1 Autocorrelation function

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

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

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

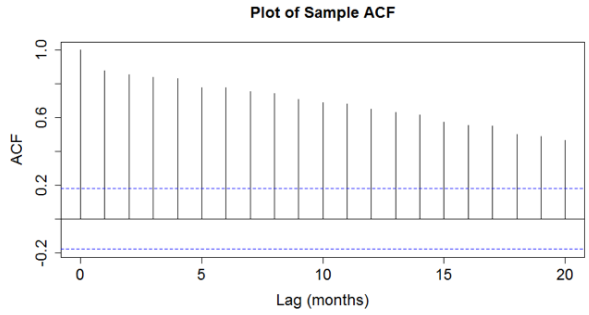
Our sample ACF is an estimate of the overall ACF, and as such we have to consider uncertainty. Typically we will compute a simple confidence interval around our point estimate at each lag as

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

where n is the number of observations in the time series.

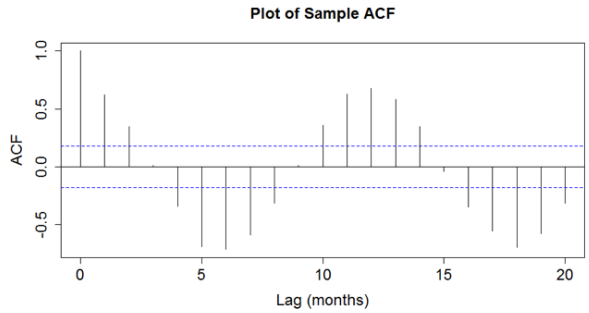
We plot the ACF function with a separate vertical line representing the size correlation at each lag, and dashed lines for the confidence intervals. If the lines lie within the confidence intervals, no autocorrelation is present.

Here, we have several lines outside the confidence intervals, so autocorrelation exists in this dataset.

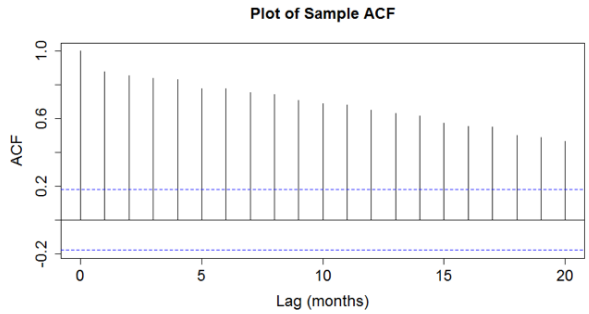


Each of the three ACFs below are examples of cases where suggest autocorrelation is present.

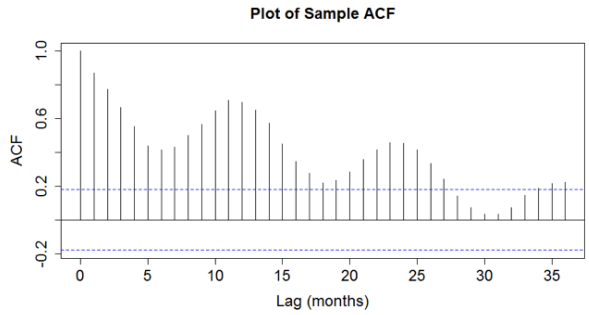
- The first has a repeating pattern — suggests seasonality:



- The second has a decreasing pattern — likely caused by trend:



- The third has both patterns — probably seasonality AND trend:



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).

2.2 Autocorrelation models

Autoregressive integrated moving average (ARIMA) models are a general class of models which 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.

2.2.1 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" which measures the strength of the autocorrelation. $\epsilon_t \sim \text{N}(0, \sigma^2)$ is simply random error, often referred to as noise.

2.2.2 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.

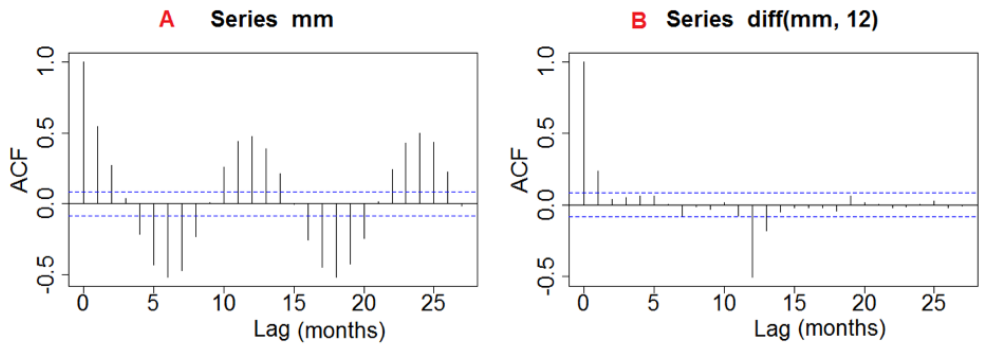
2.2.3 ARIMA

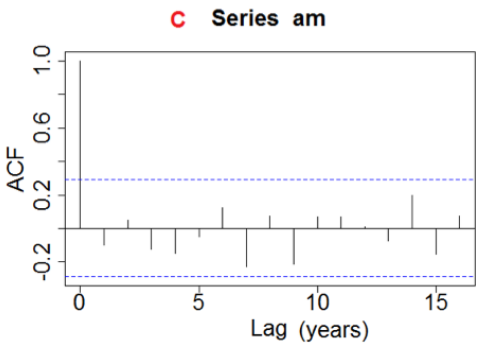
The ARIMA is a combination of AR and MA processes. The I stands for *Integrated*, which relates to "differencing", i.e., replacing a value with the difference between itself and the previous value. We write this model as ARIMA(p, d, q), where p is the order of the AR process, d is the degree of differencing and q is the order of the MA.

For example, ARIMA(1,0,0) would be equivalent to an AR(1) model and ARIMA(0,0,1) is an MA(1) model.

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 an MA is more appropriate. In practice, AR processes are less complex than MA and tend to be used more frequently as a result.

Examples of ACF plots





Plot A has a clear seasonal pattern, meaning that harmonics may be more appropriate than an ARIMA.
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 likely ignore the spike at lag 12 as just random error).
Plot C does not appear to have any correlations outside of the error bars, and so we can conclude that there is no evidence of autocorrelation. Note that the bars are wider. This is likely because we had less data available.

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).

2.2.4 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)
```

3 Changepoints

One of the main reasons we analyze 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.

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.

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.

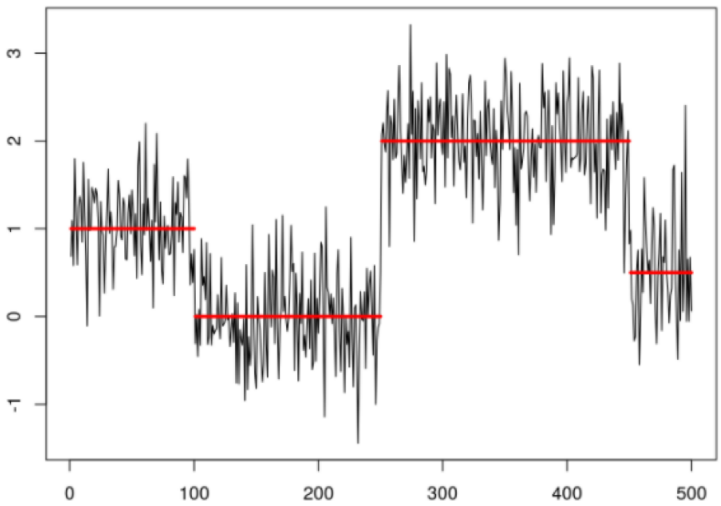


Figure 3: Plot showing shifts in mean values at changepoints. X-axis is time, y-axis is value, red lines represent the 4 mean levels, and black lines join the data points.

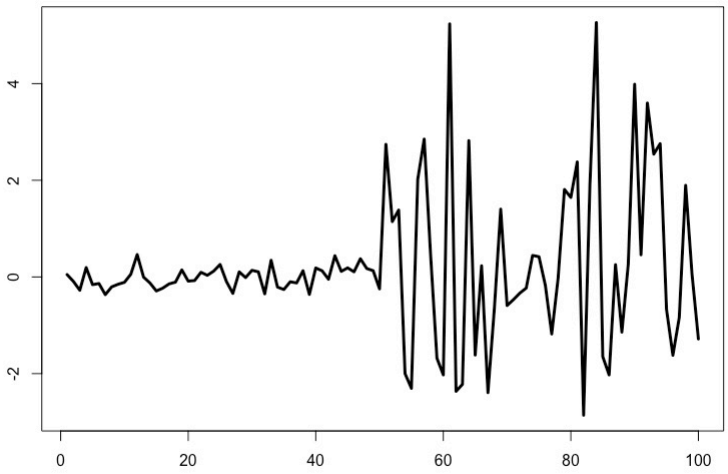


Figure 4: Plot showing shifts in variance at changepoints. X-axis is time, y-axis is value, and black lines join the data points.

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, eg slope.

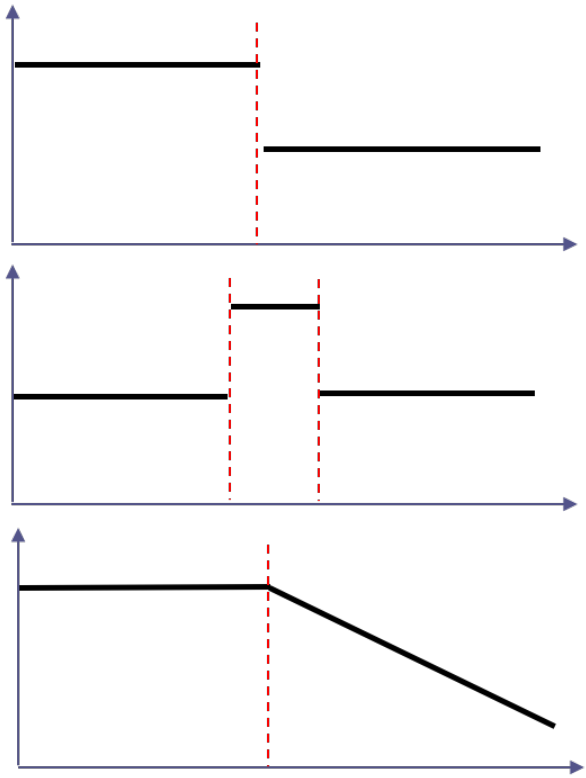
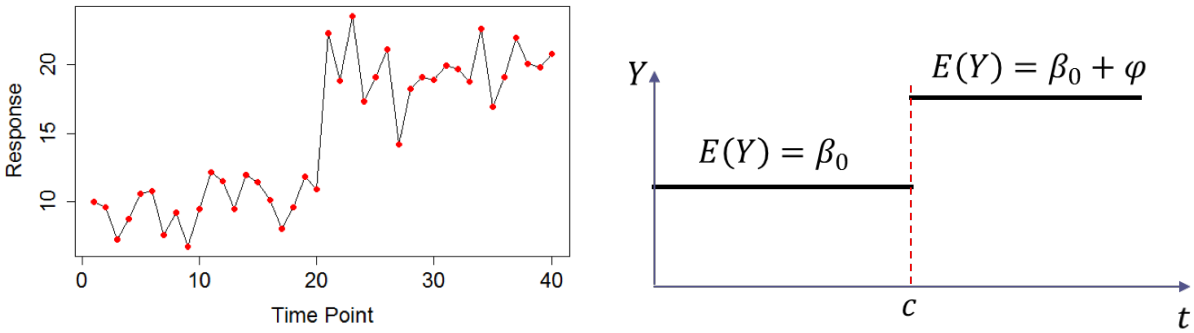


Figure 5: Illustration of the simple examples of changepoints above.

Consider a series with two different mean levels. The first 20 observations come from $\mathrm{N}(10, 1)$ and the next 20 observations come from $\mathrm{N}(15, 1)$. Our ability to detect this change depends on the size of the change and the variability in the data. The plots below illustrate these data and a possible model for these.



It can be difficult to distinguish changepoints from trend. The plots below illustrate how the magnitude of the shift in mean value can affect our ability to identify the shift in mean. The bottom right plot illustrates the additional effect of a change in variance.

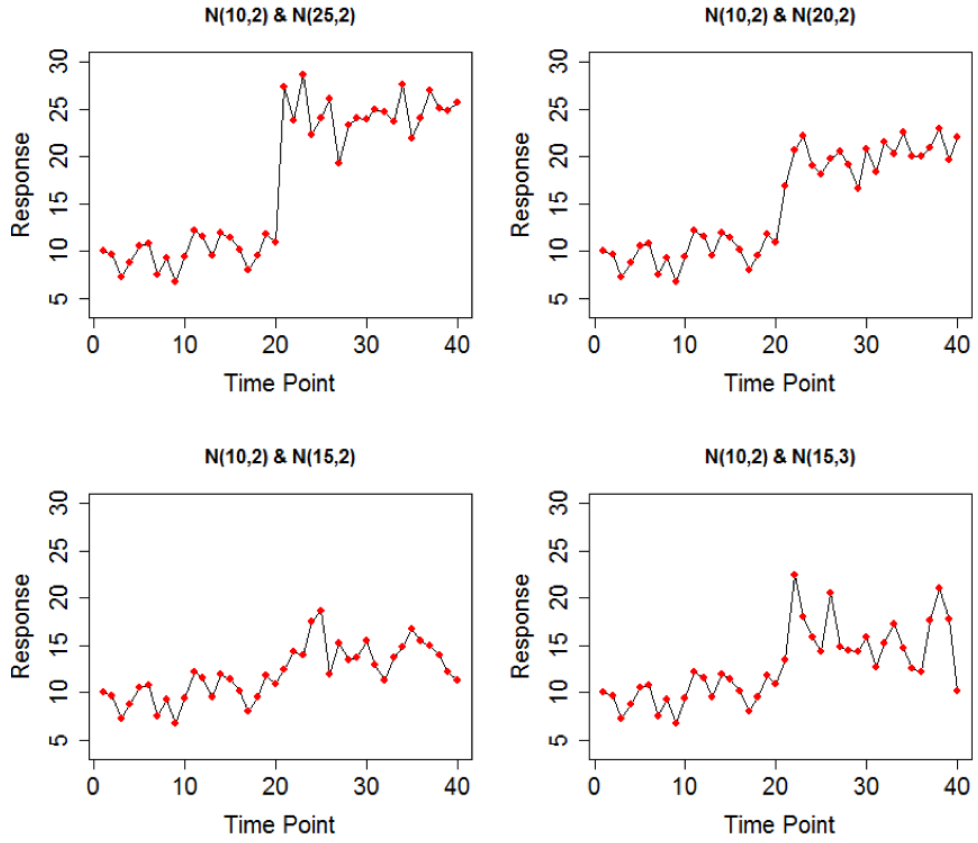


Figure 6: Plots of data (response value by time point) from different distributions with mean shifts at timepoint 20.

3.1 Known changepoint

Sometimes it is known that a change occurred at a specific timepoint, but the magnitude or shape of this change are not known.

3.1.1 Known changepoint — mean shift

Suppose that we have a series of data Y_i collected at a set of timepoints t_i with $i = 1, \dots, n$. If our known changepoint is at time c, then we can construct an indicator function

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

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

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

Here, φ, the coefficient of the indicator function ,can be described as the **intervention effect**. If this parameter is significant in our model, that implies that we have a significant change in mean at timepoint c.

3.1.2 Known changepoint — change in slope

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

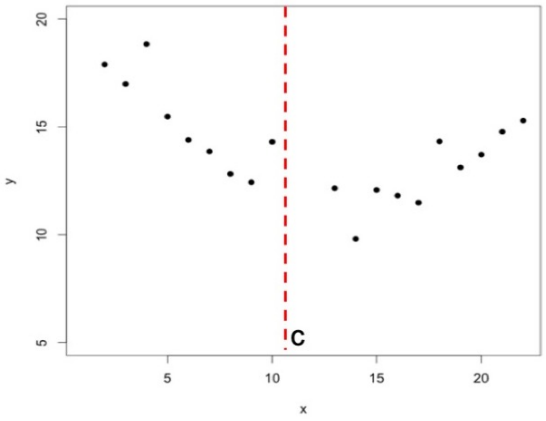


Figure 7: Plot of value by time, with data as black dots, and changepoint at time c illustrated by vertical dashed red line.

It would be possible to fit two separate regressions. However, this seems quite simplistic, and it would be better to have a single continuous model.

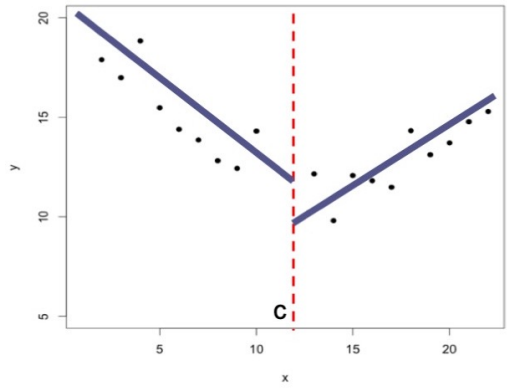


Figure 8: Plot of value by time, with data as black dots, and changepoint at time c illustrated by vertical dashed red line. Blue regression lines represent the fitted separate regression models.

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, which 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.)

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.

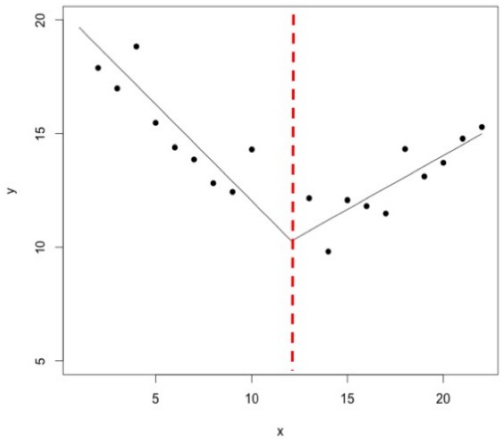


Figure 9: Plot of value by time, with data as black dots, and changepoint at time c illustrated by vertical dashed red line. Grey regression lines represent the fitted piecewise regression model.

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.

Example: Chlorofluorocarbons (CFCs)

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.

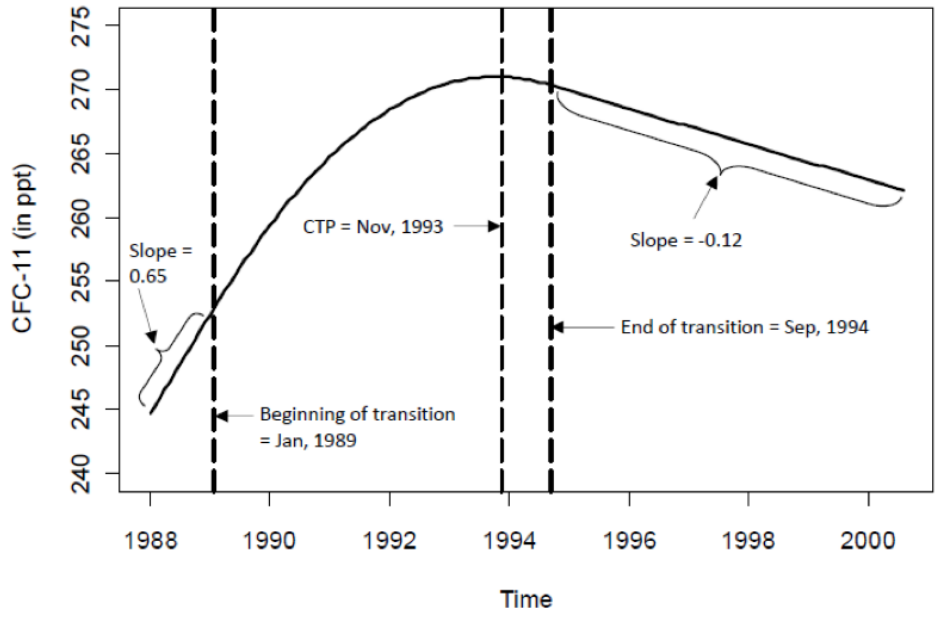


Figure 10: Diagram illustrating the bent cable model as fitted to CFC-11 (ppt) values between 1988 and 2000.

3.2 Unknown changepoint

It can be more challenging to fit a changepoint model when you don't clearly know exactly when the change occurred. One of the most popular methods is an iterative approach which searches across the entire range of our data for possible changepoints.

Example: River Nile flow data

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?

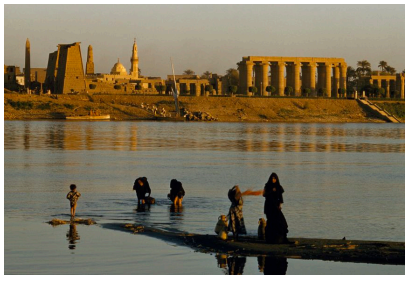


Figure 11: River Nile.

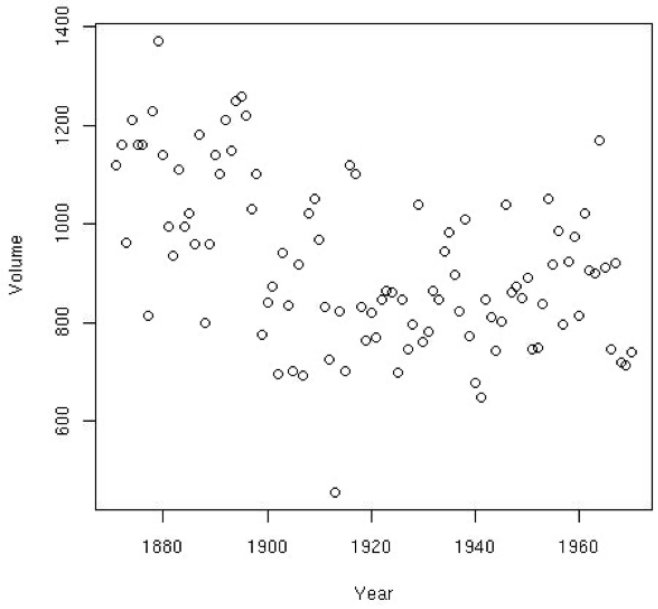


Figure 12: Plot of River Nile volume by year, Aswan.

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.

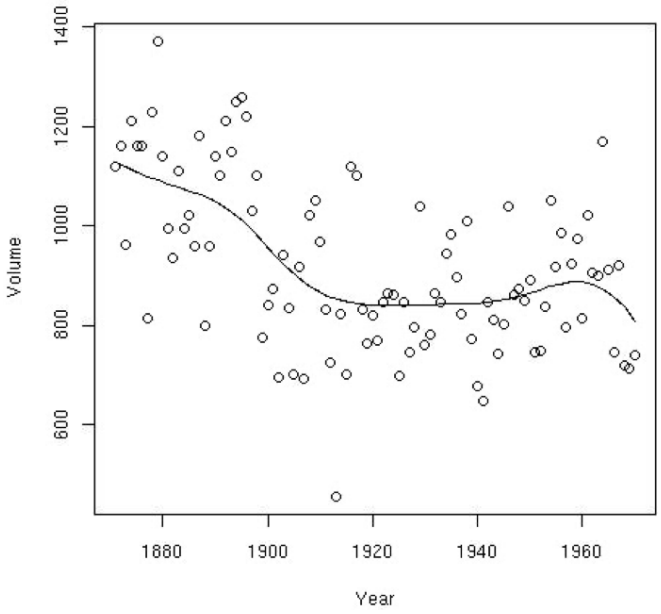


Figure 13: Fitted LOWESS curve for River Nile data.

We use the `segmented()` function in R (in the package also called `segmented`) to fit an unknown changepoint model, using the following steps:

- 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.

We run this in R:

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

```
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.
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 which introduces a mean shift, rather than a change of slope. Subject matter expertise is key!

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.

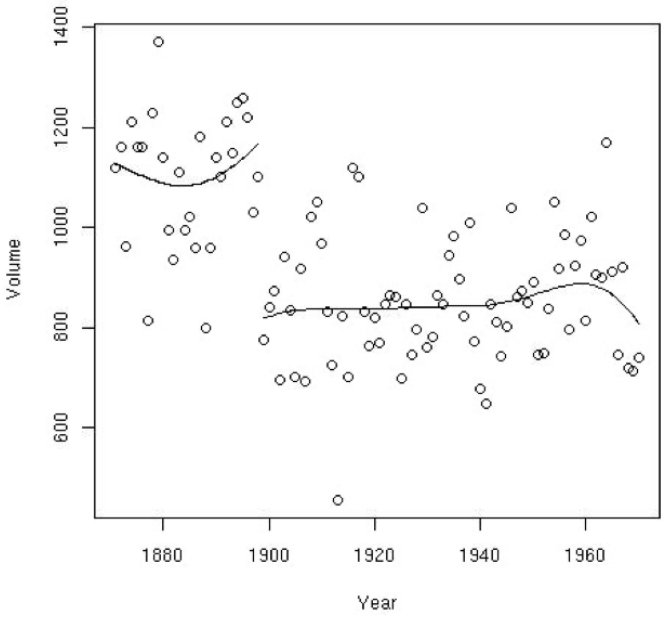


Figure 14: Fitted LOWESS curves for River Nile data (separate for before/after 1900).