Mathematics for Supply Chain

Msc Supply Chain & Purchasing Management (2022-2023)

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

1 Linear Regression: a case study

In this part, we start by studying the data related to matches that took place the previous week in order to deduce different information for the following week's matches.

1.1 A linear model to predict attendance

For this, we will consider the dataset named *attendance_ train.csv* which contains several information about matches that took place in different stadiums.

We will first study this data in order to build a model that will allow us to predict the number of spectators at a match.

Tickets and Promotions We also have the following information: the price for a place is 20 euros during the week and 25 euros during the week-end. The promotion means that all tickets are 10% cheaper

Cost Stadium We have the following information regarding the size of the stadium:

- when the stadium/hall has less than 1000 seats, the match has a cost of 50,000 euros.
- when the stadium/hall has between 1000 and 2000 seats, the match has a cost of 60,000 euros.
- when the stadium/hall has more than 2000 seats, the match has a cost of 70,000.

Other revenues (consumptions) We make the following assumptions about the consumptions (food and drink) during a match

- 40% of the seated people consume an average of 6 euros worth of drinks (2 drinks)
- 30% of the seated people consume an average of 8 euros worth of food (one sandwich)
- 80% of the people standing consume an average of 10 euros worth of drinks (3 drinks)
- 20% of those seated consume an average of 8 euros worth of food (one sandwich)

Cost employees Finally, the cost of an employee to serve the audience must be taken into account. The average cost of an employee is 500 euros per game. However, the current regulations require one employee for every 50 spectators.

Questions

1. Build your multiple linear model and estimate the parameters
2. Study the output of the model and tell which variables are significant or not, *i.e.* describe the statistical outputs of the model.
3. Given the current information, can we say that the organized meetings are profitable?
4. Evaluate the quality of the model, to do this, you have to compute the *Mean Squared Error* of the model, *i.e.*:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

   where y_i is the real attendance value and $\hat{y}_i$ is the predicted attendance value.
5. Looking at the data more closely, do you think it is possible to improve the prediction model? Suggest solutions.

1.1.1 Use our model for prediction

In the following we will try to predict the number of spectators for a set of games that will be held next week in different stadiums based on the the previous learned model.

We will use the same information as before to make this prediction. All variables are known in advance except the temperature which has been estimated using a model based on time series.

The information are available in the file *attendance_ test.csv*.

Questions

1. Use the previous learned model to predict the attendance for these future match.
2. Compute the Mean Square Error on these data. Can we say that the previous model can be used for prediction. Is it a good model ?

2 Towards Auto-Regressive Model

Here we propose to study a different approach for the prediction of the number of spectators based on auto-regressive models.

The idea behind this type of model is to try to predict the values of a time series of a single variable to be predicted according to its previous values.

In other words, this type of model (i) is used for data that depend on time only and (ii) tries to predict the values of a single variable, according to these past values. The model is the following, if consider Y the variable for which we aim to predict the values

$$Y_t = \alpha_1 Y_{t-1} + c + \varepsilon_t$$

where c is a constant, α_1 is a the parameter of the model and ε_t is a gaussian random variable with 0 mean and a given standard deviation σ which is the same for all t.

This model can be generalized as follows:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_3 Y_{t-3} + \ldots + \alpha_p Y_{t-p} + c + \varepsilon_t$$

It is then called *Autoregressive Process* of order p. And the parameters can be, for instance, estimated by a least squares method.

Questions

1. Is it possible to apply this strategy to the previous data. Explain why.

But before coming back to these models, called AR models, we will first review some standard smoothing methods for short-term prediction : the *moving average* and the *exponential smoothing*.

To study these two approaches, we will use the file entitled *Lyon.csv* which contains meteorological data on the last three years in Lyon.

On this file, we will keep only two information, the temperatures (min and max) and the dates.

2.1 Moving Average

The idea of the moving average is to use the previous values (a number that we will set ourselves) to calculate an average that will allow us to determine the value at time t. In other words, the value of Y_t at time t will be expressed as the average of the values between times $t-p$ and $t-1$:

$$Y_t = \frac{1}{p}\sum_{k=1}^{p} Y_{t-k}$$

This method is very simple to implement because it only involves calculating averages using a series of values. It can be interesting to implement when the data are simple by their structure, that they do not present any particularities. But we will come back to this point later.

It is also possible to make a weighted moving average, this time assigning different weights according to whether certain periods of the past are more interesting for the prediction at time t.

In this case, we will try to predict the value of Y_t as

$$Y_t = \sum_{k=1}^{p} w_k Y_{t-k}, \quad \text{where} \quad \sum_{k=1}^{p} w_k = 1.$$

Questions

1. Use the previous dataset, from 01/01/2020 to 31/12/2021 to predict the min and the max temperature at the following date: 01/01/2022.
   (a) using $p = 5$, *i.e.* the five previous values.
   (b) using $p = 10$,
   (c) using $p = 20$.
2. Represent the different predictions on a graph.
3. Comment the previous results (you can try higer values of p).
4. What kind of criterion can you define to measure the quality of the estimated values? Think about what we have seen in linear regression.
5. Compute this value for the different value of p.

We can also look at a particular type of smoothing, called exponential smoothing.

2.2 Exponential Smoothing

Exponential smoothing also consists in performing an average output but taking weights in a very precise way. The model looks as follows:

$$Y_{t+1} = \alpha Y_t + (1-\alpha)Y_{t-1} + (1-\alpha)^2 Y_{t-2} + \ldots + (1-\alpha)^p Y_{t-p} = \alpha Y_t + \sum_{k=1}^{p} (1-\alpha)^k Y_{t-k},$$

where α is a parameter between 0 and 1 which allows to control the weight given to the past values.

Again, this method depends on the history of the variable and the user can consider a more or less important history depending on the nature of the data and the amount of data available.

In practice, it is not necessarily useful to take a large history.

Questions

1. Describe the influence of the parameter α.
2. Use the previous dataset, from 01/01/2020 to 31/12/2021 to predict the min and the max temperature at the following date: 01/01/2022.
   (a) using $p = 5$, *i.e.* the five previous values.
   (b) using $p = 10$,
   (c) using $p = 20$.
3. Represent the different predictions on a graph.
4. Comment the previous results.

2.3 Other

The idea of this section is to get away from the values taken by our series and to focus on the different components of a time series.

In general, a time series Y_t can be written as the sum of three components

$$Y_t = T_t + S_t + \varepsilon_t,$$

where T_t represents the trend of our time series, S_t represents the seasonal component of our time series and ε_t represents a series without any particularity. This can be thought of as white noise, i.e., the errors made during data collection or the part that cannot be explained by the model.

2.3.1 Trend Study

Generalities The trend allows us to reflect the general behavior of our time series or series of values. The idea is, basically, to understand the standard behavior of the series without taking into account the punctual events that could modify its values.

Example: we can try to study the trend of the sales volume of a product without taking into account the possible discounts or offers that are made on the product.

More simply, we will try to see how our quantity evolves over time, using simple functions like a linear function

$$T_t = \beta_0 + \beta_1 t,$$

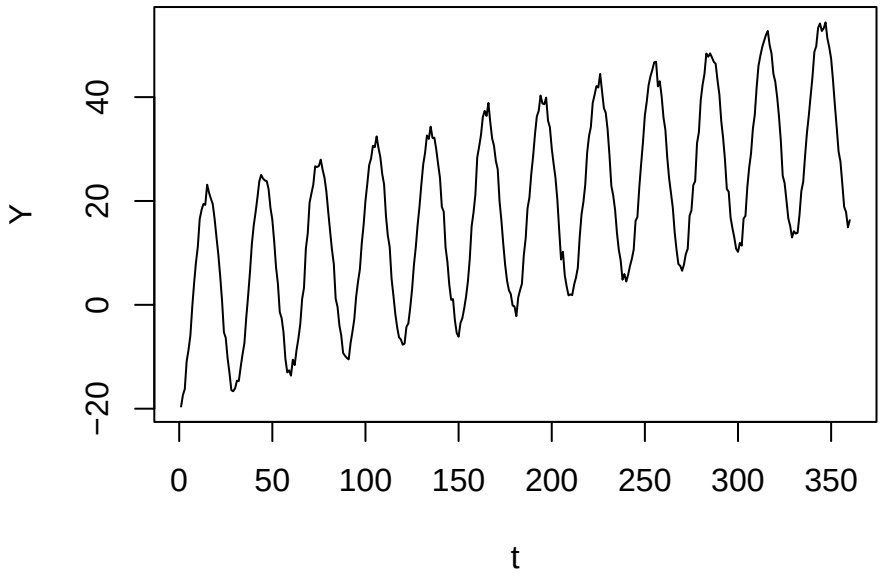
where β_0 and β_1 are two parameters. We also try to find if the trend is not linear, quadratic for instance:

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2,$$

where β_0, β_1 and β_2 are three parameters.

The parameters used to determine the trend line can be estimated using a least squares method, as can be done in the linear model. But let's now look at how we can identify the nature of the trend by considering two examples, a linear trend and a quadratic trend.

A linear trend Let us suppose that we have the series of values whose graphic representation is given below.



There is a linear upward trend for the series of values that is taken by the data. We also notice that the data period a certain cyclicity but that we will not try to study for the moment.

To determine whether our data show a linear trend, we can look at the following series Z_t

$$Z_t = Y_t - Y_{t-1},$$

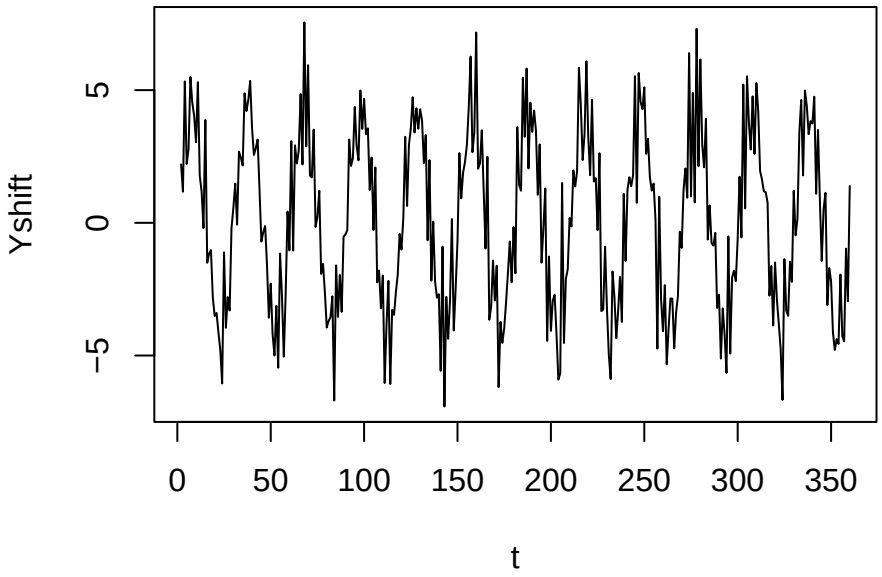
and check that this new series does not show any particular trend other than a possible seasonal behavior or noise. Indeed, for this new series, if we have

$$Y_t = \beta_0 + \beta_1 t + S_t$$

then

$$Z_t = Y_t - Y_{t-1} = (\beta_0 + \beta_1 t + S_t) - (\beta_0 + \beta_1 (t-1) + S_{t-1}) = \beta_1 + S_t - S_{t-1},$$

which has no particular behavior. On our previous example, it gives the following graph for Z_t:

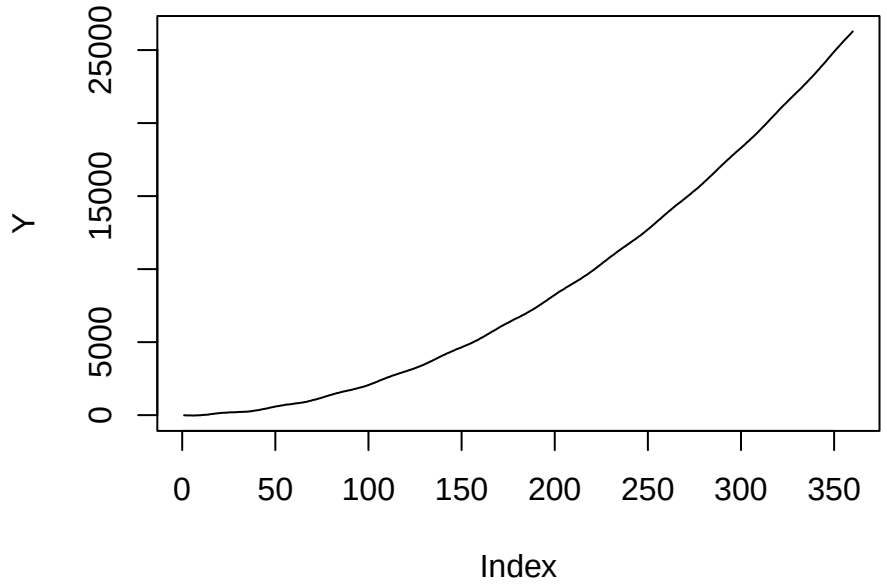


This new graph allows us to easily identify the seasonal behavior of data.

This step is important because it will allow us to build the most suitable model for prediction. Moreover, depending on the order of magnitude of our data, it will allow us to highlight possible seasonality.

But let us look at a quadratic trend first

A quadratic trend Let us suppose that we have the series of values whose graphic representation is given below.



Questions

1. Try to remove the quadratic tendance that is observed is this series based on what has been done with the linear trend.
2. Does the series have a seasonal component?

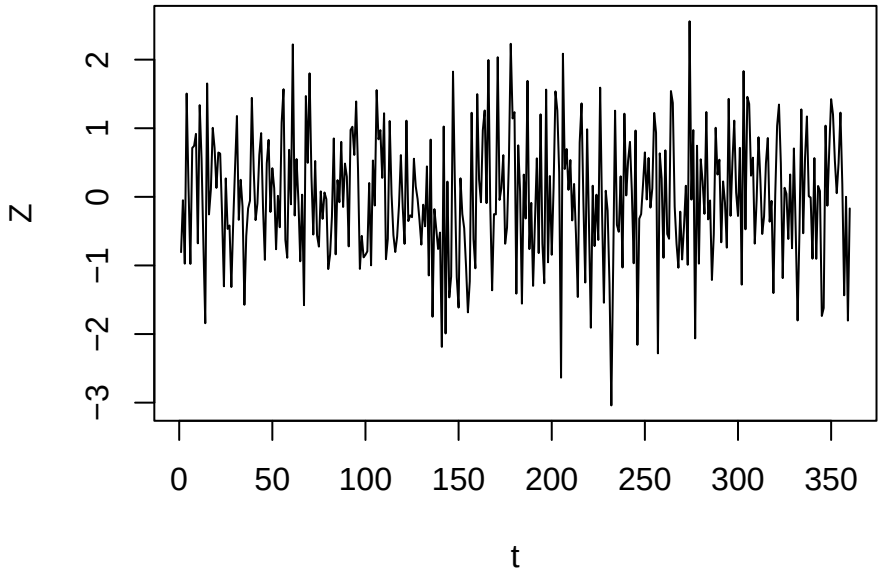
Seasonality Study Let us now start again with our series with a linear trend and look at seasonality to see how it can be taken into account in the model estimation.

What will interest us in this case is in particular the period of our seasonality (assuming there is one). If this is the case, we note T the period of our seasonality, and suppose that during the study, this periodicity is observed N times, i.e. the cycle occurs N. For example, a cycle occurs 12 times a year.

Then for all times between j between 1 and T, the impact of seasonality in the model will be revealed as the mean value. In other words, for all $j \in [\![1, T]\!]$:

$$S_j = \frac{1}{N} \sum_{k=1}^{N} S_{jk},$$

where S_{jk} denotes the values of the series S at each j component over all the N observed periods. If we go back to our previous example, with the linear trend and we apply the previous remark. It then results into the following graph:



And we do not observe anymore trend or seasonality in this dataset.

Questions

1. Apply the same process to the dataset used for the quadratic trend.
2. Use the learned model to mlake predictions for t from 361 to 400.

You are now ready to make predictions using these phenomena. If you want to go further in the analysis of that particular type of data, you can have look at the following models :

- AR
- ARMA
- ARIMA