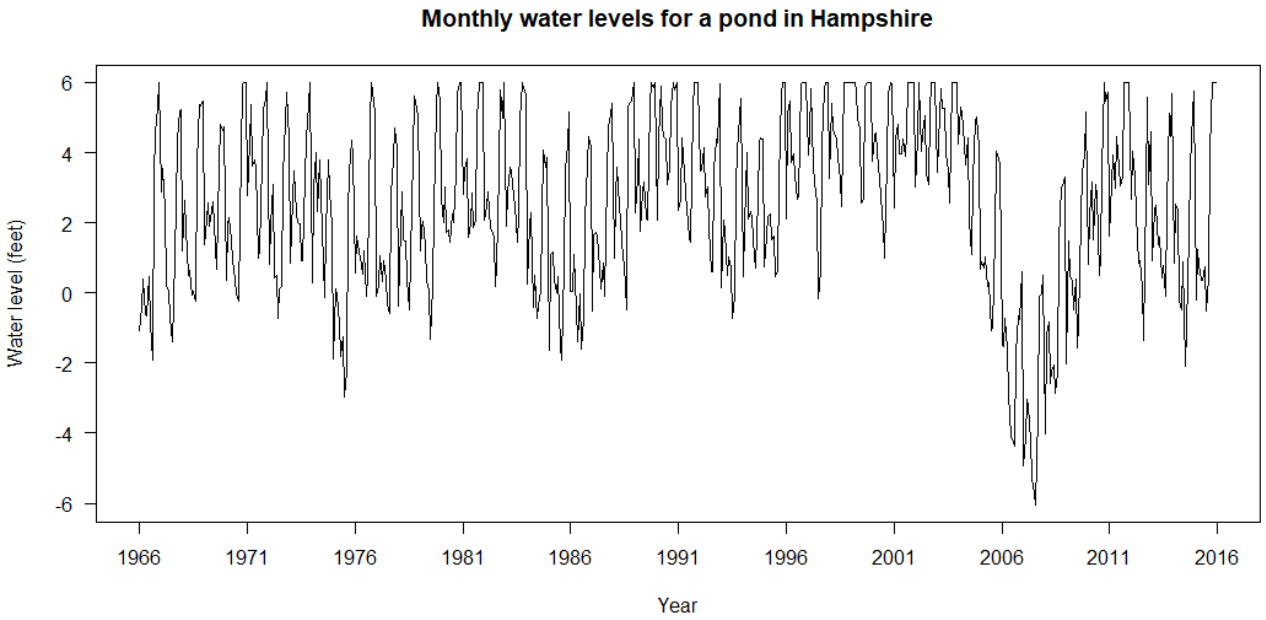
Investigating monthly water levels for a pond in rural Hampshire

Report summary

The aim of this report is to investigate the time series data of monthly water levels for a small pond in rural Hampshire. A linear trend and seasonal effects can be observed from the data. Removing these and analysing the residuals leads to an autoregressive model of order 2 being fitted. This means the data can be summarised as being a weighted average of its own past, plus a linear trend, weighted seasonal effects and some random noise. All of this will be discussed in more detail throughout the report.

First impressions of the dataset

Using R to plot the data, the following graph shows the monthly water levels for the pond:



The monthly water level in the dataset varies from -6 feet to 6 feet. This indicates that 0 must be some defined base level chosen at the start of record-keeping and +6 and -6 feet shows the fluctuations from the base level. The peaks regularly hit 6 feet but never exceed this amount and the troughs which usually hover around the -2 feet mark never goes below 6 feet. This indicates the maximum depth of the pond is 12 feet.

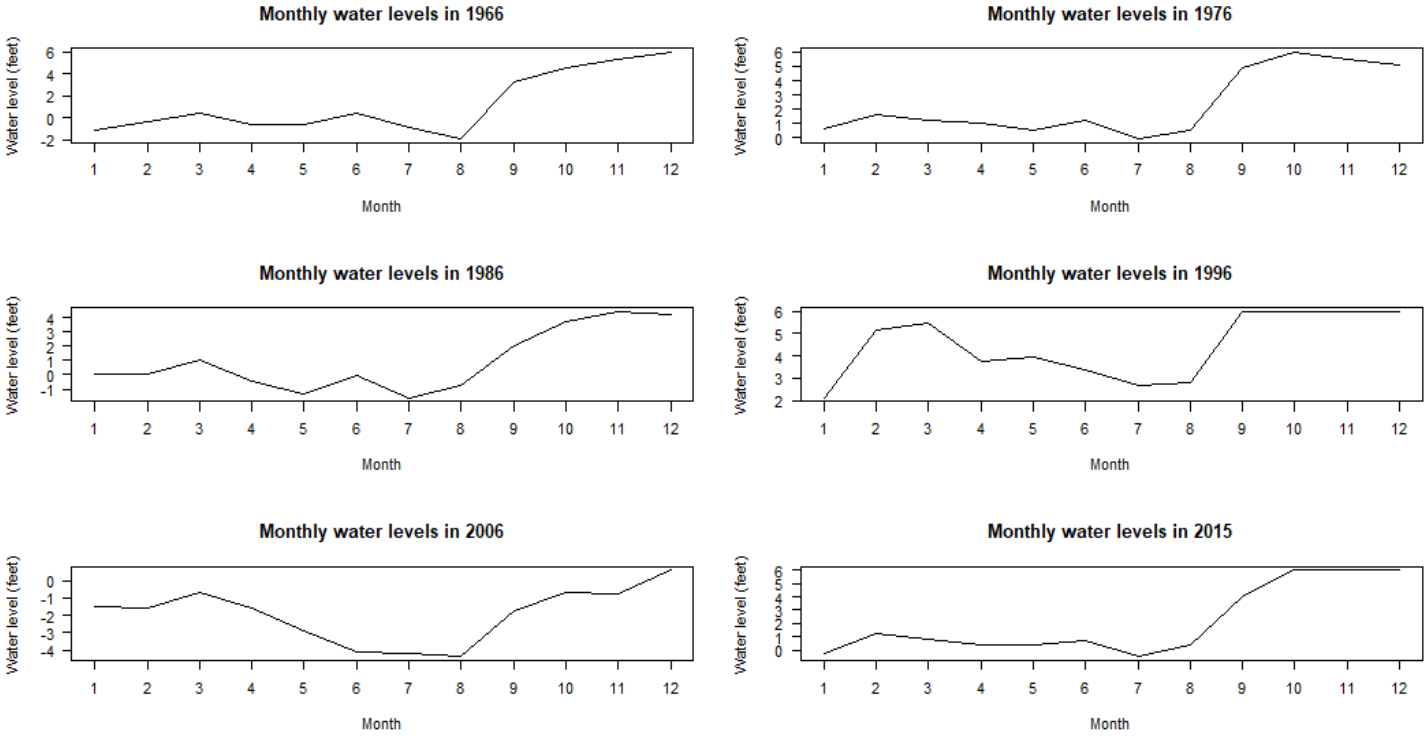
The data has a total period of 50 years (1966 - 2016), over which the trend and seasonality has remained fairly consistent apart from a large dip between 2004 – 2009 which suggests a freak event/anomaly given there is only one occurrence over 50 years. The dip in overall water level between 2004-2009 culminating in the lowest observed value of -6 feet could have occurred due to extreme drought/ adverse weather conditions. The data then appears to resume the general pattern post 2010.

The mean value of 2.52 is close to the median value of 2.64 indicating the data is rather symmetrically distributed. This is supported by the first quartile value (0.71) and third quartile value (4.45 being distributed almost evenly from the median.

The standard deviation of water level data is 2.48 which suggests the water level stays fairly consistent around the mean through the 50-year period.

Evidence of seasonal effects

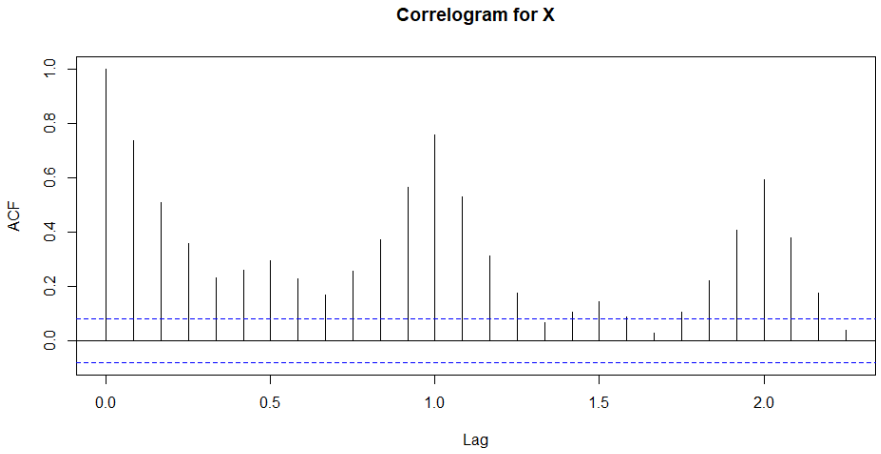
The plots below show a further analysis of the data focusing on the 12-month period every ten years starting from 1966 up until the final year of records.



The plots show that the water level is at its lowest in months 7 and 8 (July and August) which coincides with the summer period where less rainfall is to be expected.

It can also be observed how months 11 and 12 (November and December) have the highest water level which is to be expected given the greater level of rainfall in these months and the general weather conditions.

These yearly plots further indicate that there is a seasonal component in the data. The correlogram on the right side also supports this as there are high autocorrelations around lags 1 and 2.



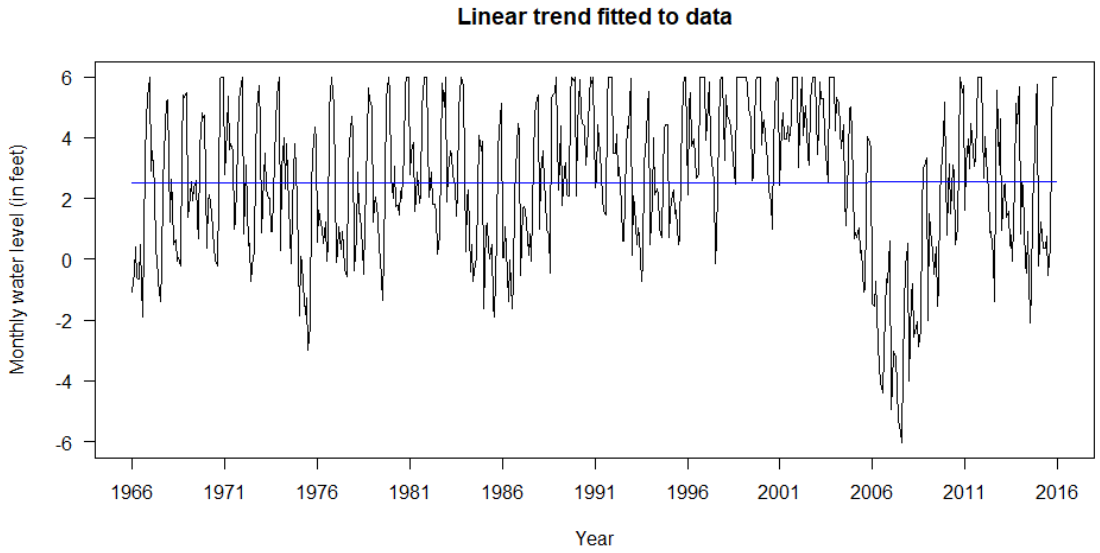
Removing trend and seasonal effects from the data

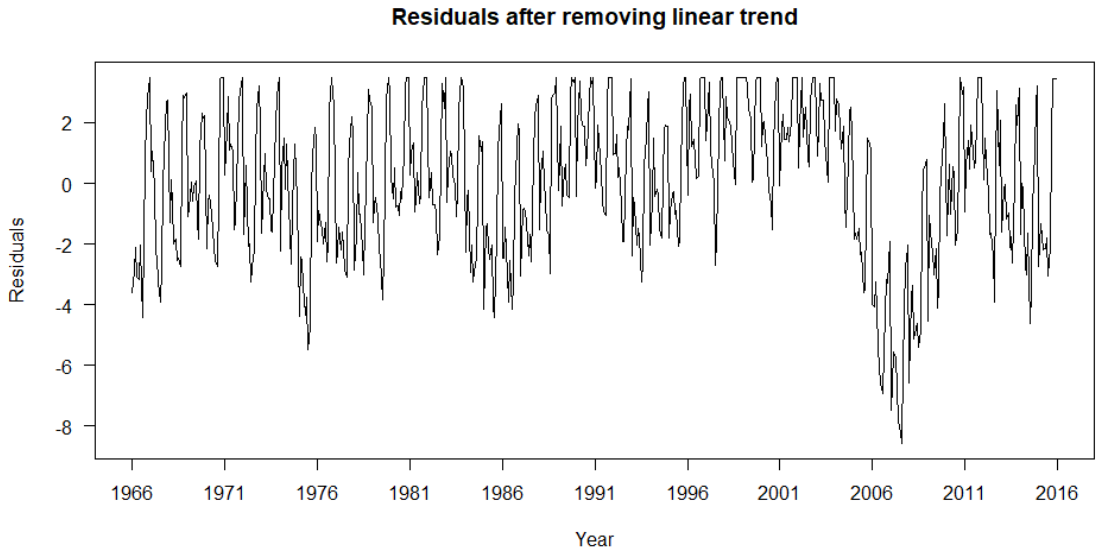
To fit a linear trend to the data, the following equation has to be solved: $\mu(t) = \alpha + \beta t$

The coefficients for the linear trend are estimated using linear regression which can be done easily in R. The following equation is obtained: $\hat{\mu}(t) = 2.5 + 0.0000665686t$

The $\hat{\beta}$ is quite small so the βt term can be disregarded giving the final equation for the trend: $\hat{\mu}(t) = 2.5$

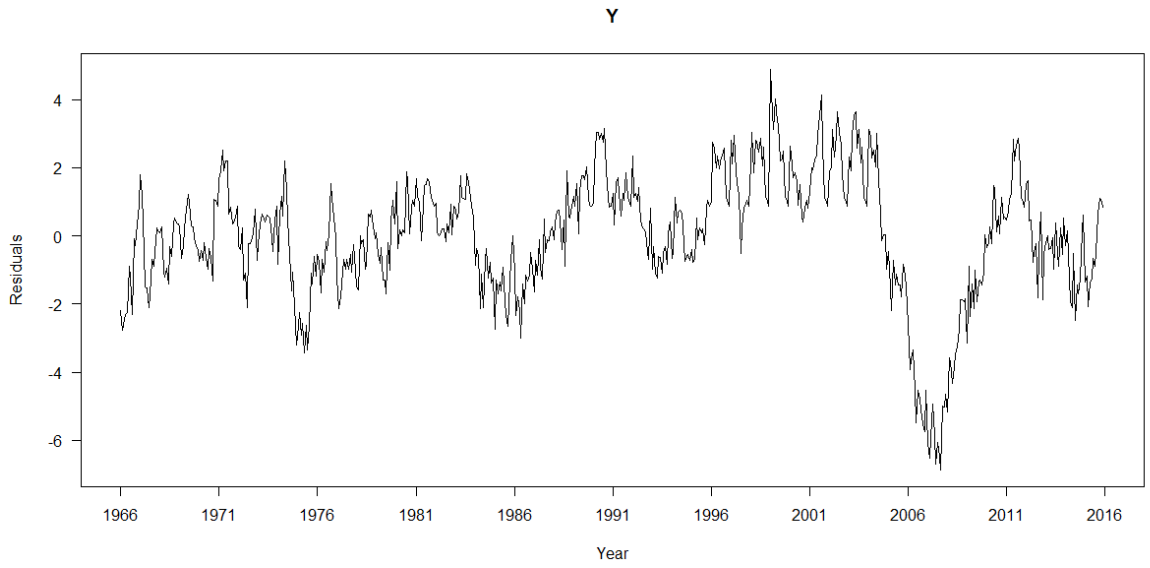
The fitted trend gives us a trend line that stays quite close to the mean value of 2.5. Plotting the residuals shows the trend has had little to no effect on the data and the variance has roughly stayed the same. This is to be expected as the trend is constant. The fitted trend and the residuals after removing the trend can be seen in the plots below.



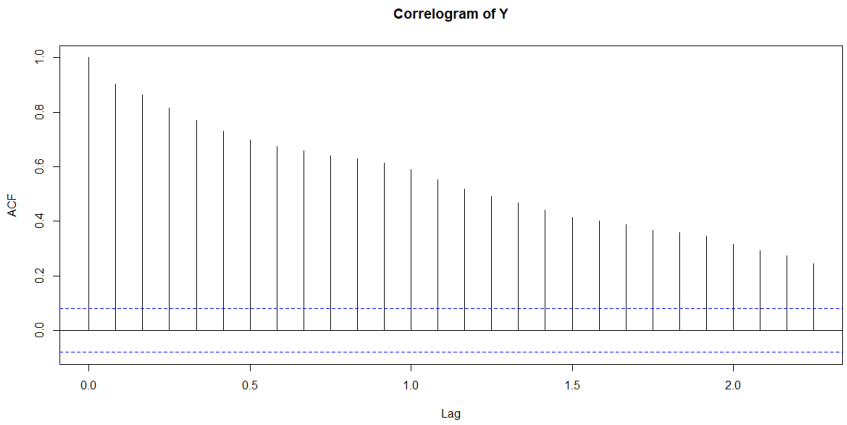


There are still seasonal effects apparent in the data, and removing them leaves us the below graph of the residuals Y. There is no clearly visible structure remaining. Using r to use linear regression to work out the coefficients gives the following equation for the seasonal effects:

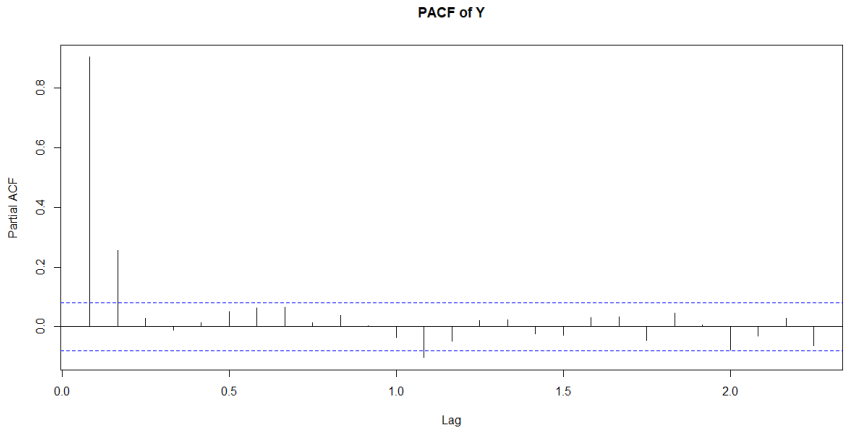
$$\hat{s}(t) = -1.40 \cdot Jan - 0.15 \cdot Feb + 0.35 \cdot Mar - 0.79 \cdot Apr - 0.9 \cdot May - 1.13 \cdot Jun - 2.15 \cdot Jul - 2.11 \cdot Aug + 0.88 \cdot Sep + 2.35 \cdot Oct + 2.44 \cdot Nov + 2.61 \cdot Dec$$



Inspecting the process Y



It can be seen from the correlogram for Y that the residuals are not a white noise process. Since the correlogram doesn't cut off quickly, an MA process would not be a suitable model. MA processes are stationary and have a mean value of 0. The process Y doesn't have a 0 mean value further indicating it's not a MA process.



Plotting the PACF of Y suggests that an AR process would be suitable to fit to Y. This is because the PACF cuts off approximately after lag 2 . If the process was a MA or an ARMA process, the PACF would show exponential decay or damped oscillations.

Further investigation is required to determine the order of the AR(p) model to fit to the data.

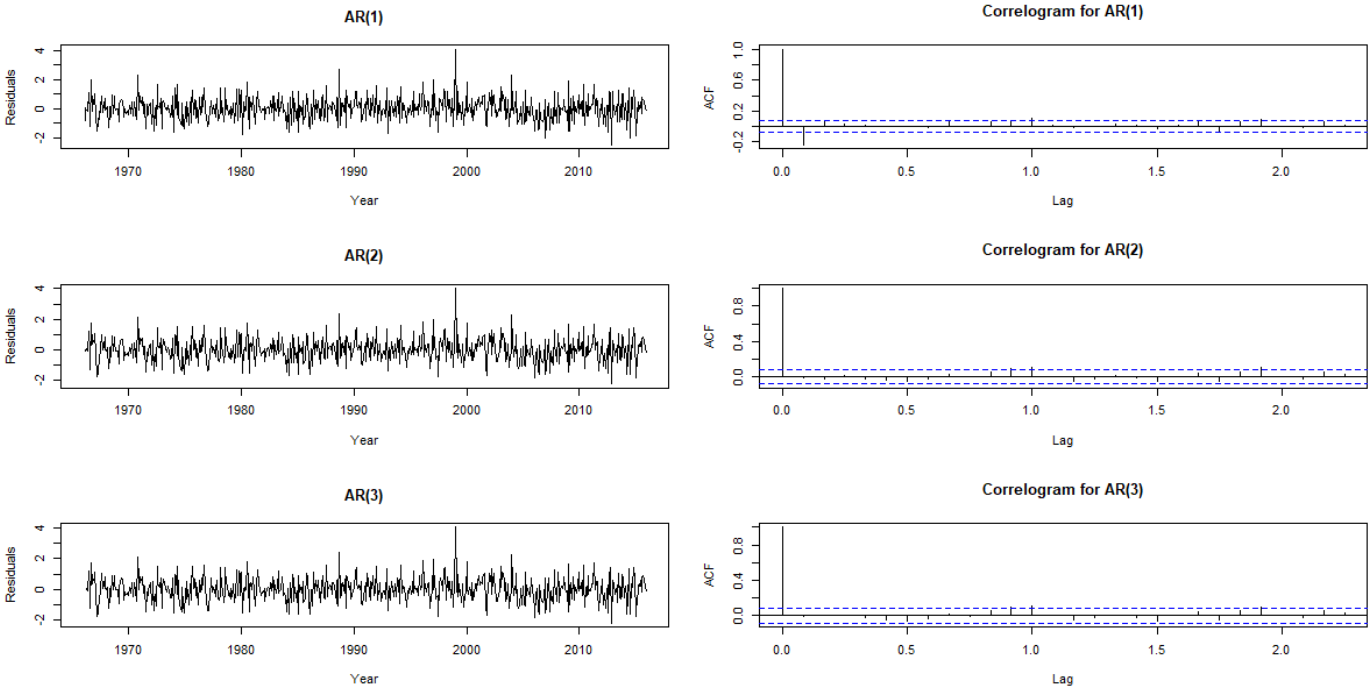
Fitting AR(p) models

To choose the correct AR(p) model to best fit the time series Y, the Yule-Walker equations are used for each of p=1,2,3 to fit an AR(p) model to the time series Y where p is the order of the model.

Using the R function 'AR', the following results in the table to the right are obtained for the estimated coefficients and variance for each order:

Model	Estimated Coefficients			Estimated Variance
	$\widehat{\alpha_1}$	$\widehat{\alpha_2}$	$\widehat{\alpha_3}$	$(\widehat{\sigma_\varepsilon^2})$
AR(1)	0.9025	-	-	0.6292
AR(2)	0.6708	0.2568	-	0.5887
AR(3)	0.6634	0.2374	0.0288	0.5892

Below are the plot of the residuals and corresponding correlograms for each model:



The AR(1) model has the largest variance and its correlogram suggests it's not a white noise process as the lag doesn't cut off at 0 given there is a sizeable lag value at 0.1. The correlograms for AR(2) and AR(1) are quite similar and both resemble white noise processes given that the lag cuts off at 0. The AR(2) model has a smaller variance than AR(1), also for AR(3) the α_3 coefficient is very small so the third term can be disregarded and therefore it would be preferable to go with the simpler model and choose AR(2). The residuals plot for AR(2) resembles white noise further indicating it is a good fit.

The model with the best fit is the AR(2) model and will be chosen as Z.

The complete model

The complete model X can be summarised as having three parts:

1. A linear trend: $\hat{\mu}(t) = 2.5$
2. Seasonal effects: $\hat{s}(t) = -1.40 \cdot Jan - 0.15 \cdot Feb + 0.35 \cdot Mar - 0.79 \cdot Apr - 0.9 \cdot May - 1.13 \cdot Jun - 2.15 \cdot Jul - 2.11 \cdot Aug + 0.88 \cdot Sep + 2.35 \cdot Oct + 2.44 \cdot Nov + 2.61 \cdot Dec$
3. Fluctuation Y where $Y_t = 0.67Y_{t-1} + 0.26Y_{t-2} + Z_t$ where Z_t is white noise.

The following formula shows the complete formula for X

$$X_t = 0.67Y_{t-1} + 0.26Y_{t-2} - 1.40 \cdot Jan - 0.15 \cdot Feb + 0.35 \cdot Mar - 0.79 \cdot Apr - 0.9 \cdot May - 1.13 \cdot \text{Jun} - 2.15 \cdot \text{Jul} - 2.11 \cdot \text{Aug} + 0.88 \cdot \text{Sep} + 2.35 \cdot \text{Oct} + 2.44 \cdot \text{Nov} + 2.61 \cdot \text{Dec} + \varepsilon_t$$

This shows that the data can be summarised as being a weighted average of its own past, plus a linear trend, weighted seasonal effects and some random noise.