STAT 153 & 248 - Time Series
Lecture Two

Fall 2025, UC Berkeley

Aditya Guntuboyina

September 2, 2025

We shall start the first topic of the course: Linear Regression. We start our discussion with simple linear regression (where there is a single covariate), and then extend to multiple linear regression (where there are multiple covariates).

1 Simple Linear Regression

We want to learn the relationship between two variables y and x, with the aim of predicting y given the value of x. y is called the response variable, and x is called the covariate. For example (this was one of the original applications of regression), y denotes the height of an adult man, and x denotes the height of their father. The linear regression model assumes that y is related to x via the equation:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

where β_0 and β_1 are parameters, and ϵ denotes an error term which captures deviations of y from the assumed equation $\beta_0 + \beta_1 x$. The parameters β_0 and β_1 can be interpreted as follows: β_0 denotes the value of y when $x = 0$ and β_1 represents the change in y when x changes by one unit.

We observe data $(x_1, y_1), \dots, (x_n, y_n)$ on the covariate and response variables corresponding to n instances (in the height example, we have data on heights for n father-son pairs). Writing the equation (1) for each individual pair (x_i, y_i) we get

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{2}$$

The observed data $(x_1, y_1), \dots, (x_n, y_n)$ will be used to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ (as well as uncertaintly quantification) for the parameters β_0 and β_1. After obtaining these estimates, one can predict the value of the response variable for a possibly new covariate value x_{new} by $\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$.

We will implement the equations for obtaining $\hat{\beta}_0$ and $\hat{\beta}_1$ from the observed data $(x_1, y_1), \dots, (x_n, y_n)$ using the Python library `statsmodels`. We will also study the math behind this process.

To apply linear regression, we need data on both y and x. In the time series context, the observed data is $y_1, \dots, y_n$ which represent observations for a single variable y. There is no additional data on another variable x. In order to apply regression methods to time series, we need to create a covariate variable x. There are two main ways of doing it:

1. **Time as covariate**: Here we take the time index as the covariate x. For example, in the time series dataset on the population of the United States for each month from January 1959 to December 2024: n denotes the total number of data points, $x_i = i$ and y_i denotes the observed population data for the i^{th} month (first month is January 1959, second month is February 1959 and so on).

2. **Lagged y as covariate**: Here we take $x_i = y_{i-1}$. In other words, the covariate equals the response at the previous time point. This kind of regression is called Lagged Regression or, more commonly, AutoRegression.

For now, we shall focus on the first kind of regression (time as covariate). We shall study AutoRegression in more detail later.

2 Estimation of β_0 and β_1

2.1 Least Squares Estimates

The estimates of β_0 and β_1 reported by standard libraries (such as `statsmodels`) are obtained using the method of least squares. This involves minimizing the sum of squares criterion:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{3}$$

over all values of β_0 and β_1. It is left as an exercise to verify that:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where

$$\bar{y} = \frac{y_1 + \cdots + y_n}{n} \quad \text{and} \quad \bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

2.2 MLE under Normality of Errors

Suppose we assume that the error terms $\epsilon_1, \dots, \epsilon_n$ in (2) are i.i.d normal with mean zero and some variance σ^2:

$$\epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2). \tag{4}$$

Then the least squares estimates of β_0 and β_1 coincide with the Maximum Likelihood Estimates (MLEs).

Another way of writing the model (2) and (4) is:

$$y_i \overset{\text{independent}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).$$

To obtain the MLEs of the parameters (β_0, β_1 as well as σ), we need to write the likelihood function and then maximize it. The likelihood function is the joint density of the data for fixed values of the parameters β_0, β_1, σ:

$$
\begin{aligned}
f_{y_1,\ldots,y_n|\beta_0,\beta_1,\sigma}(y_1,\ldots,y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right) \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right) \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left(-\frac{S(\beta_0,\beta_1)}{2\sigma^2}\right) \qquad (5)
\end{aligned}
$$

where $S(\beta_0, \beta_1)$ is the sum of squares (3).

$$S(\beta_0,\beta_1) := \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2.$$

To write this likelihood, we are assuming that $x_1, \ldots, x_n$ are fixed. This assumption is fine if $x_i = i$ (regression with time as covariate) but not strictly true when $x_i = y_{i-1}$ (auto-regression). We shall see how it is still approximately true in the case of AutoRegression later.

Maximization of the likelihood is a three variable optimization problem (the variables being β_0, β_1, σ). The optimal values of β_0 and β_1 in this problem coincide with the least squares estimate. We shall see why in the next lecture.