TIME SERIES ANALYSIS

Victor Hugo F. Francheto
Institute of Mathematical and Computer Sciences
University of São Paulo
São Carlos, Brazil
victor.francheto@usp.br

February 25, 2025

Abstract: This material is a guide for students who wish to learn more about time series analysis. Our focus here is on presenting the fundamental concepts involved in time series.

Keywords: *write*

1 Introduction

2 Linear Regression

Before diving into time series analysis, it's important to revisit the concepts of linear regression. Linear regression provides a fundamental basis for understanding how variables relate to each other over time, serving as a stepping stone to more complex models.

Many time series techniques rely on estimating trends and relationships between variables—an idea already explored in linear regression. Moreover, time series models like ARIMA and dynamic regressions often incorporate linear components to capture patterns in the data.

Reviewing the fundamentals of linear regression helps reinforce key concepts such as minimizing the mean squared error, interpreting coefficients, and understanding residuals—all of which are crucial in time series analysis. With this foundation in place, transitioning to specific time series models becomes more natural and intuitive.

Regression analysis is a statistical tool that utilizes the relationship between two or more quantitative variables to make predictions. For example, if one knows the relationship between advertising expenditures and sales, one can predict sales using regression analysis once the level of advertising expenditures has been set.

2.1 Relation Between Variables

A functional relation between two variables is expressed by a mathematical formula. If X is the *independent* variable and Y is the *dependent* variable, a functional relation between X and Y is of the form:

$$Y = f(X).$$

Give a particular value of Xm the function f indicates the corresponding value of Y.

Example: Consider the relationship between dollar sales Y of a product sold at a fixed price and the number of units sold X. If the selling price is \$2 per unit, then the relationship is given by:

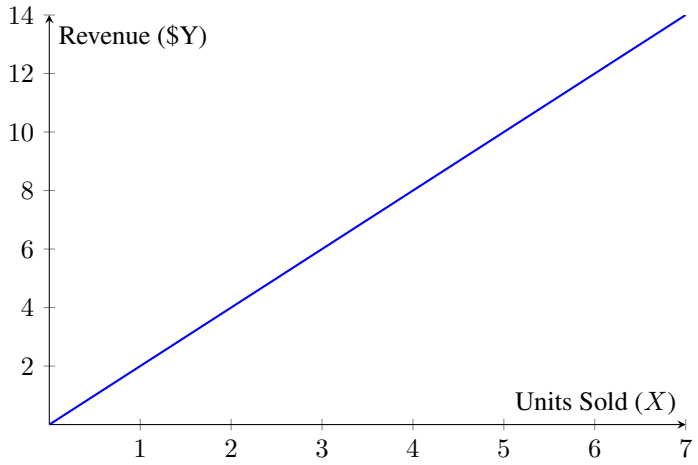
$$Y = 2X.$$

This means that for every unit sold, the total revenue increases by \$2. The table below shows the number of units sold (X) and the corresponding revenue (\$Y):

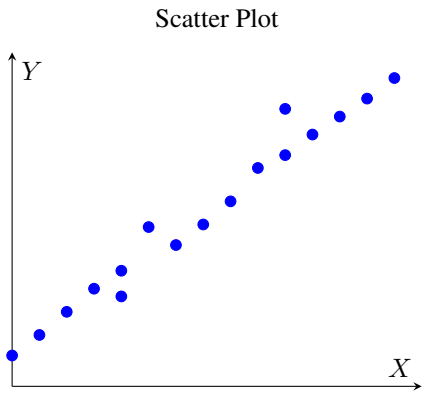
Units Sold (X)	Revenue (\$Y)
0	0
1	2
2	4
3	6
4	8
5	10
6	12

Table 1: Revenue (\$Y) as a function of units sold (X) for a price of \$2 per unit.

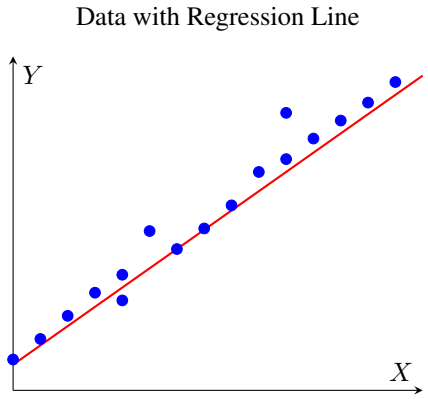
The graph below represents the function $Y = 2X$, showing how revenue (\$Y) increases linearly with the number of units sold (X).



A statistical relation, unlike a functional relation, is not perfect. In general, observations in a statistical relation do not fall directly on the curve of the relationship.



(a) Only data points.



(b) Data points with regression line.

2.2 Regression Models and Their Uses

A regression model is a formal approach to representing the fundamental relationship between two statistical quantities.

1. A tendency of the dependent variable Y to vary with the independent variable or variables in a systematic fashion;

2. A scattering of observations around the curve of the statistical relationship.

These two characteristics are embodied in a regression model by postulating that:

1. In the population of observations associated with the sampled process, there is a probability distribution of Y for each level of X.
2. The means of these probability distributions vary systematically with X.

The expressions *independent variable* or **predictor variable** for X, and *response variable* for Y in a regression model are merely conventional labels. There is no implication that Y causally depends on X in a given case. No matter how strong the statistical relationship, a regression model does not necessarily imply a cause-and-effect pattern.

In some applications, an independent variable may actually be causally dependent on the response variable. For example, when estimating temperature (the response) from the height of mercury (the independent variable) in a thermometer.

2.3 Regression Models with Multiple Independent Variables

Regression models may contain more than one independent variable. For example:

1. **Predicting Used Car Prices:** A dealership wants to predict the prices of used cars to adjust its sales strategy. To achieve this, it considers factors such as mileage, model year, engine power, and the number of previous owners as independent variables. Each of these variables helps understand how different characteristics influence the perceived value of vehicles;
2. **Predicting Urban Real Estate Prices:** A real estate company aims to estimate the prices of residential properties in a specific city. The independent variables considered include total built area, location (neighborhood), proximity to schools or green areas, property age, and the presence of a garage. These variables provide insights into which factors have the greatest impact on property valuation;
3. **Predicting Stock Prices in the Financial Market:** A team of financial analysts wants to forecast stock price fluctuations for technology sector companies. They consider independent variables such as quarterly profit, company debt, annual growth rate, brand reputation, and the overall economic outlook. These variables help understand the factors that may influence investor behavior and stock prices.

References

[1] Samuraí Brito, Askery Canabarro, Rafael Chaves, and Daniel Cavalcanti. Statistical properties of the quantum internet. *Physical Review Letters*, 124(21), May 2020.

[2] Minsuk Kim and Filippo Radicchi. Shortest-path percolation on random networks. *Physical Review Letters*, 133(4), July 2024.

APPENDIX

Master's Research Project: Quantum Networks

Next, we will briefly introduce the network I plan to study in my master's project. Understanding its properties requires knowledge of the concepts discussed earlier, especially communities and network mathematics. The dynamics of the network to be studied are described below.

A network model is defined by a set of N nodes connected by vertices according to a given probabilistic rule. The central goal of network science is to understand the asymptotic properties of networks as the number of nodes increases. A particularly relevant example is provided by random networks, defined by a model where each pair of nodes is connected with probability p. The hallmark of random networks is that for a sufficiently large number of nodes N, the probability of finding a node with k connections $\mathcal{P}(k)$, known as the degree distribution, can be approximated by a Poisson distribution

$$\mathcal{P}(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}$$

where $\langle k \rangle = p(N-1)$ is the average connectivity of the network. Despite its simplicity, the random network model exhibits rich statistical phenomena. For instance, it undergoes a phase transition: there exists a critical probability p_c such that if $p < p_c$, the network consists of small disconnected clusters, and if $p > p_c$, a giant cluster of size comparable to the entire network emerges.

Another notable feature is the phenomenon known as small-world. This refers to the fact that the average shortest path length (i.e., the shortest path between two nodes) scales logarithmically with N, meaning that typical distances between pairs of nodes are very short compared to the network size. Another important network property is the average clustering coefficient. It measures how interconnected the neighbors of each node are, on average. Specifically, we define the local clustering coefficient of node i as

$$\mathcal{C}_i = \frac{2n_i}{k_i(k_i-1)}$$

where n_i is the number of edges among the k_i neighbors of node i, and $k_i(k_i-1)/2$ is the total possible number of edges among them. If $\mathcal{C}_i = 0$, there are no links among the neighbors of i, while $\mathcal{C}_i = 1$ indicates that these neighbors form a fully connected graph. The average clustering coefficient is defined as

$$\langle \mathcal{C} \rangle = \frac{1}{N} \sum_i \mathcal{C}_i.$$

For random networks, $\langle \mathcal{C} \rangle = \frac{\langle k \rangle}{N}$, demonstrating a decrease with network size.

Next, we propose a model to simulate the quantum internet and use it to predict these properties for photonic networks. As we will see, these networks exhibit similarities and differences compared to random networks. Our model considers a network built from optical fibers, the primary medium for carrying quantum information encoded in photons. Other technologies, such as quantum satellites, are also under consideration and are likely to be integrated with the fiber-optic infrastructure. Therefore, the results presented here can be viewed as a benchmark to be refined by additional technologies. Our model is defined by the following steps:

Step-1 - Fiber-optic network simulation: Initially, we distribute N nodes uniformly within a disk of radius R. We simulate the distribution of optical fibers among these nodes using the Waxman model, which assumes that each pair of nodes i and j is connected by a fiber with a probability given by $\Pi_{ij} = \beta e^{-\frac{d_{ij}}{\alpha L}}$, where d_{ij} is the Euclidean distance between i and j, L is the maximum distance between any two nodes, $\alpha > 0$ controls the typical edge length of the network (maximum distance between directly connected nodes), and $0 < \beta \leq 1$ controls the average degree of the network. The constants α and β have been estimated for specific optical fiber networks, such as the US fiber-optic network where $\alpha L = 226$ km and $\beta = 1$. We will use these values in the numerical simulations presented here.

Step-2 - Photonic network simulation: Once we have generated the fiber-optic network, we simulate the transmission of photons through it. Photonic losses increase exponentially with fiber length. Specifically, the transmissivity, determining the fraction of energy received at the output of a fiber link connecting nodes i and j, is given by

$p_{ij} = 10^{-\frac{\gamma d_{ij}}{10}}$, where d_{ij} (km) is the Euclidean distance between i and j, and the fiber loss parameter γ depends on the photon wavelength. For example, for silica fiber, losses are minimized at a wavelength of 1550 nm, achieving $\gamma \approx 0.2$ dB/km, the value used in our simulations. Even with future advancements, the intrinsic physical loss limit of silica optical fibers is estimated to be between 0.095 to 0.13 dB/km.

Finally, we define the probability P_{ij} that two nodes are connected as

$$P_{ij} = 1 - (1 - p_{ij})^{n_p}.$$

The parameter n_p controls how many photons are sent between each node in attempts to establish a photonic link; specifically, two nodes are connected if at least one of n_p photons is transmitted between them. For illustrative purposes, we have chosen $n_p = 1000$ in the figures presented here, as this value ensures that connections over 100 km, typical of state-of-the-art quantum communication experiments, are established. It should be noted that extensive simulations have also been conducted with different values, demonstrating that the qualitative features described below for photonic networks are universal and independent of the value of n_p.

imgs/quantum.jpg

We use this model to predict global properties of typical photonic networks, including:

- Network connectivity;
- Robustness to failures;
- Presence or lack of community structures;
- Synchronization of information in the network;

- Shortest Path Percolation, i.e., how vertex disconnection affects network synchronization.

More details about the considered quantum network and the percolation process we are working on can be found in references [1] and [2].

It is important to note that the work is still in its early stages, where we are validating hypotheses about the type of network structure we aim to explore.