

Lecture One

STAT 153/STAT 248, Fall 2025

Aditya Guntuboyina (28 August 2025)

What is Time Series?

Time series is a set of observations each one being recorded at a specific time

The websites <https://fred.stlouisfed.org/> and <https://trends.google.com> have many examples of time series that we will use in this class

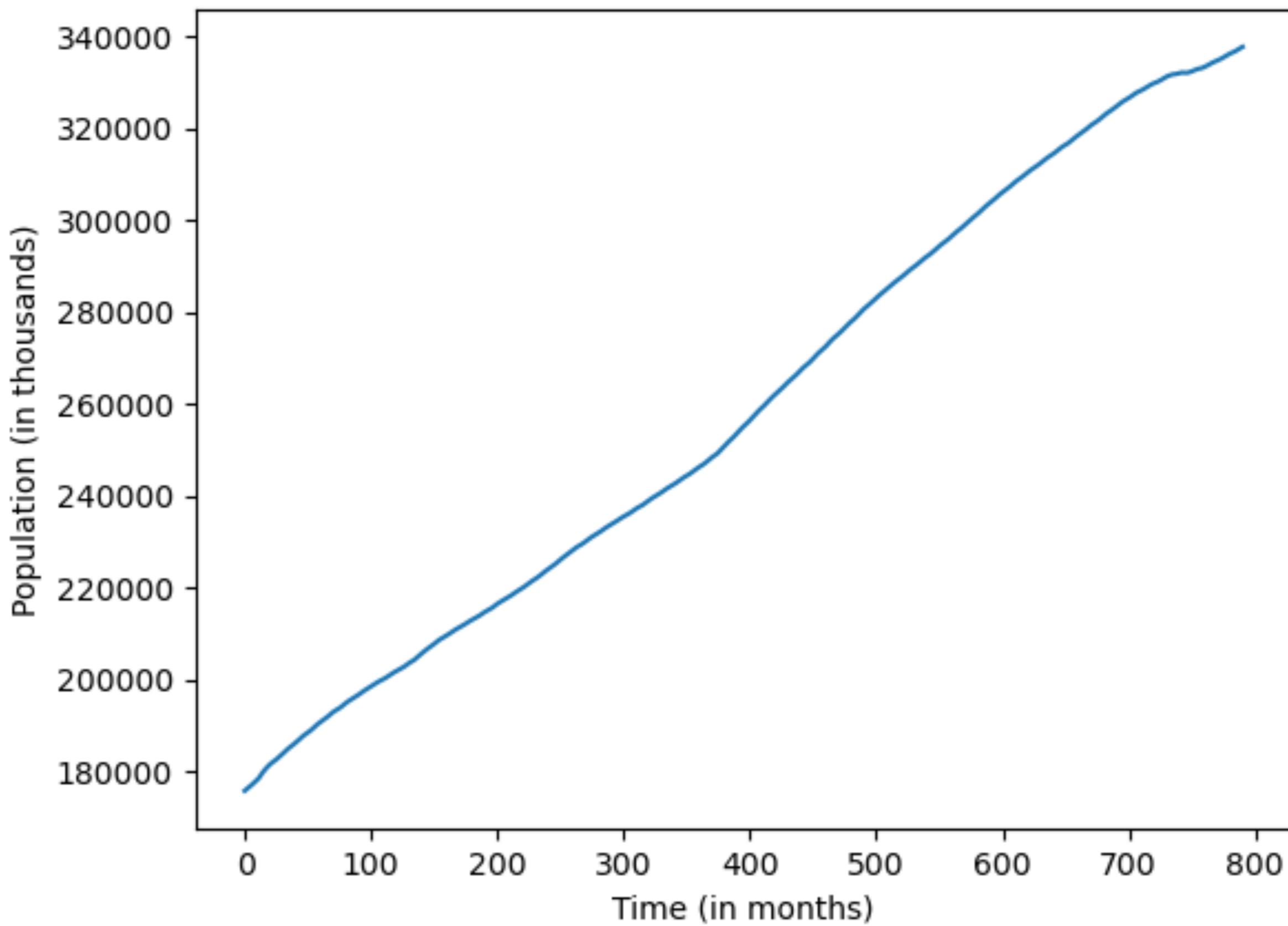
US Population

Population of the United States

observation_date	POPTHM
1959-01-01	175818
1959-02-01	176044
1959-03-01	176274
1959-04-01	176503
1959-05-01	176723
1959-06-01	176954
1959-07-01	177208
1959-08-01	177479
1959-09-01	177755
1959-10-01	178026
1959-11-01	178273
1959-12-01	178504
1960-01-01	178925
1960-02-01	179326
1960-03-01	179707
1960-04-01	180067
1960-05-01	180408
1960-06-01	180728

- Units are thousands so that 300,000 actually refers to 300 million
- This dataset is downloaded from FRED

US Population Data



Questions

- Prediction: what is an estimate of the population at a future point (e.g., Jan 2040)?
- What is the rate of growth of the American population?
- Has the growth rate been roughly constant over time?
- What is the period where the population grew the fastest? Slowest?

Time Series Analysis answers such questions by fitting statistical **models** to observed time series data

Time Series Prediction/Forecasting

One of the most important questions in time series analysis is that of prediction (estimating future values based on the given data)

What's the next number?

- 1.) 5, 8, 11, 14, 17, __
- 2.) 42, 32, 23, 15, 8, __
- 3.) 3, 6, 12, 24, 48, __
- 4.) -3, 7, -6, 11, -9, 15, __, __
- 5.) 4, 16, 36, 64, 100, __

Basic Time Series Prediction Problem

Find the next number: 1, 4, 9, 16, 25, #

The answer is 36 but how did we arrive at it?

We noticed that y_t is a function of the time t

In other words, we performed a (quadratic) regression of y_t over time t

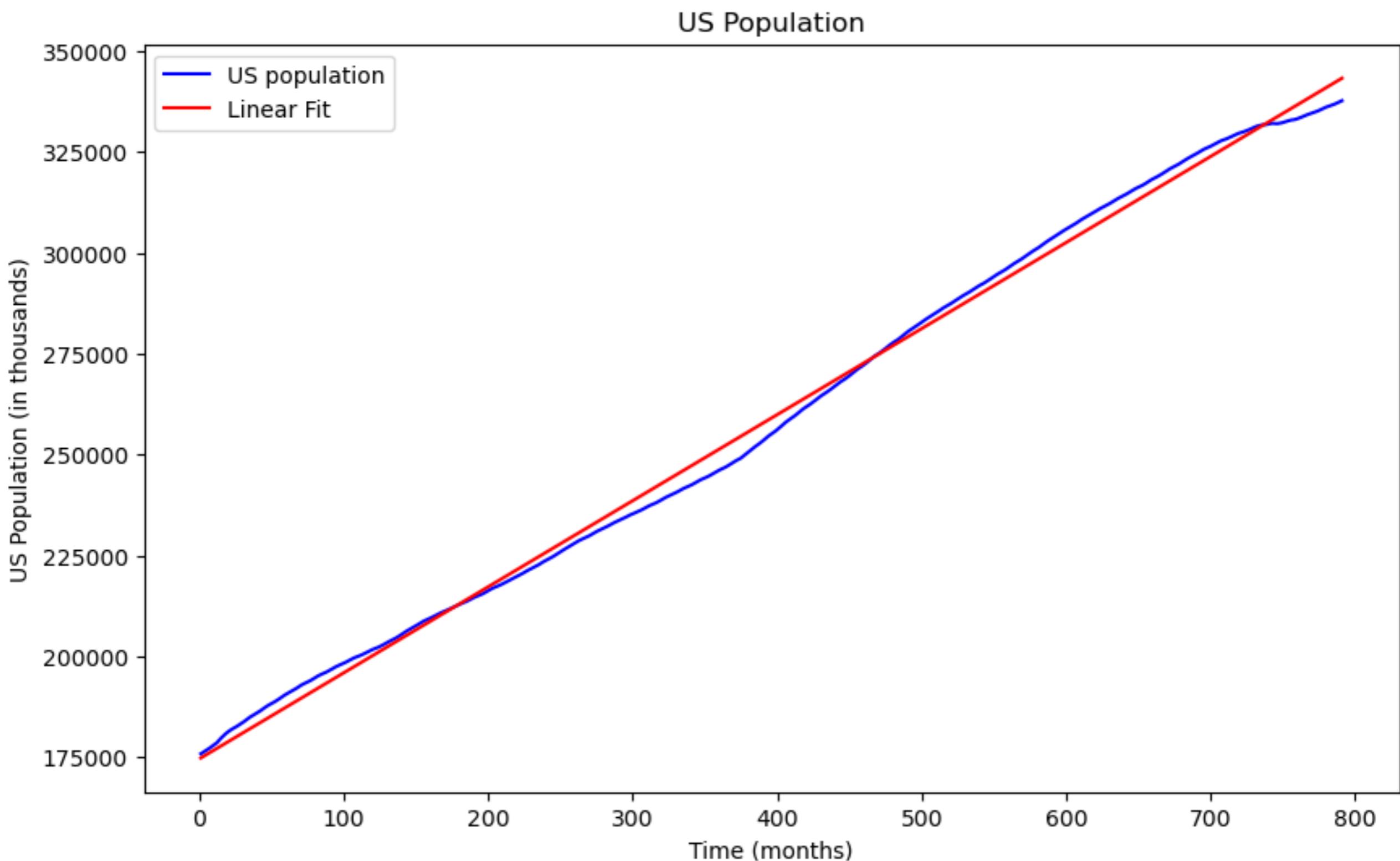
Regression over time

The first time series technique we will study is regression over the time variable t

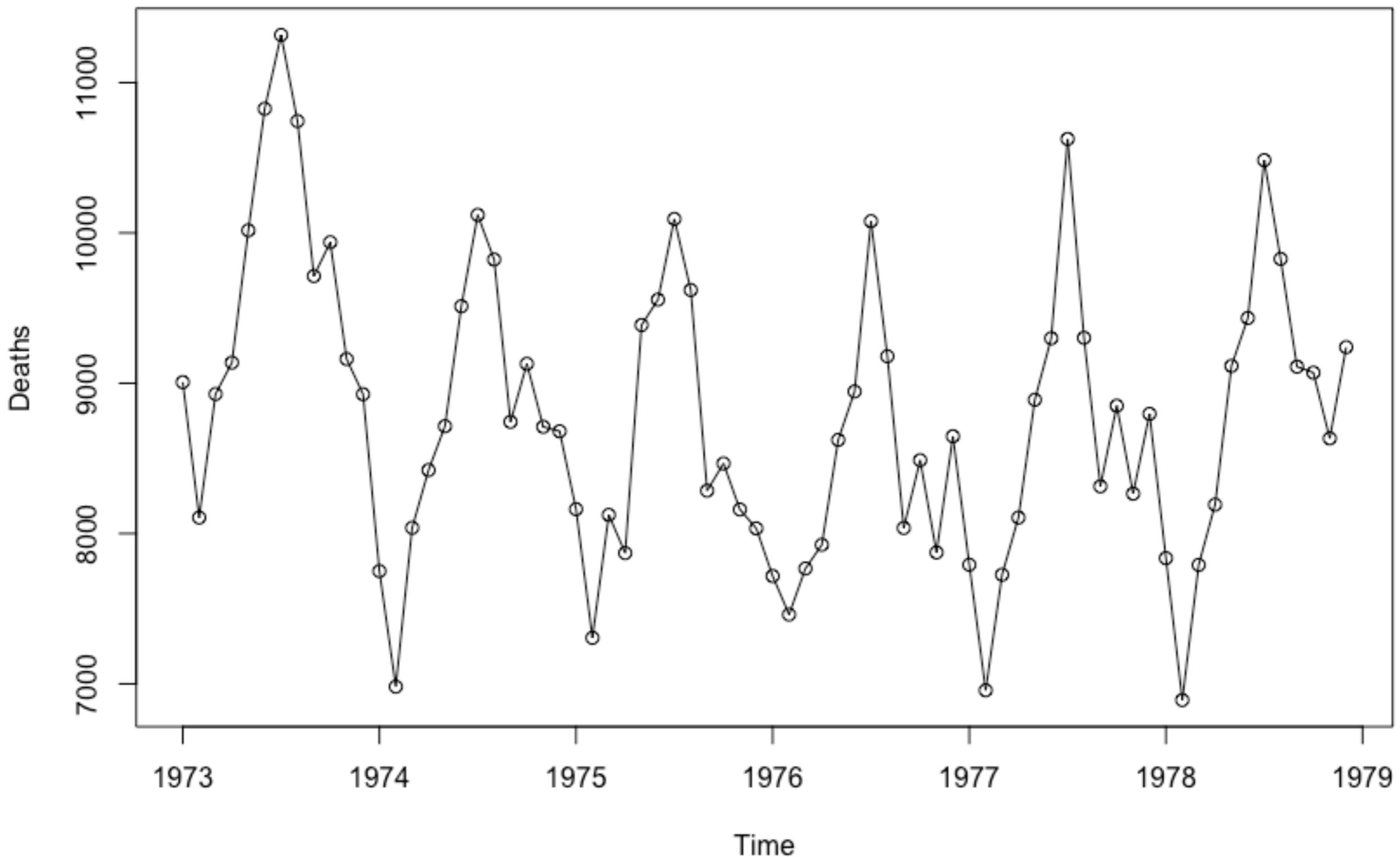
Simple linear regression of the time series y_t on the time variable t will fit a line to the observed time series data

Multiple linear regression of y_t on t and other functions of t (such as powers, sinusoids etc) will fit more general functions to the data

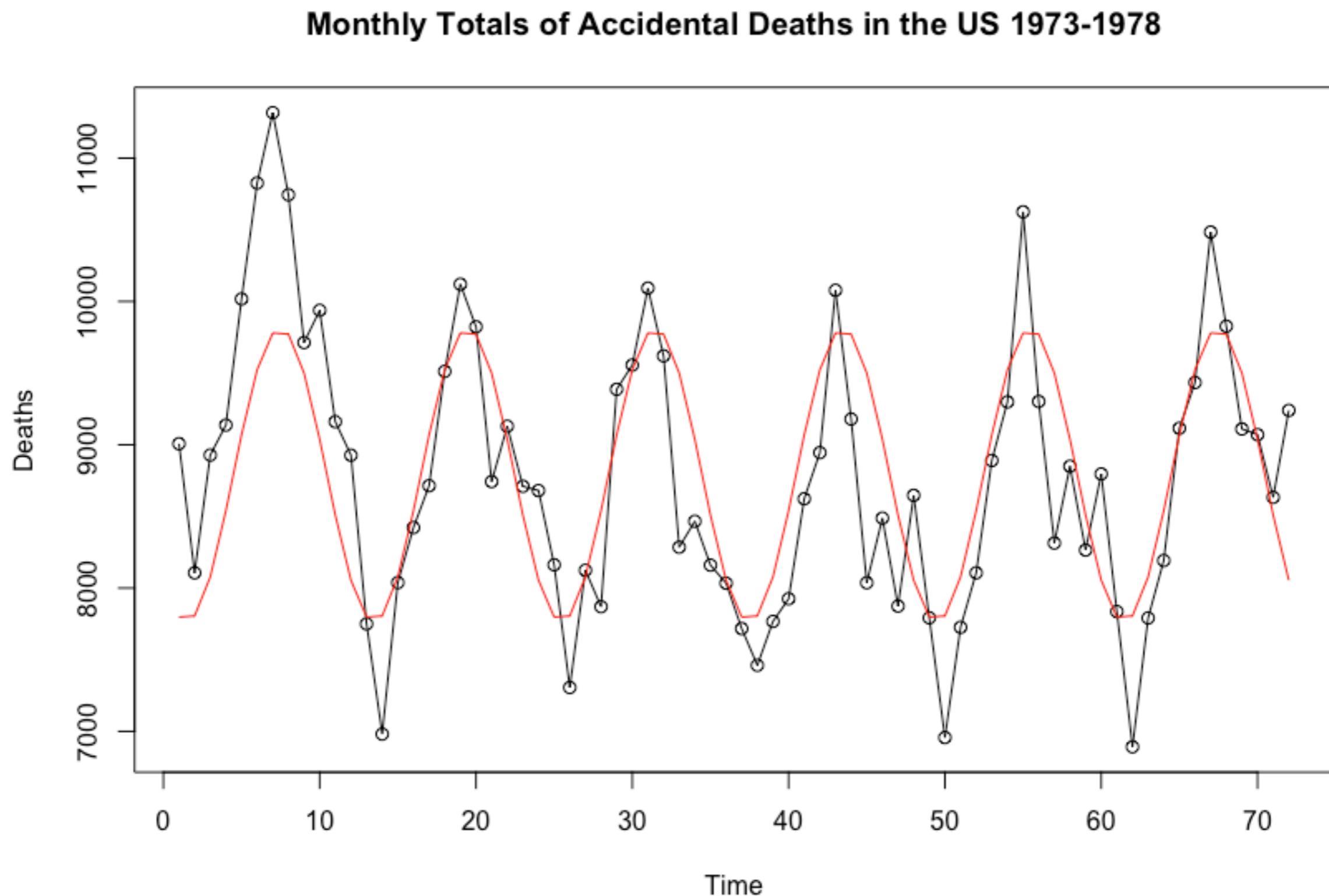
Linear regression of y_t on $1, t$



Monthly Totals of Accidental Deaths in the US 1973-1978

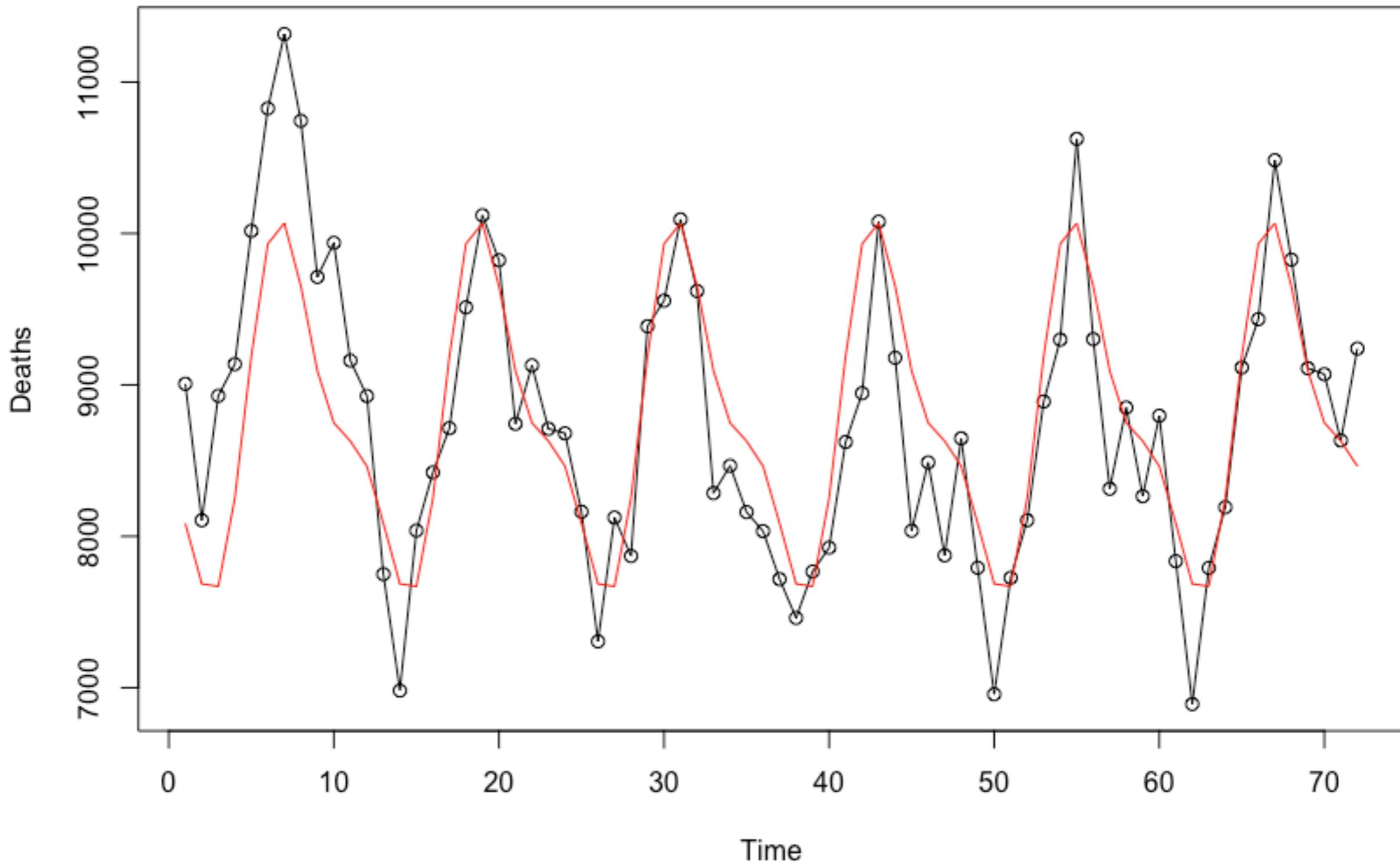


Linear regression of y_t on 1, $\cos(\pi t/6)$, $\sin(\pi t/6)$:



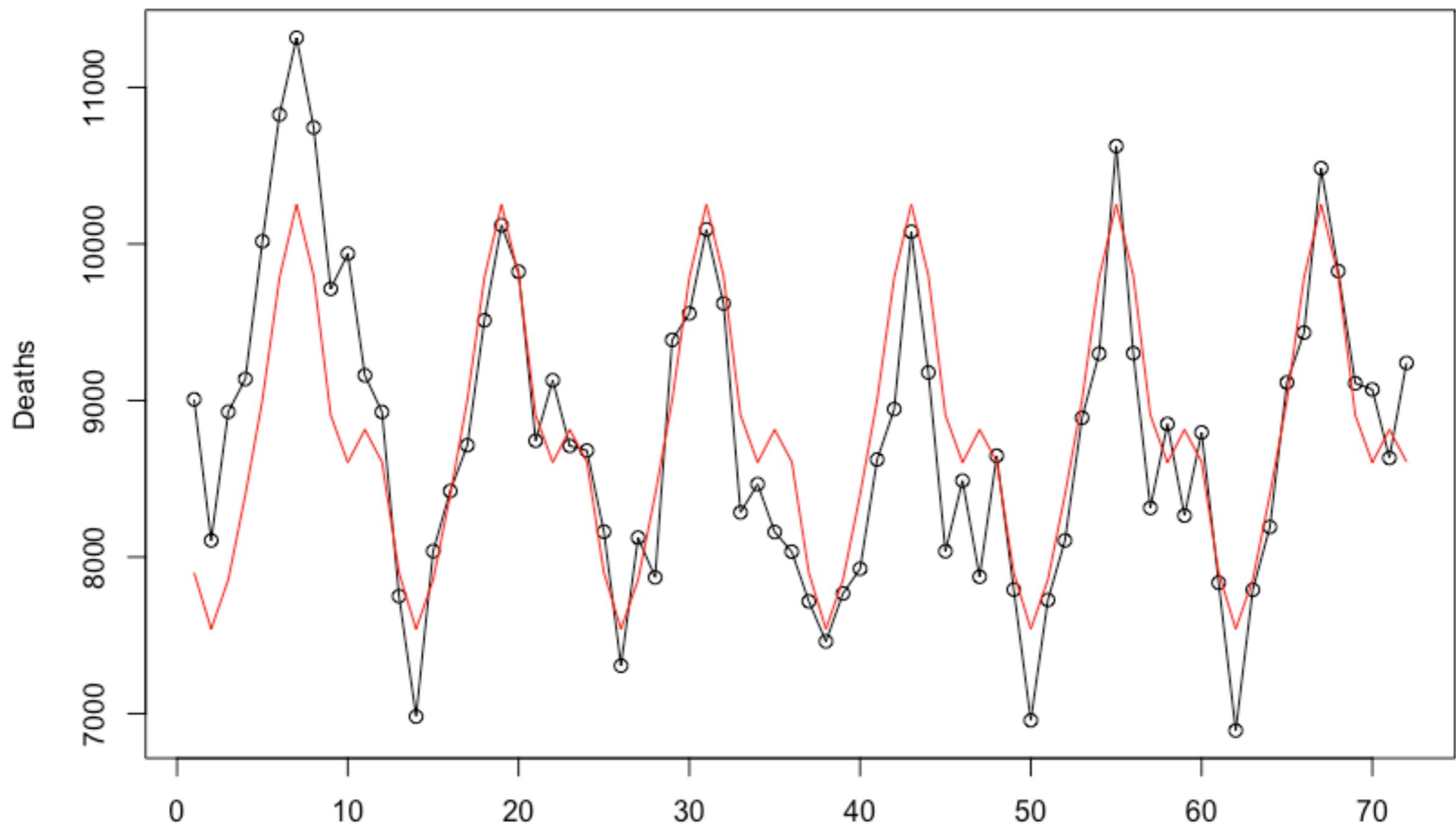
$$y_t \sim 1, \cos(\pi t/6), \sin(\pi t/6), \cos(\pi t/3), \sin(\pi t/3)$$

Monthly Totals of Accidental Deaths in the US 1973-1978



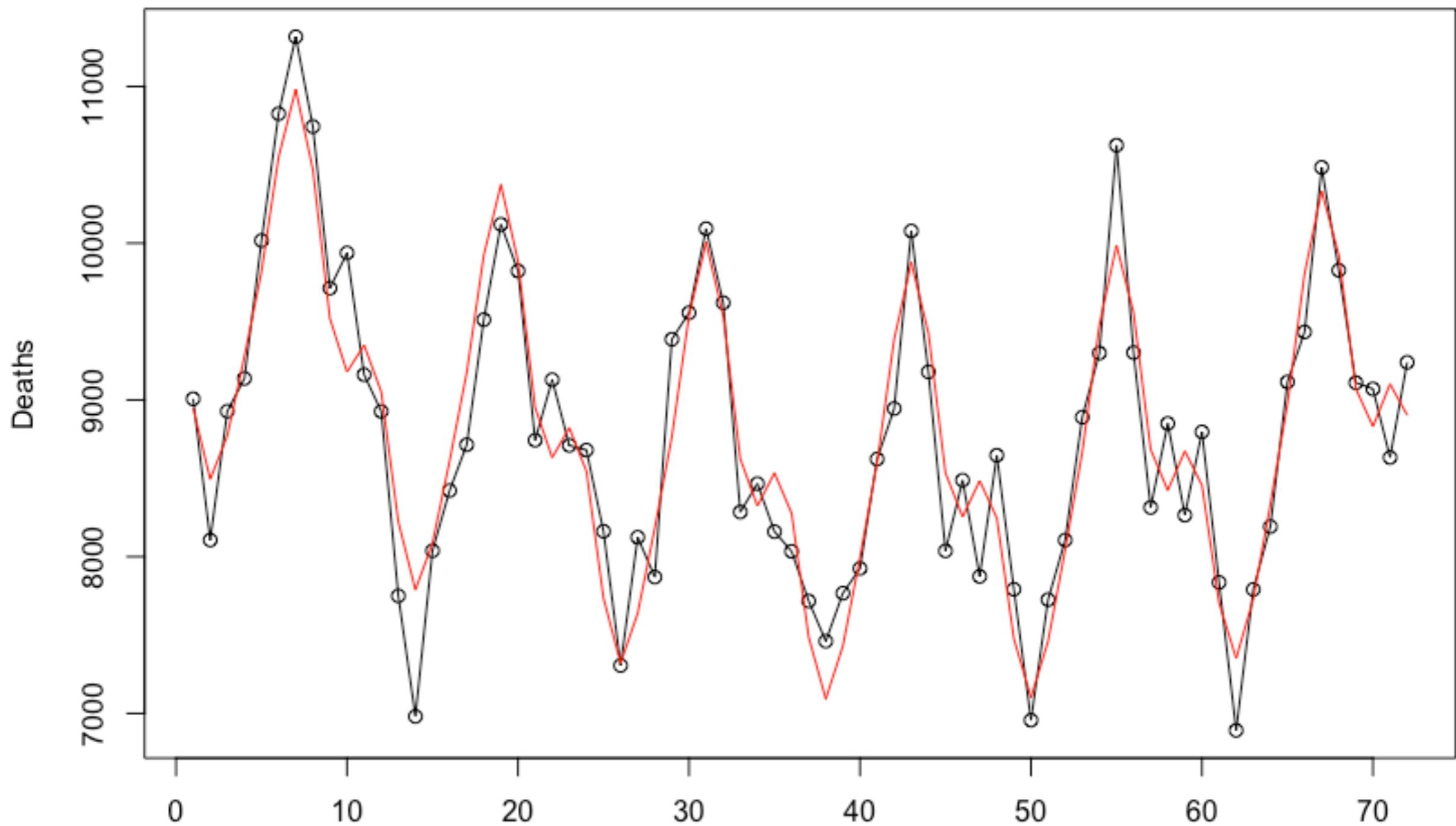
$\cos(\pi t/6), \sin(\pi t/6), \cos(\pi t/3), \sin(\pi t/3), \cos(\pi t/2), \sin(\pi t/2)$

Monthly Totals of Accidental Deaths in the US 1973-1978



$y_t \sim \text{sinusoids} + \text{quadratic}$

Monthly Totals of Accidental Deaths in the US 1973-1978



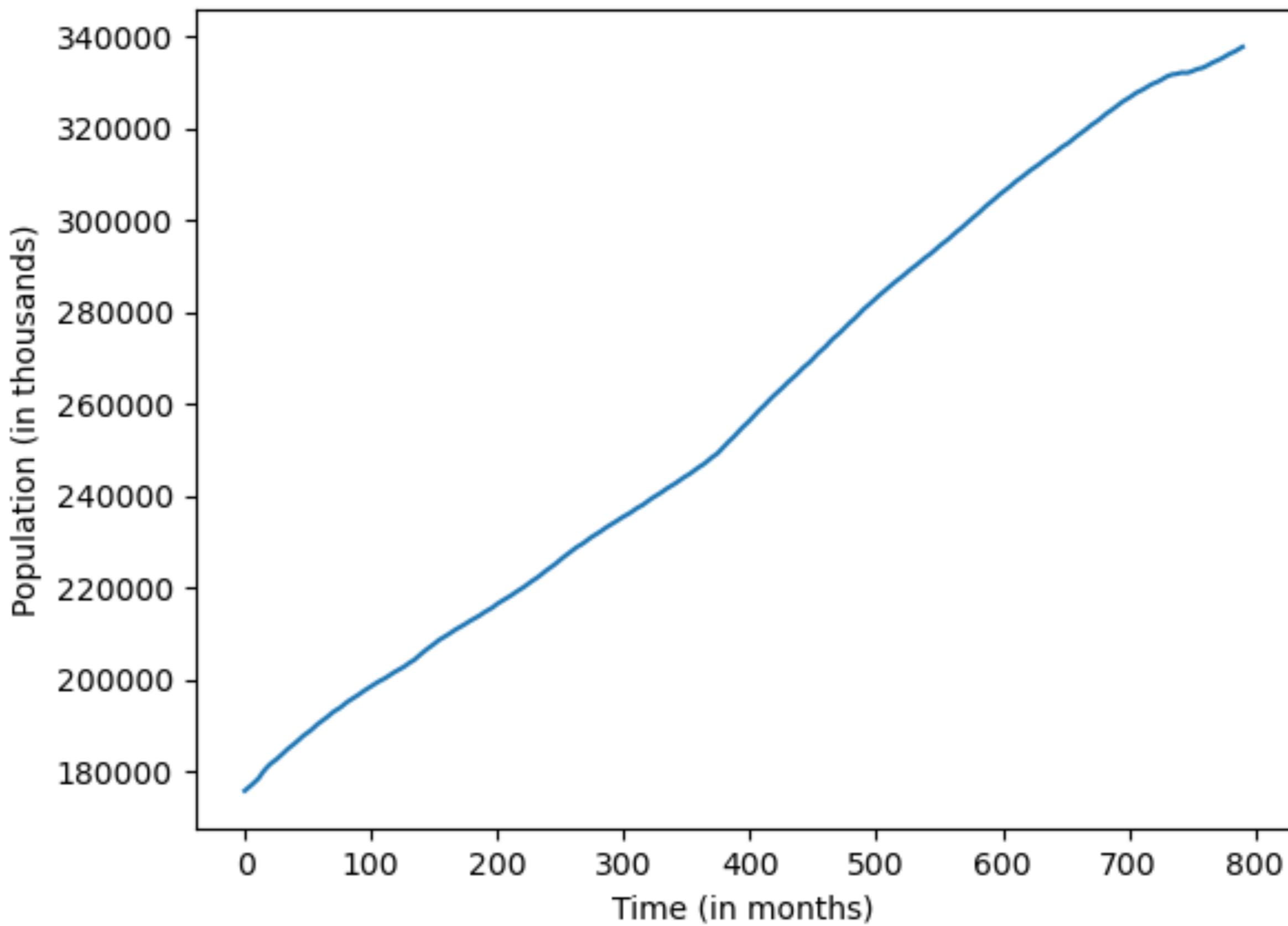
Topic One: Multiple Linear Regression

- These models clearly give (sometimes reasonable) solutions to the prediction problem
- The first topic in this course is multiple linear regression
- We will go over the usual frequentist inference but also discuss in detail Bayesian inference for linear regression

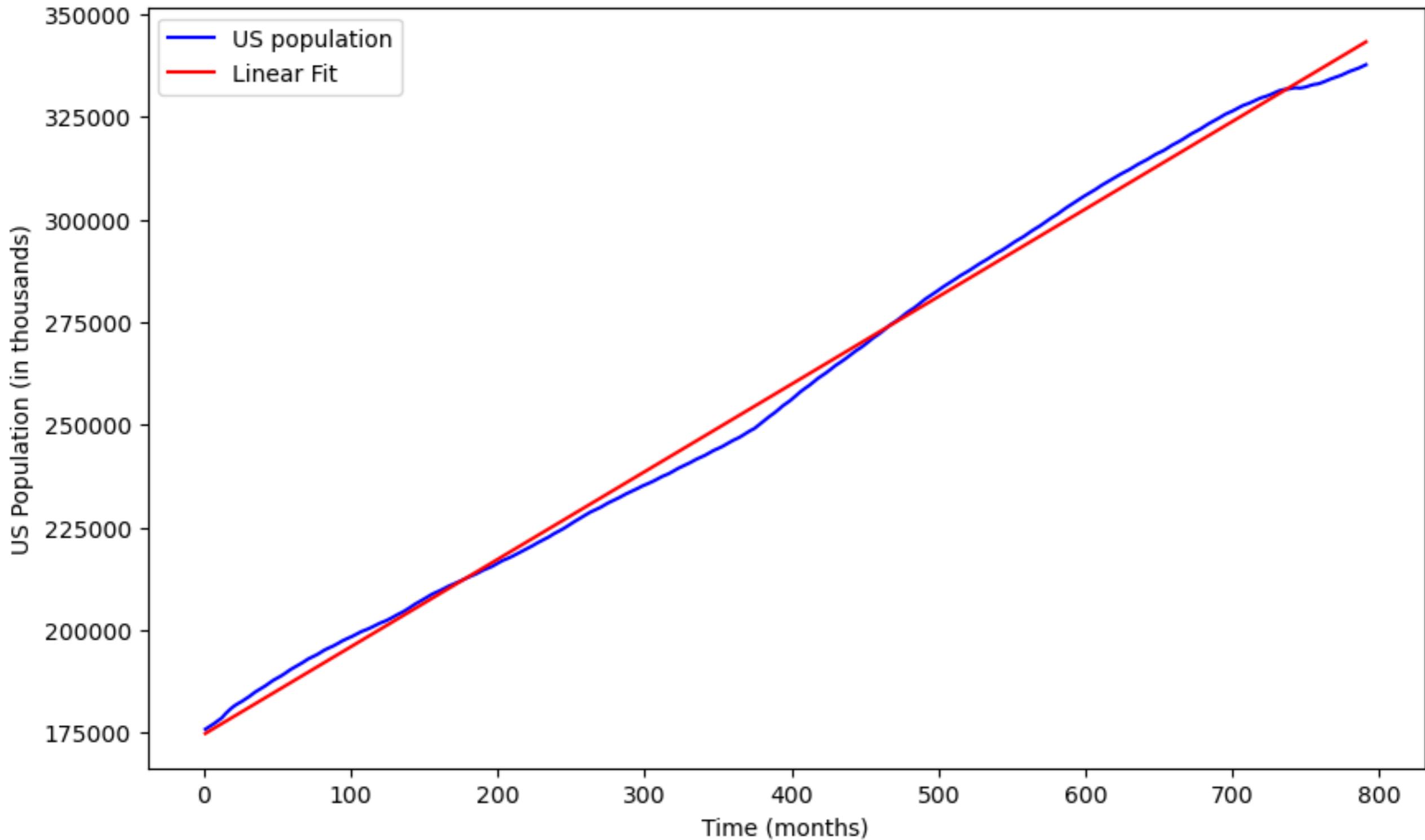
Topic Two: Nonlinear Regression

For many time series, nonlinear regression over t leads to much more useful and realistic models

US Population Data



US Population



- For the US population dataset, simple linear regression over t fits one line to the entire data which is clearly unrealistic
- Quadratic (and higher order polynomial) regression does not seem ideal either. These models are also not very interpretable in terms of growth rates

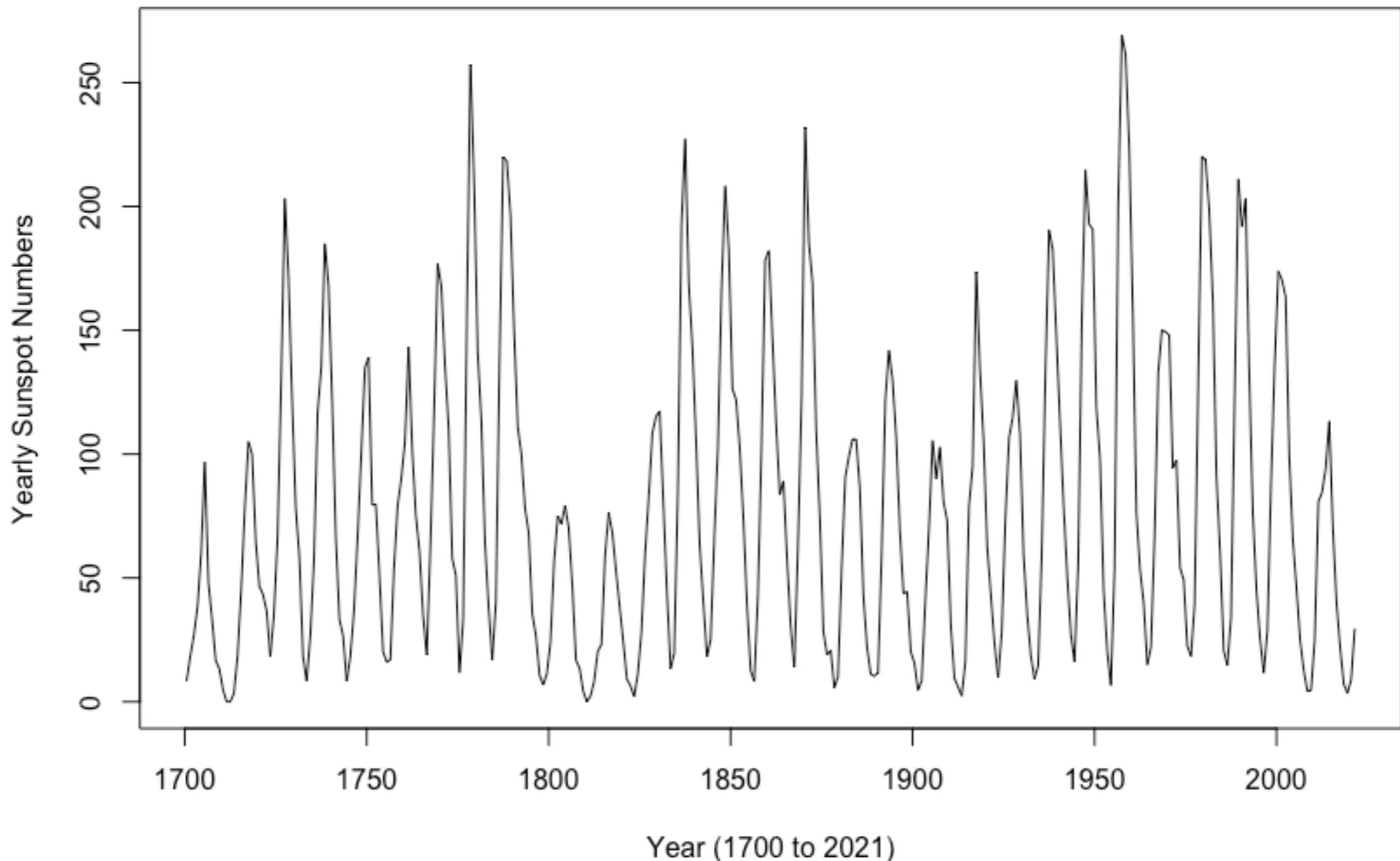
- A more realistic model here is:

$$y_t = \beta_0 + \beta_1 t + \alpha_1(t - c_1)_+ + \alpha_2(t - c_2)_+ + \text{error}$$

- This model allows for three different slopes (growth rates)
- This is a nonlinear regression model with parameters $\beta_0, \beta_1, \alpha_1, c_1, \alpha_2, c_2$

Annual Sunspots Data

Sunspot Data



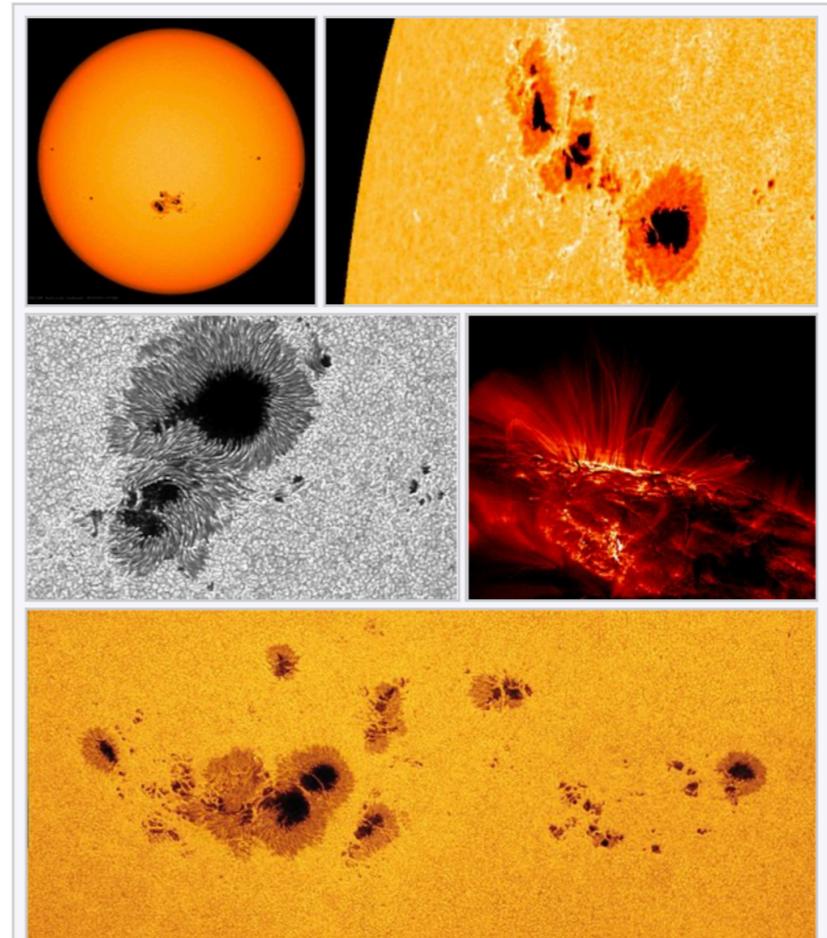
From Wikipedia, the free encyclopedia

For other uses, see [Sunspot \(disambiguation\)](#).

Sunspots are temporary spots on the [Sun's surface](#) that are darker than the surrounding area. They are one of the most recognizable [Solar phenomena](#) and despite the fact that they are mostly visible in the [solar photosphere](#) they usually affect the entire [solar atmosphere](#). They are regions of reduced surface temperature caused by concentrations of [magnetic flux](#) that inhibit [convection](#). Sunspots appear within [active regions](#), usually in pairs of opposite [magnetic polarity](#).^[2] Their number varies according to the approximately 11-year [solar cycle](#).

Individual sunspots or groups of sunspots may last anywhere from a few days to a few months, but eventually decay. Sunspots expand and contract as they move across the surface of the Sun, with diameters ranging from 16 km (10 mi)^[3] to 160,000 km (100,000 mi).^[4] Larger sunspots can be visible from Earth without the aid of a [telescope](#).^[5] They may travel at [relative speeds](#), or [proper motions](#), of a few hundred meters per second when they first emerge.

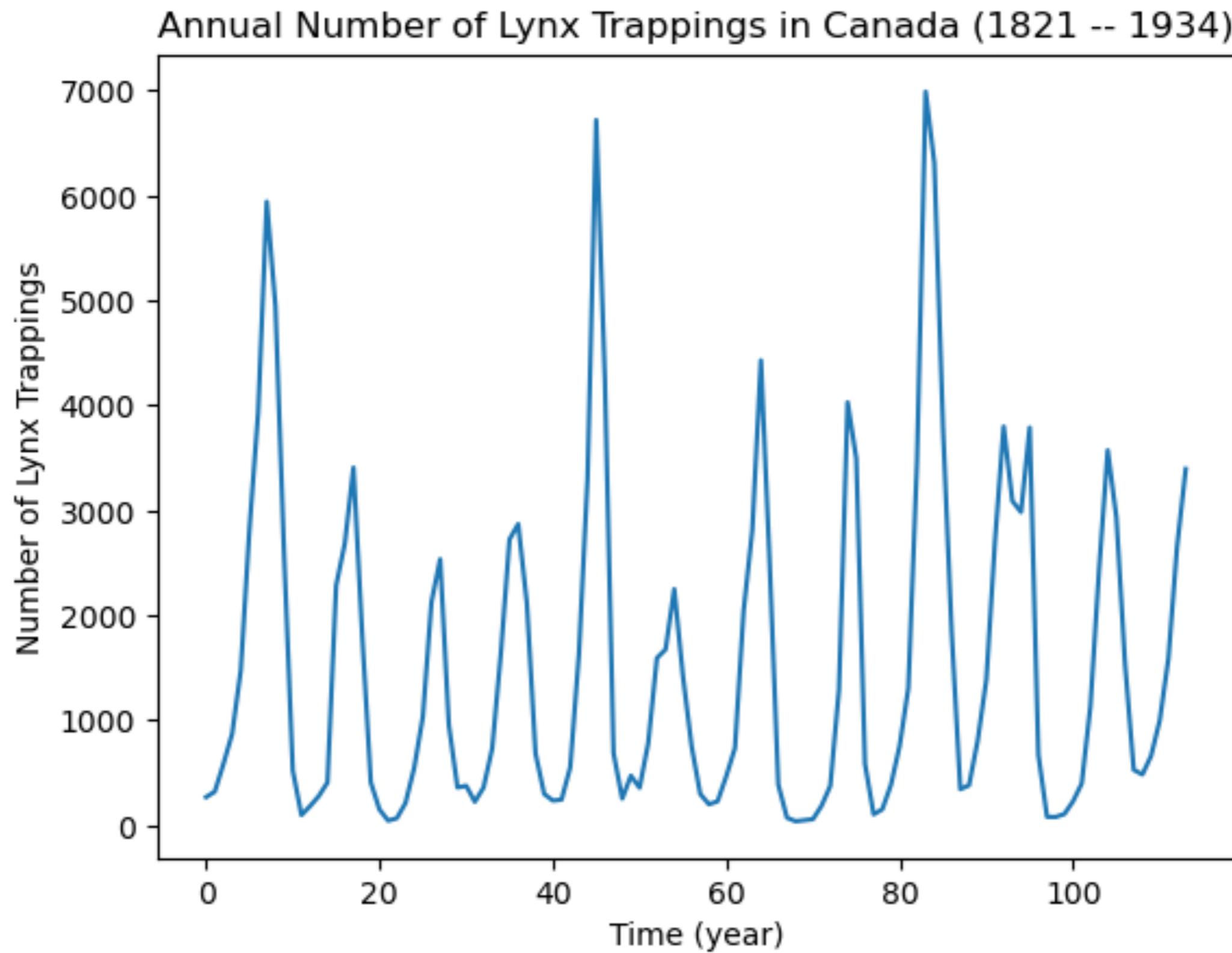
Indicating intense magnetic activity, sunspots accompany other active region phenomena such as [coronal loops](#), [prominences](#), and [reconnection](#) events. Most [solar flares](#) and [coronal mass ejections](#) originate in these magnetically active regions around visible sunspot groupings. Similar phenomena indirectly observed on [stars](#) other than the Sun are commonly called [starspots](#), and both light and dark spots have been measured.^[6]



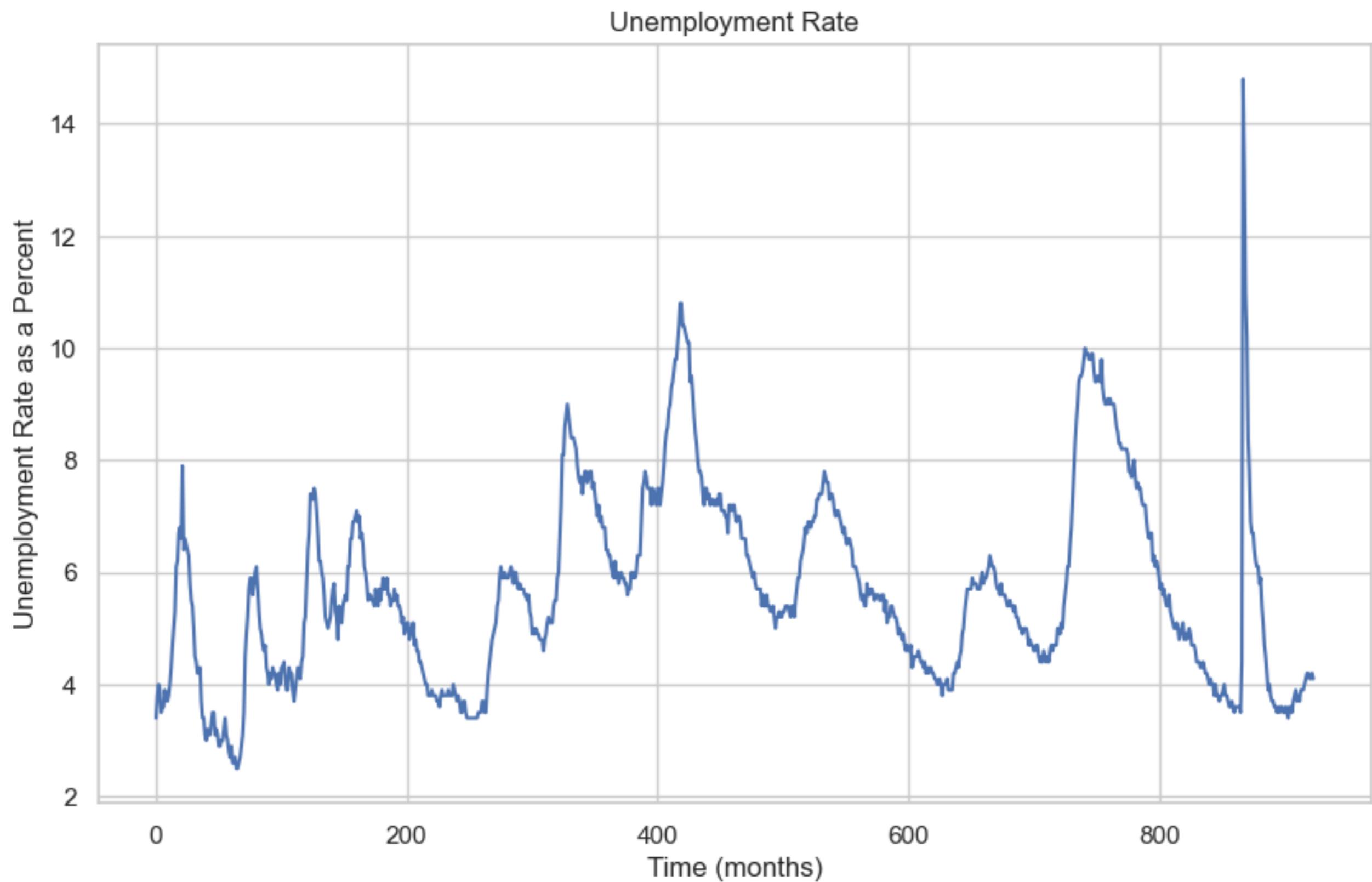
- Top: active region 2192 in 2014 containing the largest sunspot of [solar cycle 24](#)^[1] and active region 1302 in September 2011.
- Middle: sunspot close-up in the visible spectrum (left) and another sunspot in [UV](#), taken by the [TRACE](#) observatory.
- Bottom: a large group of sunspots

- Wikipedia says that the number of sunspots varies according to the 11 year solar cycle
- Why should the periodicity be exactly 11? Why not 10.5 or 11.5? What is the uncertainty around 11?
- Can the periodicity be figured out from the dataset?
- One way to do this is to fit the model:
$$y_t = \beta_0 + \beta_1 \cos(\omega t) + \beta_2 \sin(\omega t) + \text{error}$$
- This is a nonlinear regression model with parameters $\beta_0, \beta_1, \beta_2, \omega$

Lynx Trappings Dataset



Unemployment Rate from FRED

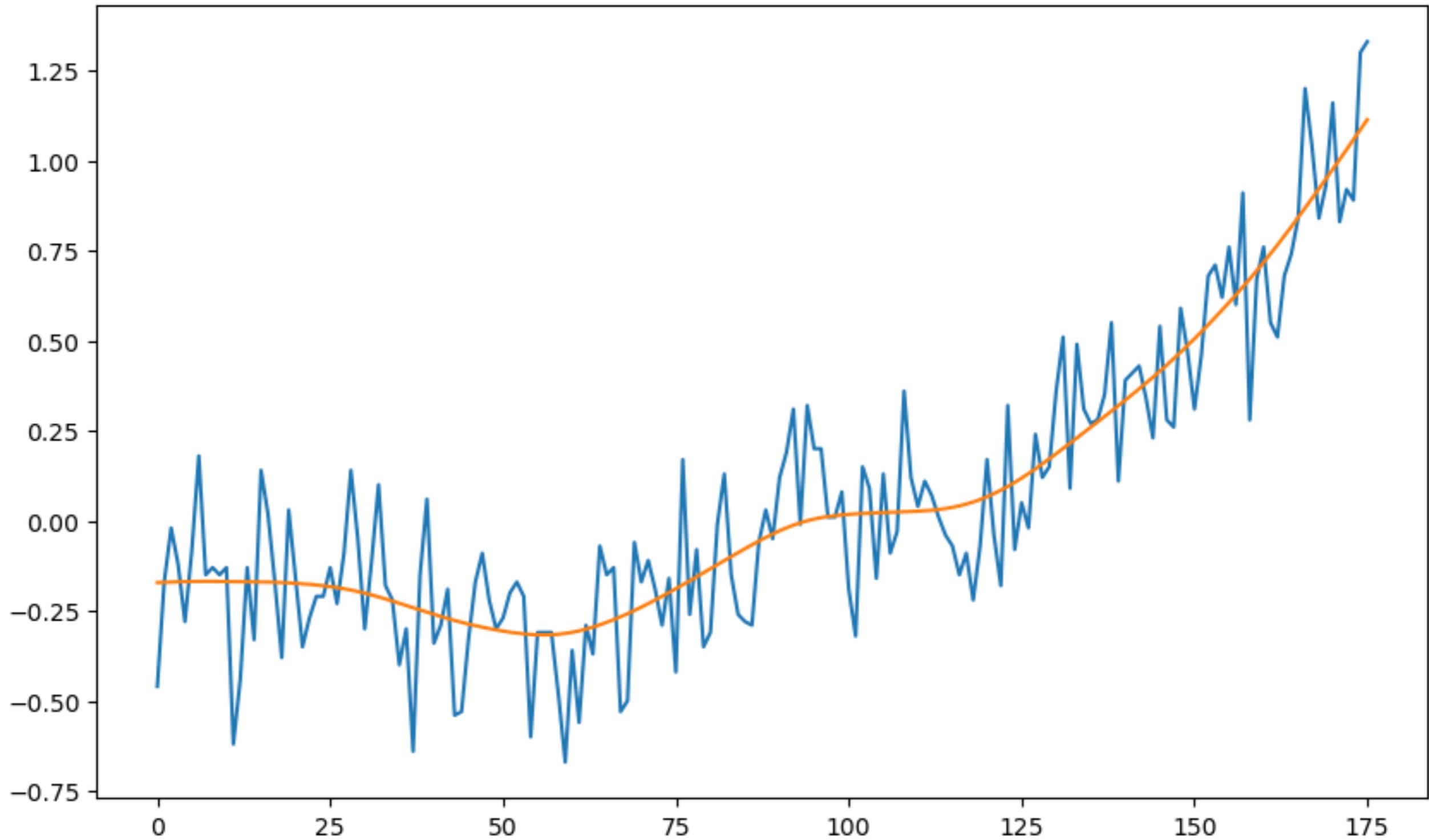


- The model:
 $y_t = \beta_0 + \beta_1 \cos(\omega t) + \beta_2 \sin(\omega t) + \text{error}$ is closely related to **Fourier Analysis**
- We will study concepts such as **Fourier Frequencies**, **Discrete Fourier Transform** and the **Periodogram**
- These are extensively used in engineering time series analysis
- We shall look at some practical applications of these concepts

Topic Three: High-dimensional Regression

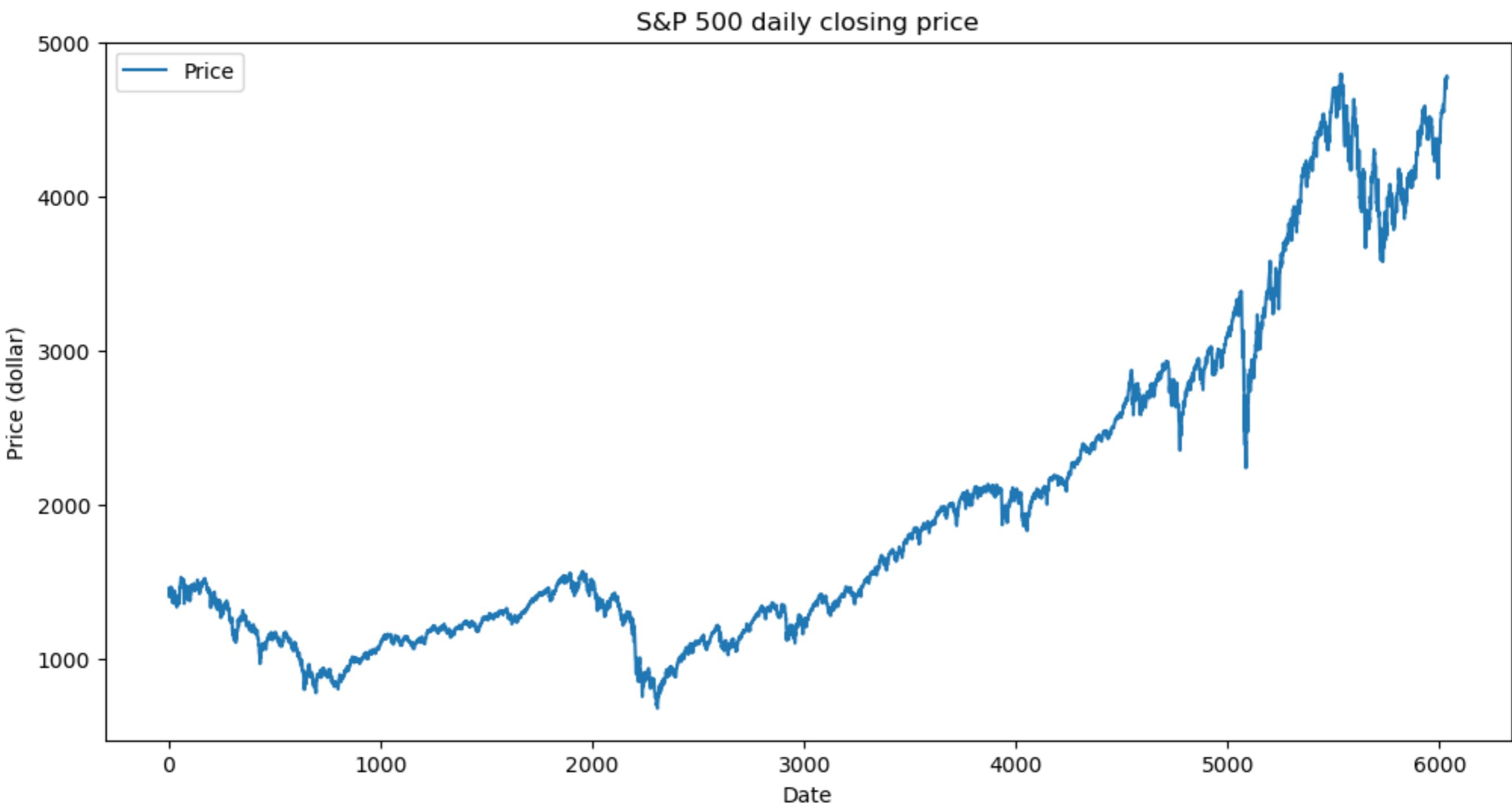
- It is sometimes tempting to throw in a large number of variables while regressing over time
- For example, consider the model:
$$y_t = \beta_0 + \beta_1 t + \beta_2(t - 2)_+ + \beta_3(t - 3)_+ + \dots + \beta_n(t - n)_+ + \epsilon$$
- This model allows a different growth rate between every two time points
- To fit such models sensibly, one would need to employ regularization
- We shall study the Ridge and LASSO regularizations

Here is this model (with Ridge regularization)
applied to a dataset on temperature anomalies



Topic Four: Variance Modeling

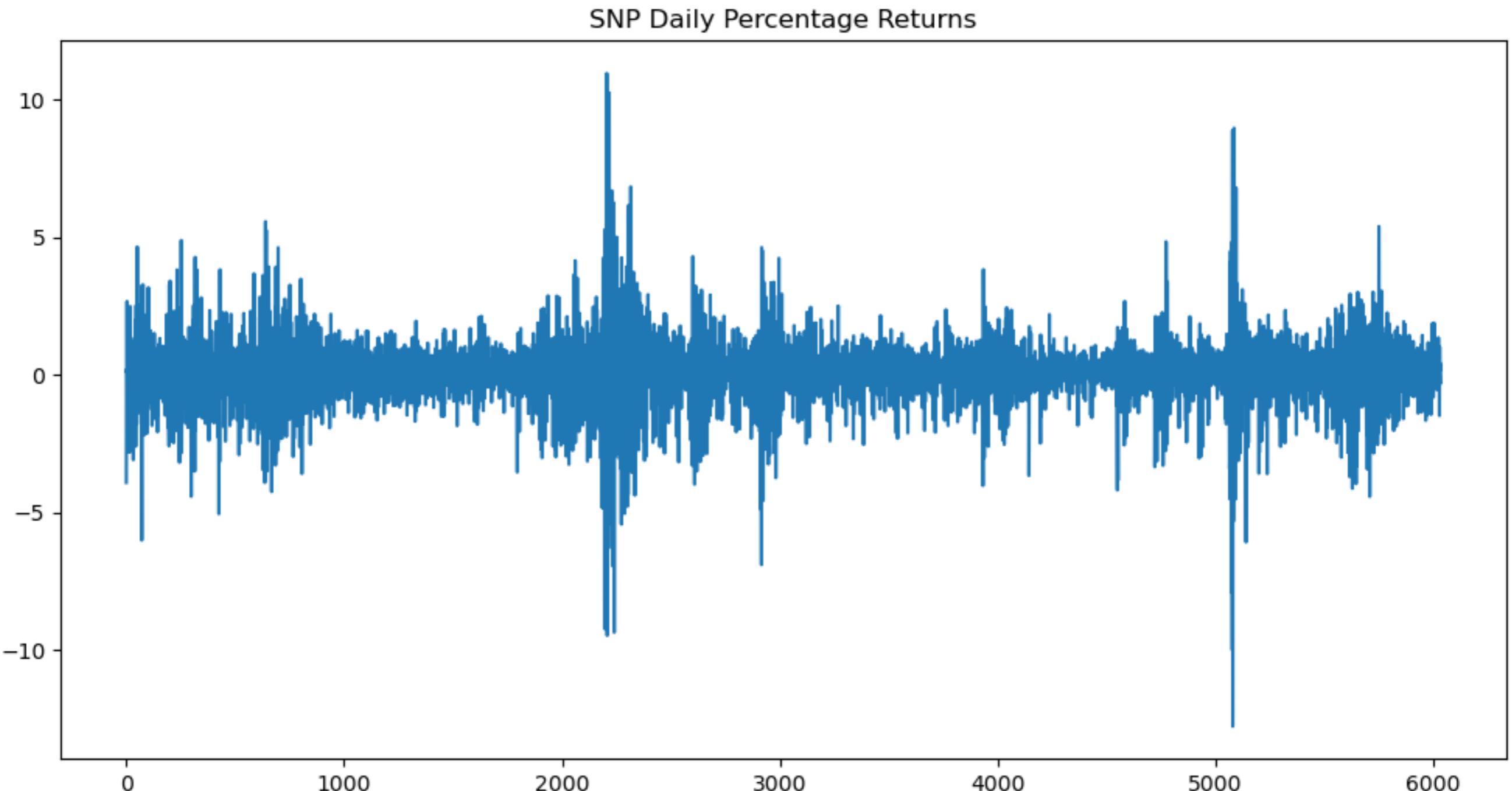
Daily closing prices P_t of S&P 500 (2000-01-01 to 2024-01-01)



Stock Returns r_t

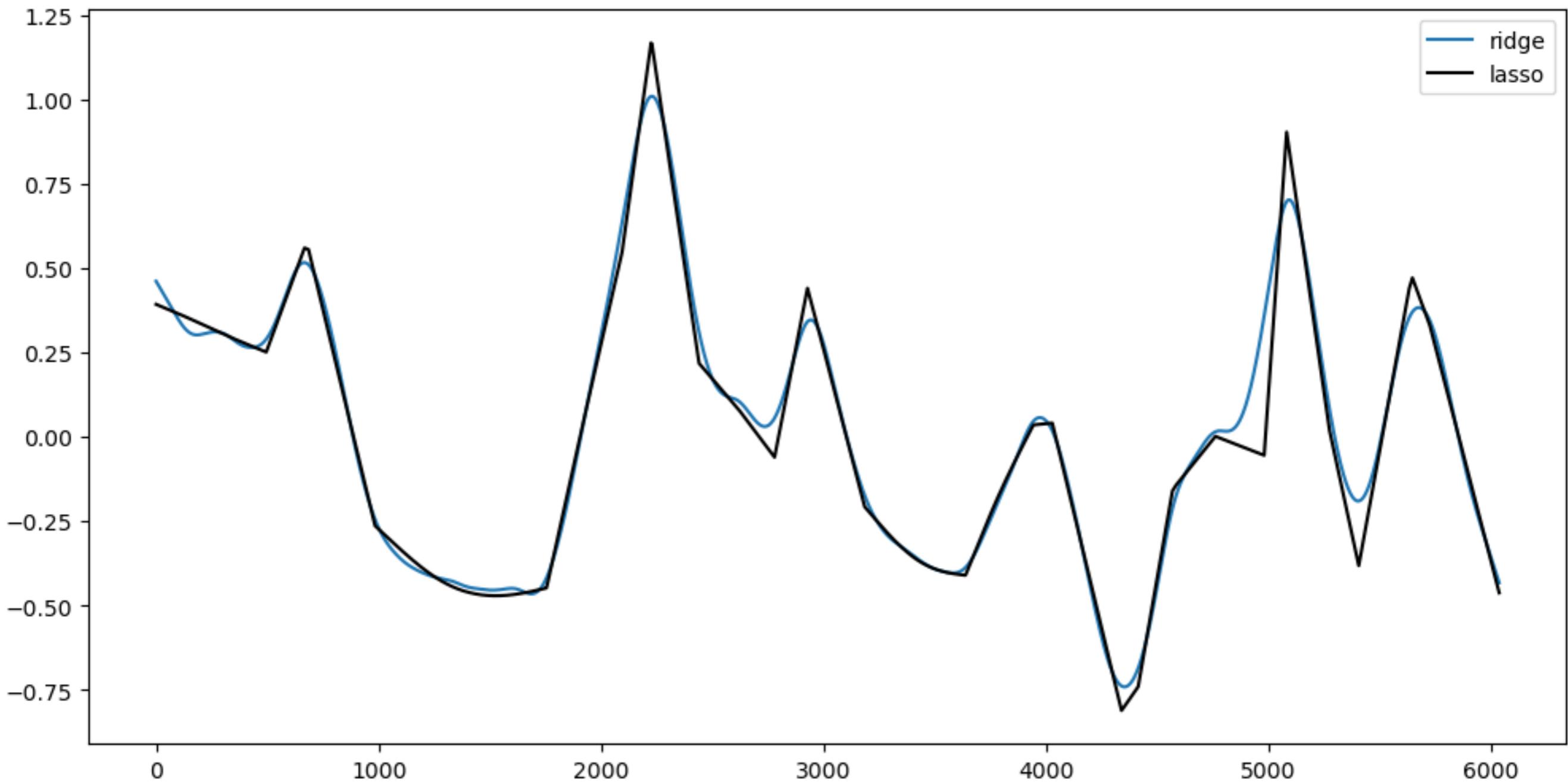
$$r_t = 100 \times (\log P_t - \log P_{t-1})$$

Note $r_t \approx 100 \times \frac{P_t - P_{t-1}}{P_{t-1}}$

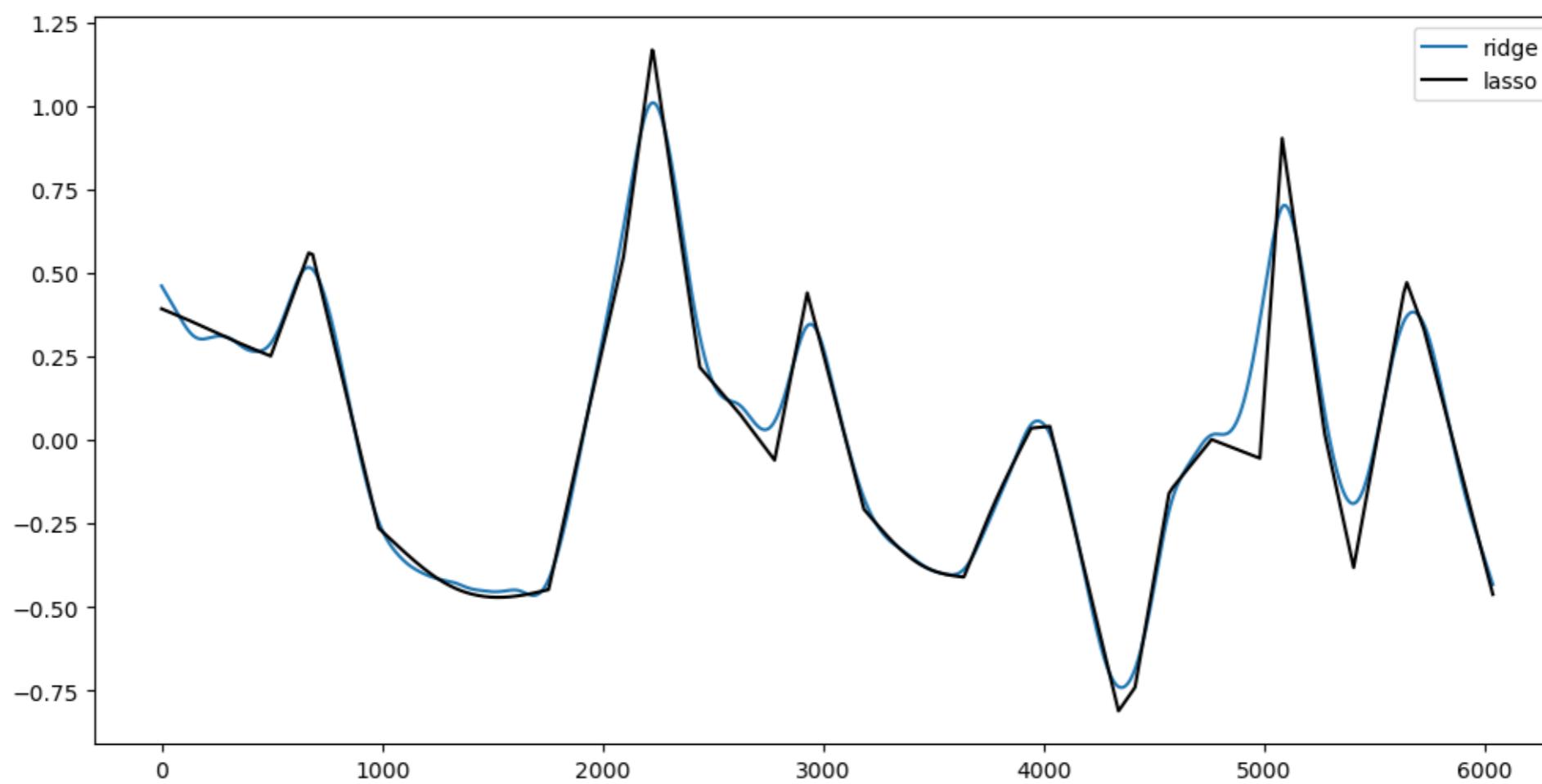
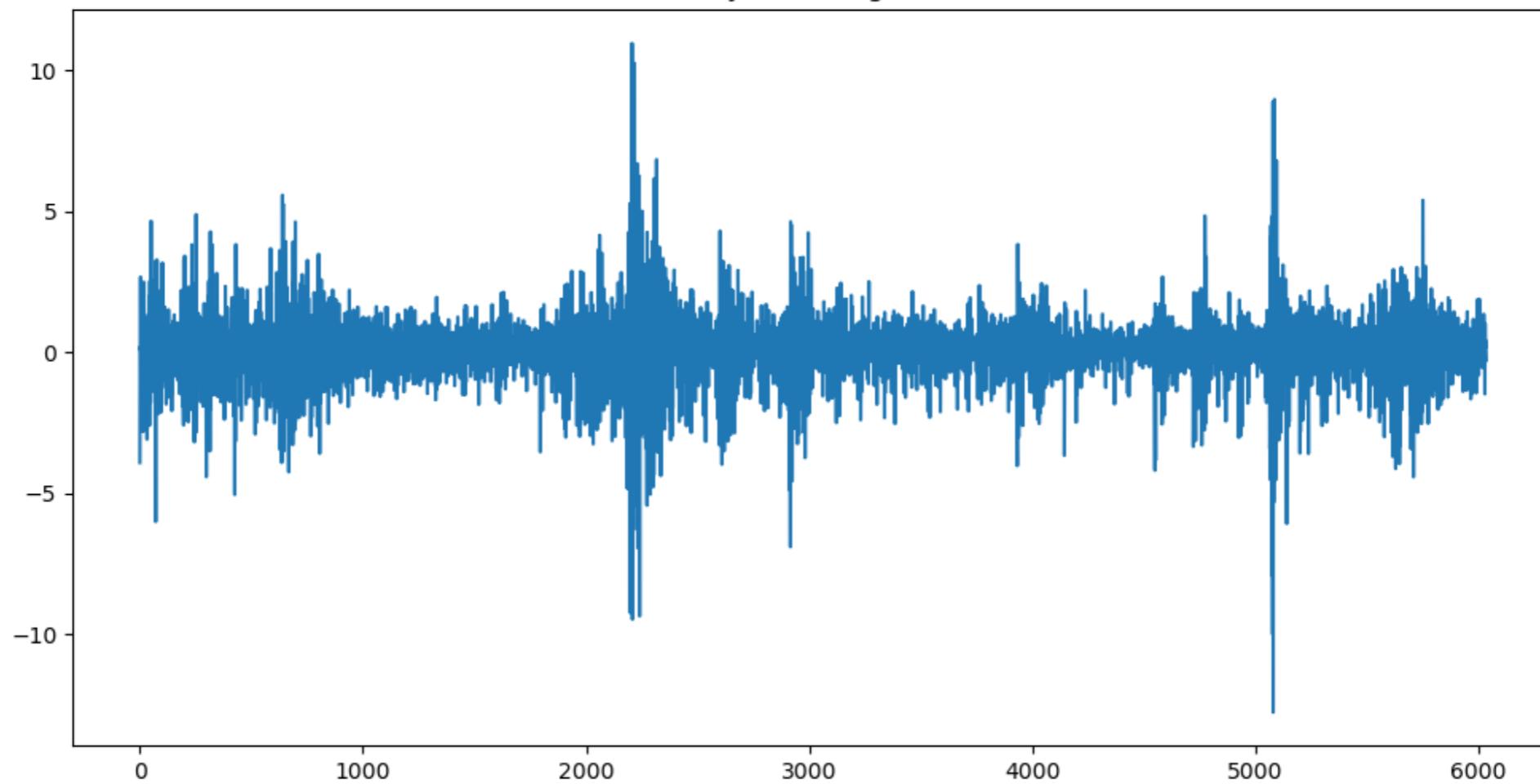


- Financial analysts also study the volatility of stock price returns
- For estimating volatility, it is common to use the model: $r_t \sim N(0, \sigma_t^2)$ and then to model σ_t (which is a proxy for volatility) as a function of t
- These are examples of variance models as opposed to the mean (regression) models we saw so far

For the S&P returns data, below is an estimate of $\log \sigma_t$



SNP Daily Percentage Returns

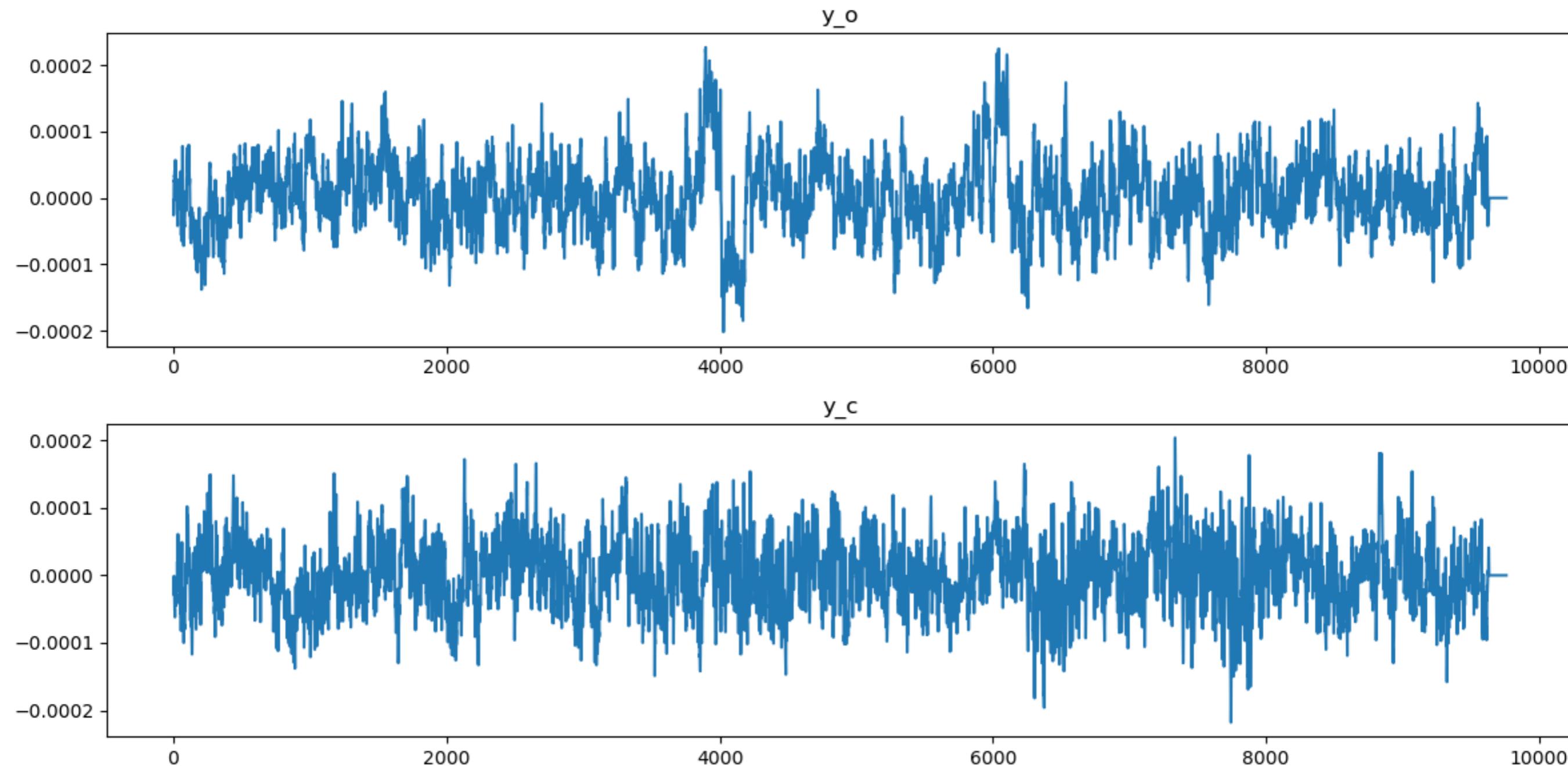


Spectral Analysis

Spectral Analysis is a special case of variance modeling where the variance model is applied to the Discrete Fourier Transform of the data (as opposed to the data directly)

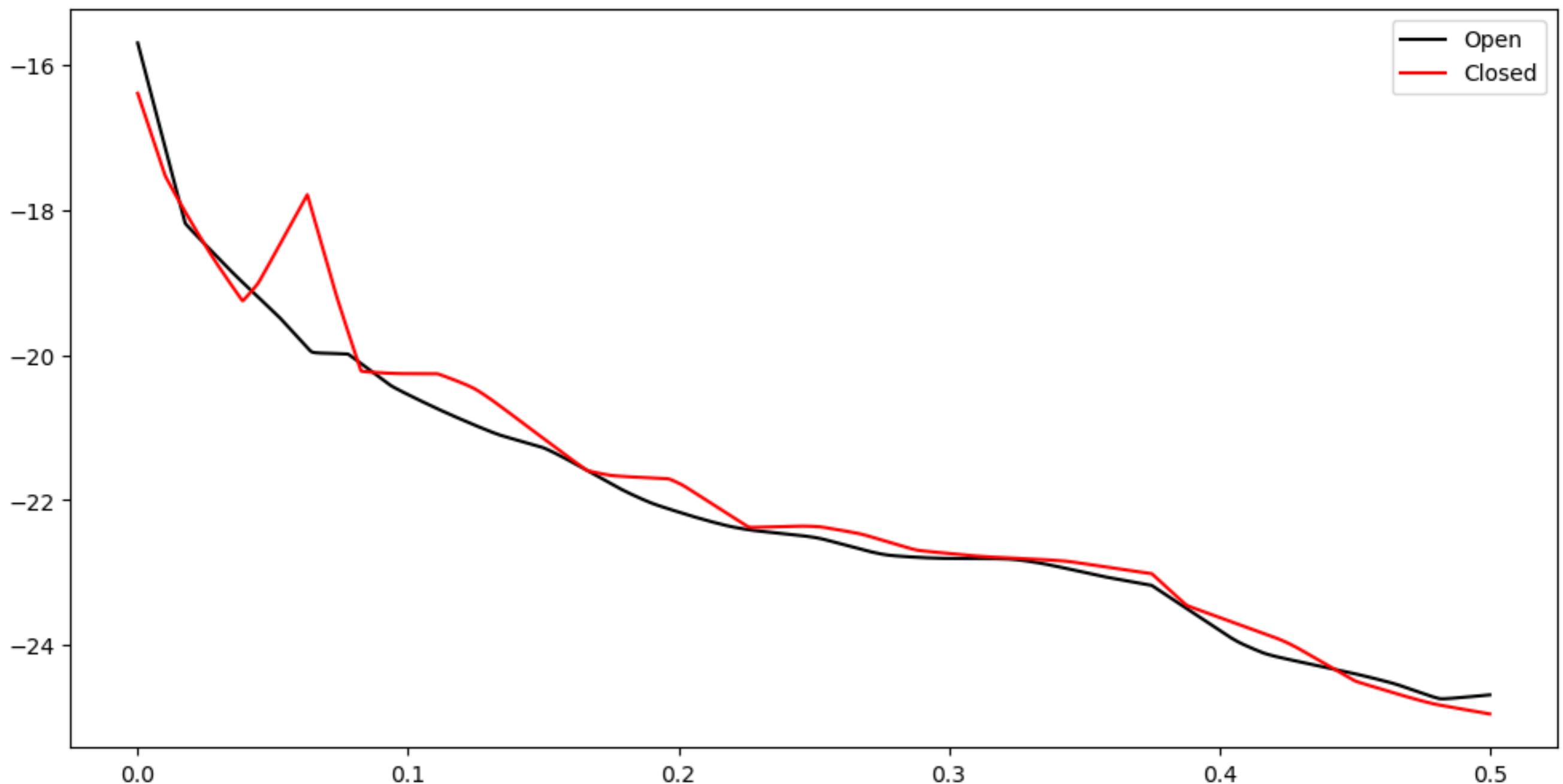
This is one of the most important tools used by Engineers for Signal Processing

EEG Example: Eyes open vs Closed



How to quantify the differences between these two time series?

Spectra estimates (multiply x axis by 160 for Hz):



There is a clear difference in the spectra at around the 10 Hz frequency which is a known fact from cognitive neuroscience.

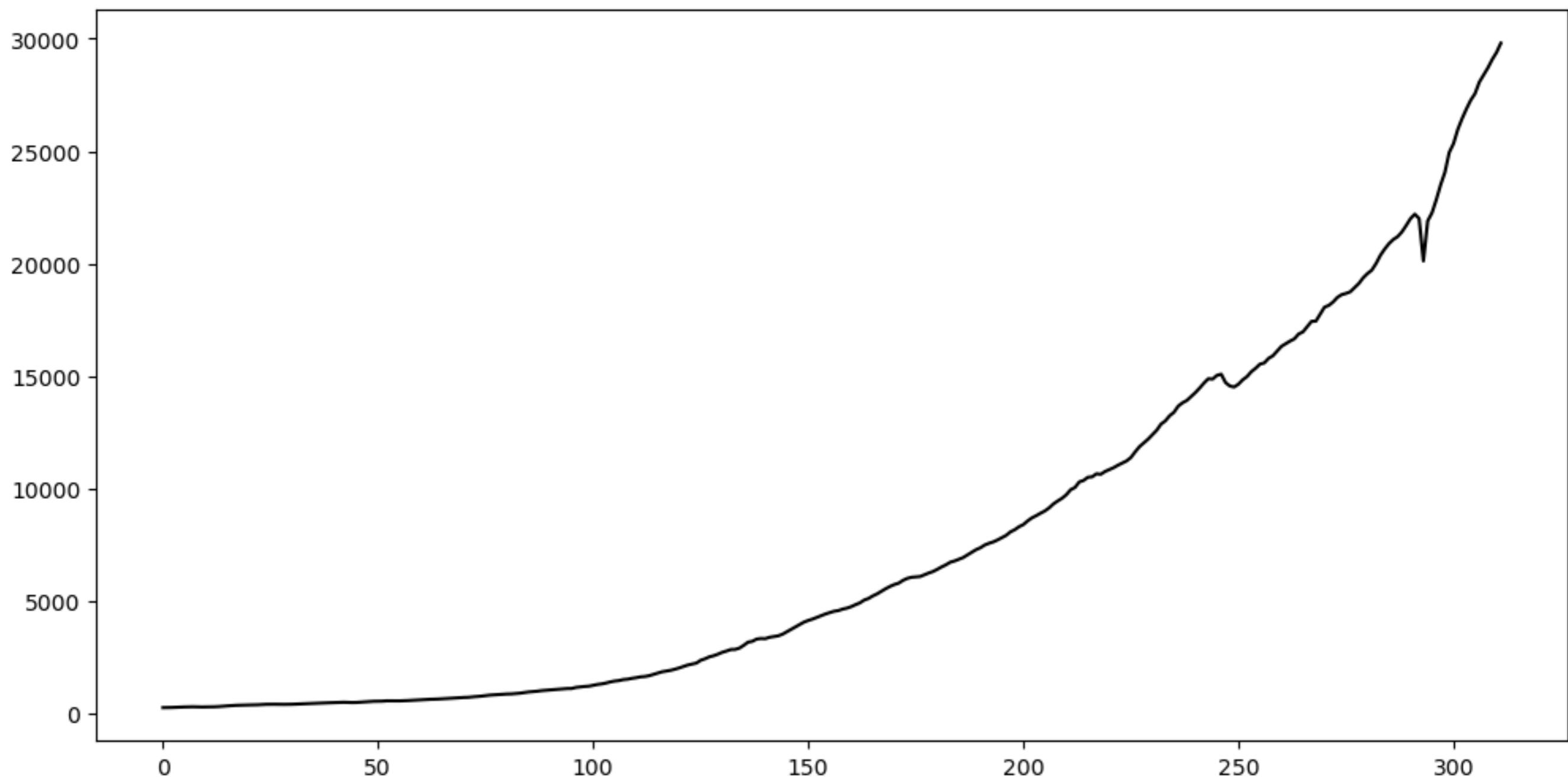
Topic Five: Lagged Regression (ARIMA)

- Find next number: 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, #
- This is the Fibonacci sequence ($y_t = y_{t-1} + y_{t-2}$) and the next number is 144
- Here regression over time will not work
- Instead, we have to regress y_t over its own lagged values y_{t-1} and y_{t-2}
- This is called Lagged Regression or AutoRegression

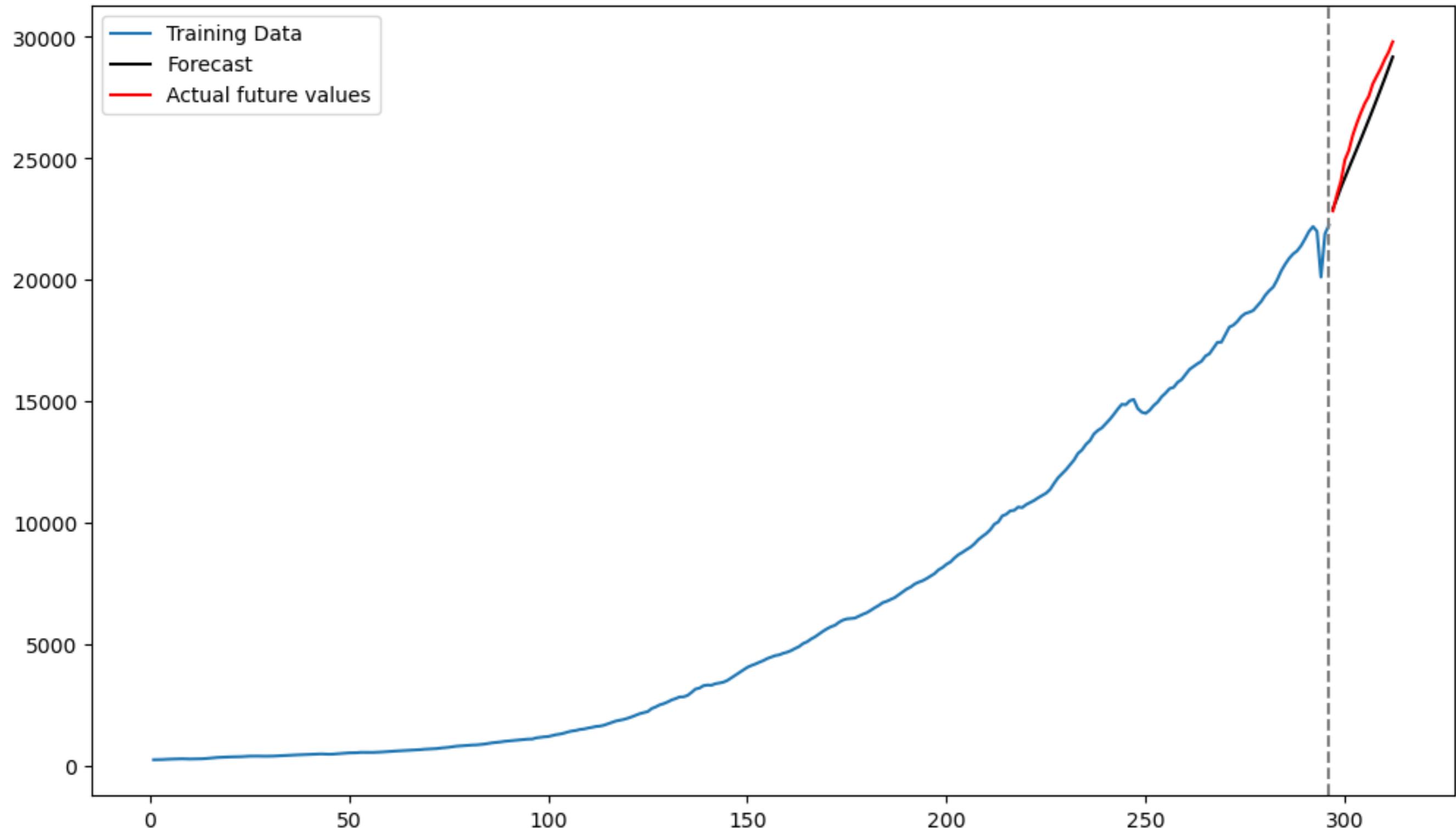
- AutoRegression is the main idea behind the ARIMA class of models which are widely used for time series prediction
- ARIMA stands for AutoRegressive Integrated Moving Average Models
- The basic idea is simply to do a linear regression of y_t on $x_t = (1, y_{t-1}, \dots, y_{t-p})$ for some fixed lag $p \geq 1$
- ARIMA models give decent predictions on real datasets and are used extensively

ARIMA Example One

FRED data on GNP (seasonally adjusted, billions of dollars). This is a quarterly dataset (four quarters in a year).

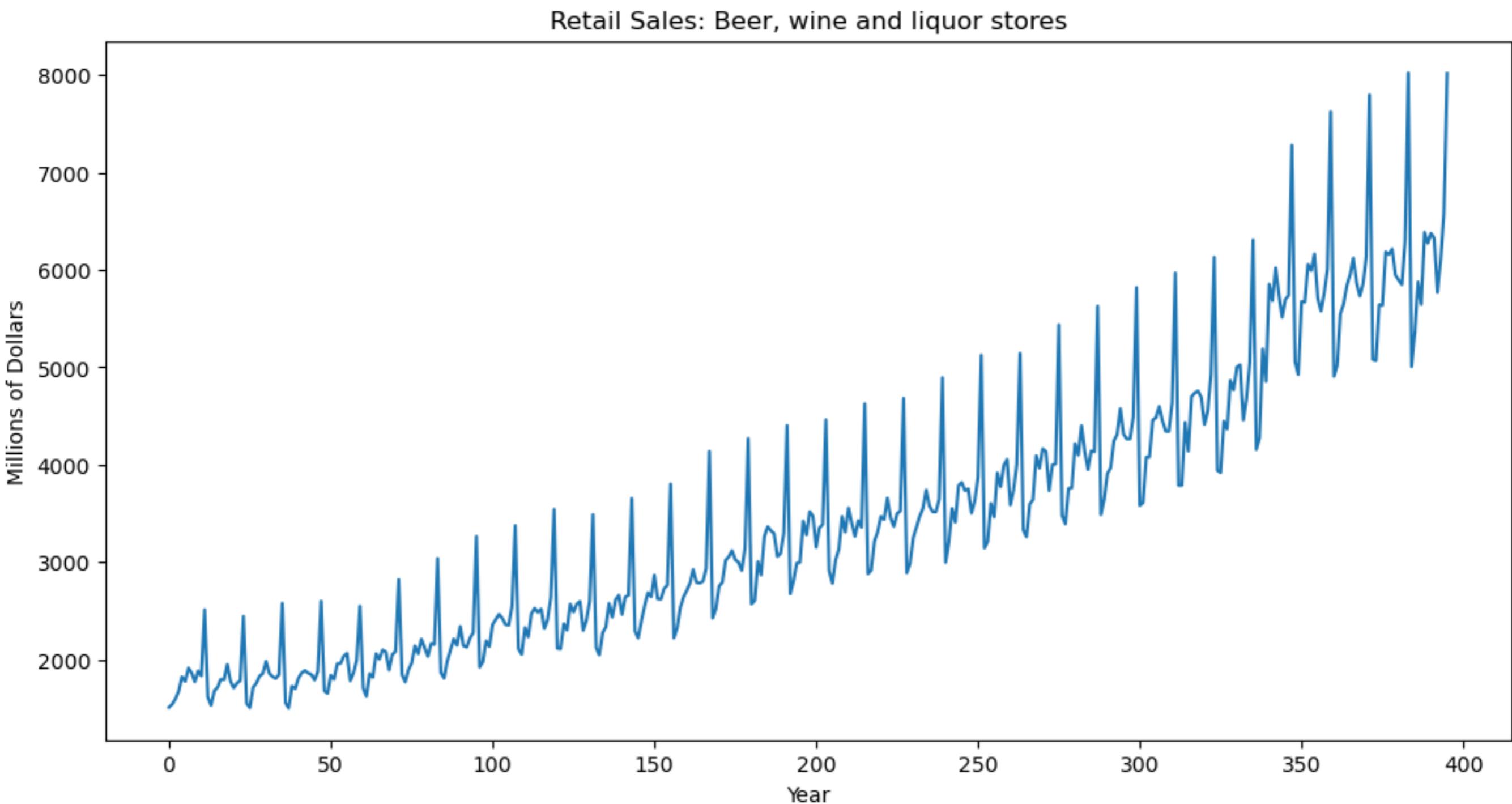


Here are the predictions given by an AR model for the last four years (= 16 quarters).

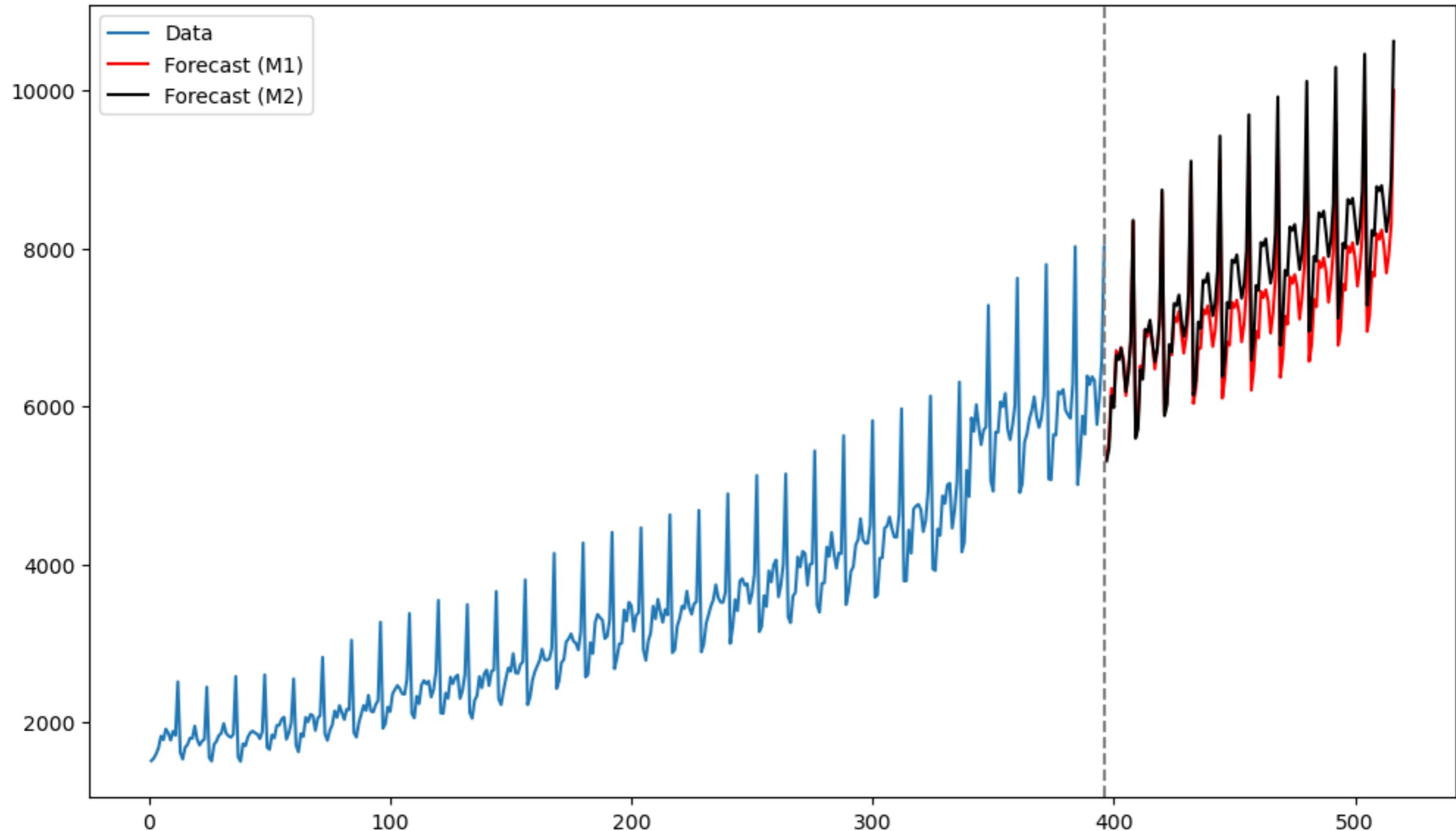


ARIMA Example Two

FRED data on monthly retail sales (in millions of dollars) for beer, wine and liquor stores (typo on axis label: should be month not year)



Below we plot the data along with predictions by two different ARIMA models



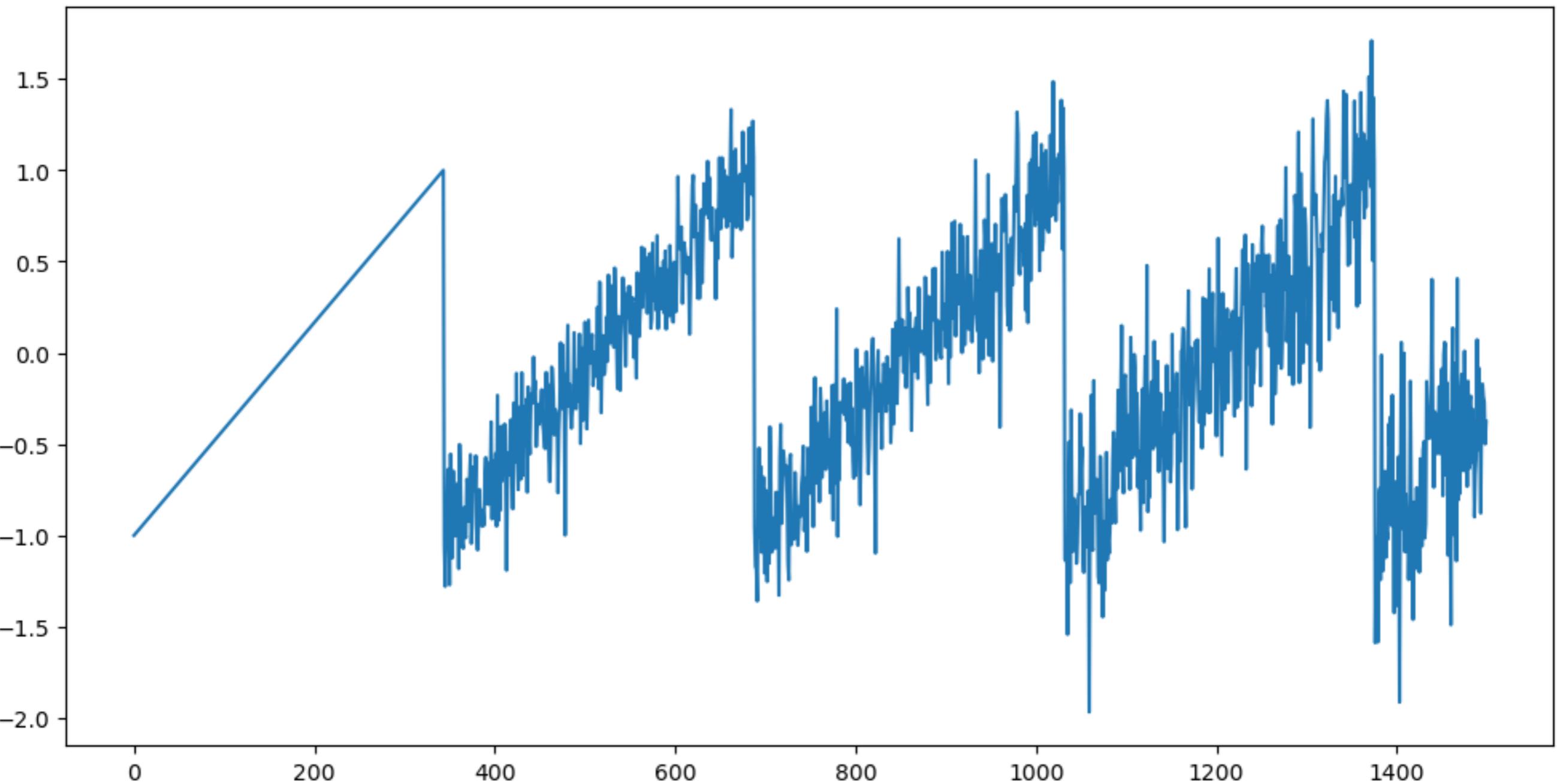
Topic Six: Recurrent Neural Networks

- AR models are simply linear regression of y_t on lagged covariates $x_t := (1, y_{t-1}, \dots, y_{t-p})$
- It is natural to use nonlinear regression of y_t on x_t for improved predictive power
- We perform this nonlinear regression using **single hidden-layer** neural networks
- These is Nonlinear AutoRegression which fits functions $y_t = f(x_t) = f(y_{t-1}, \dots, y_{t-p})$

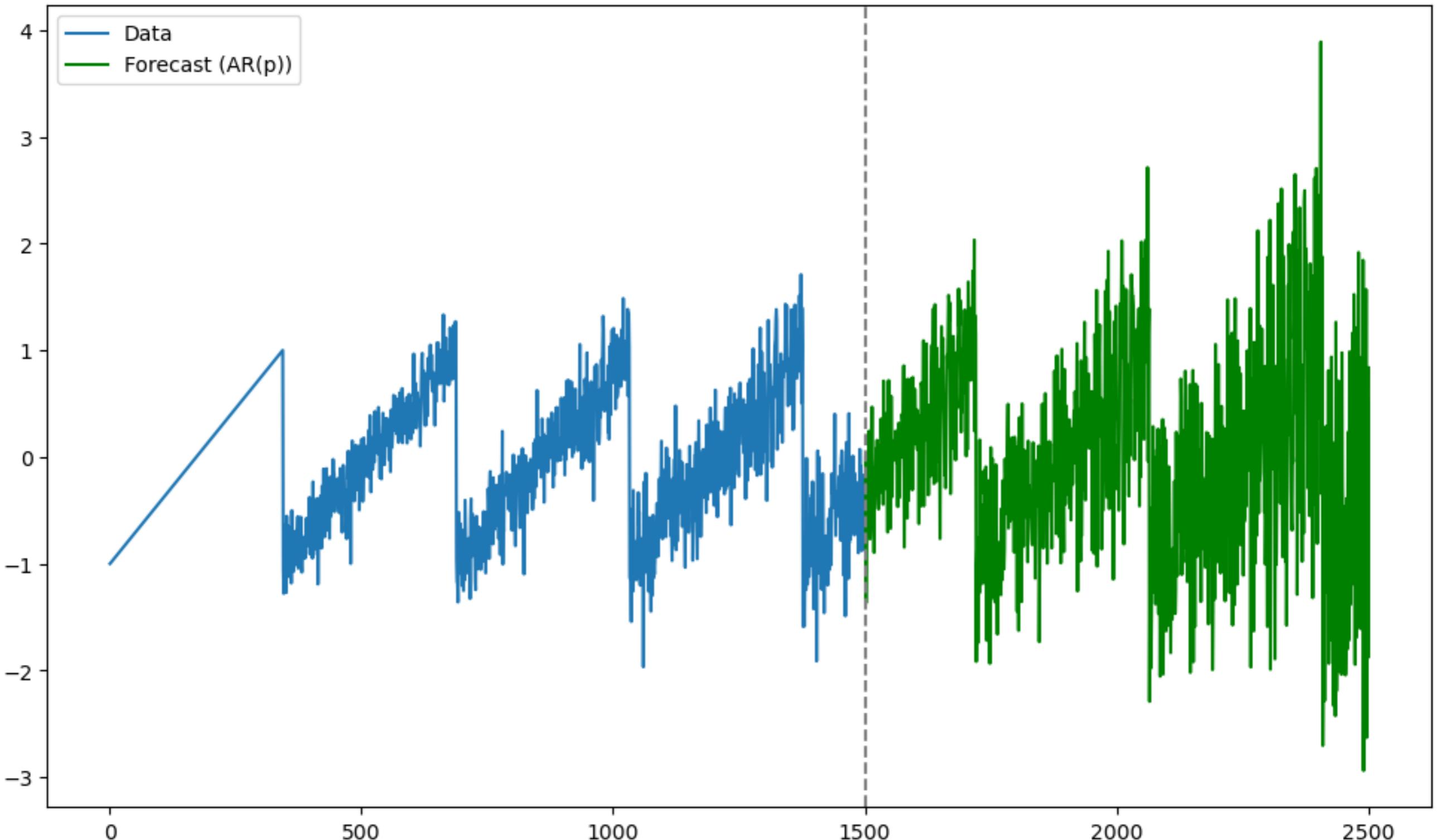
Topic Six: Recurrent Neural Networks

- A drawback of these models is that we have to select the value of p appropriately
- Even if p is chosen well, the dependence of y_t on its past values abruptly cuts off at p
- RNNs are a way of fitting functions which use all the past values: $y_t = f_t(x_t, x_{t-1}, \dots, x_1)$
- We shall study vanilla RNNs and more fancy variants such as LSTMs and GRUs

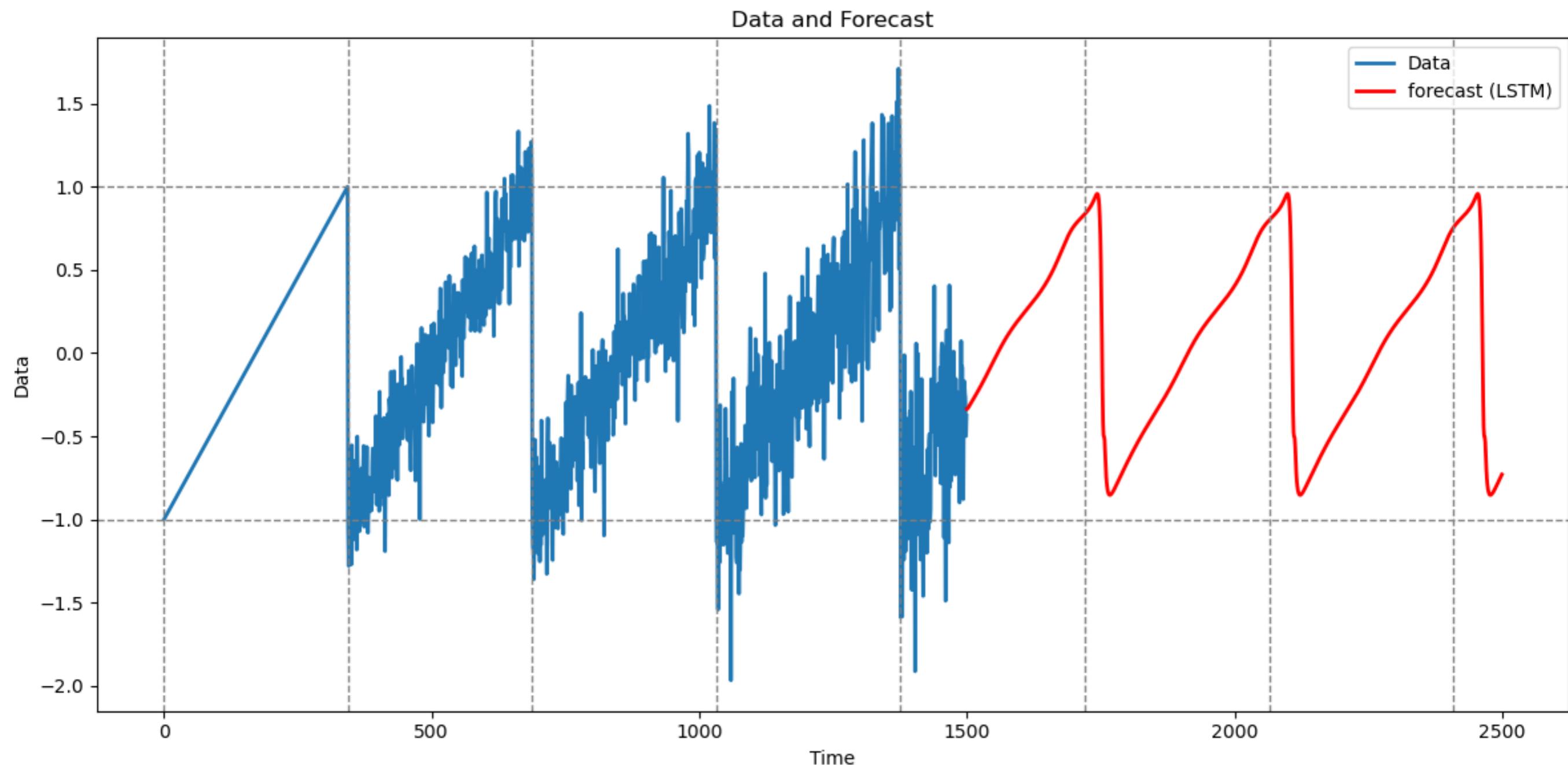
Here is a simulated data generated as: $y_t = y_{t-p} + \epsilon_t$ for $p = 344$



Here are the forecasts by the AR model with the correct p :



Here are the forecasts by an LSTM (which does not use knowledge of the true p):



- Topic 1: Multiple Linear Regression
- Topic 2: Nonlinear Regression
- Topic 3: High-dimensional Regression
- Topic 4: Variance Models and Spectral Analysis
- Topic 5: ARIMA modeling
- Topic 6: Recurrent Neural Networks

- We will work with a variety of models and will apply them to a variety of real time series datasets
- We fit models to data mostly using likelihood maximization (sometimes with additional regularization terms)
- When uncertainty quantification is desired, we employ Bayesian methods