Lecture One

STAT 153/STAT 248, Spring 2025

Aditya Guntuboyina (21 January 2025)

What is Time Series?

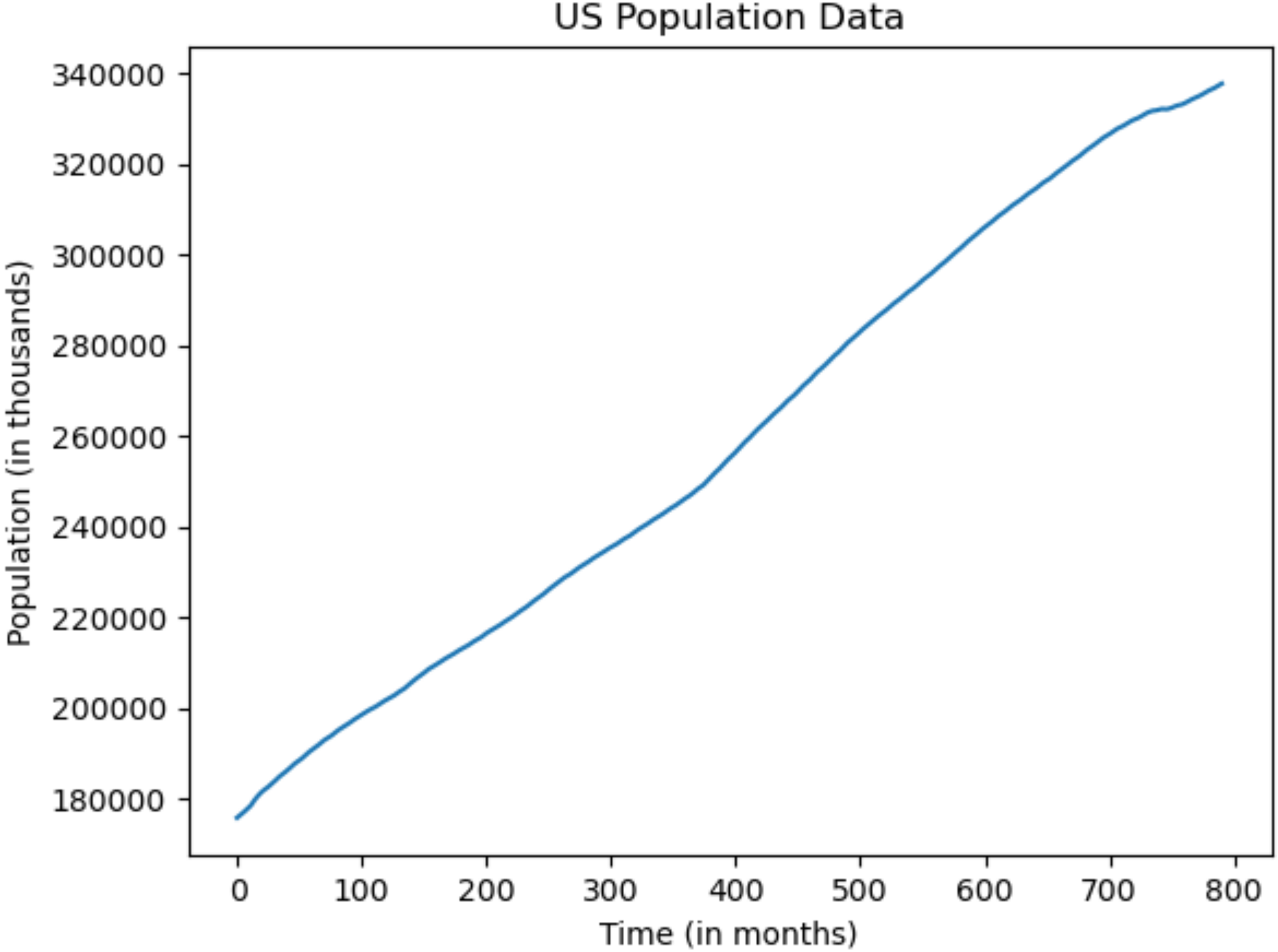
Time series is a set of observations each one being recorded at a specific time

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
340000
320000
300000
280000
260000
240000
220000
200000
180000
Population (in thousands)
0
100
200
300
400
500
600
700
800
Time (in months)

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)

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

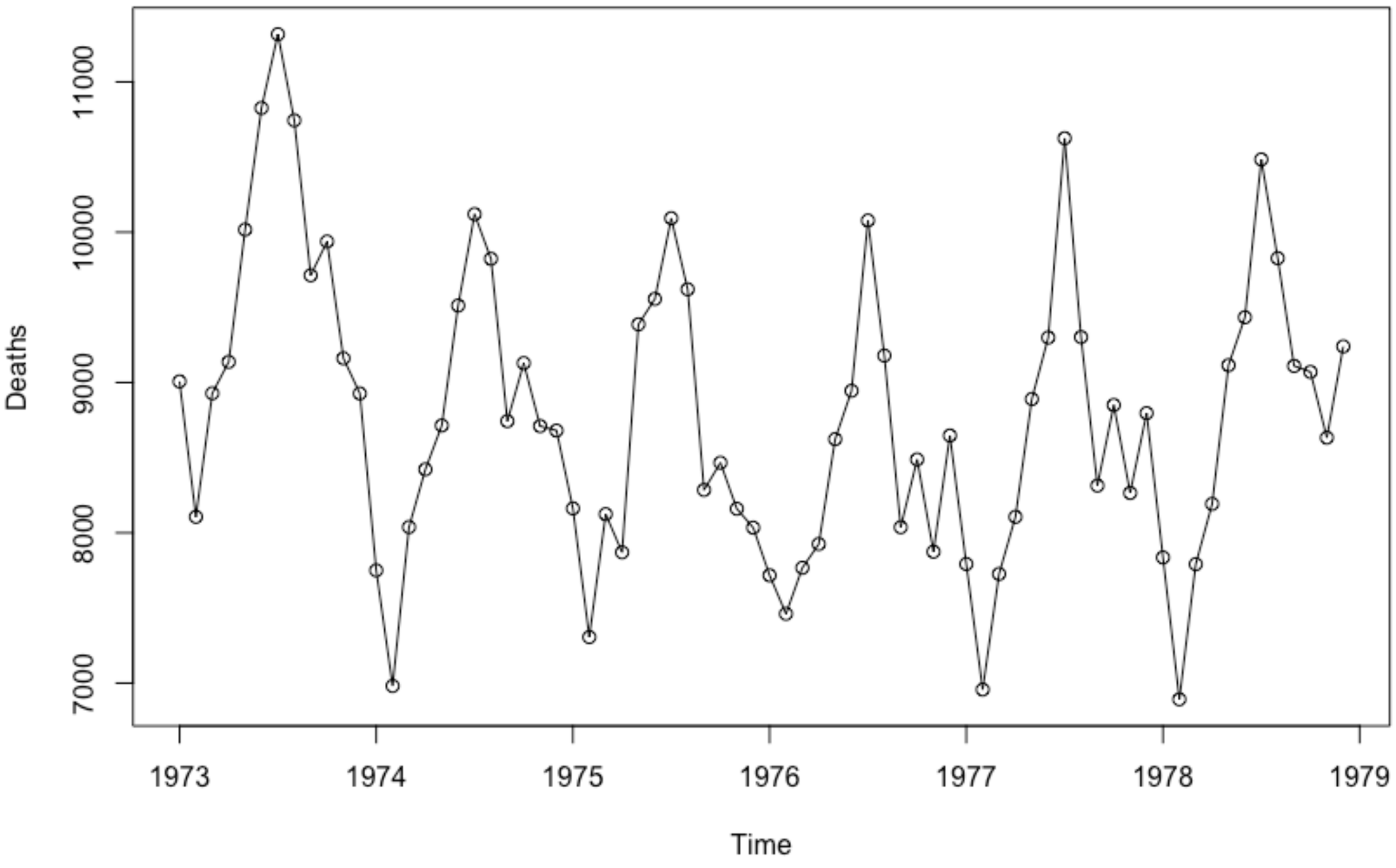
In other words, we performed a (quadratic) regression of Y_t over time t

Regression over time

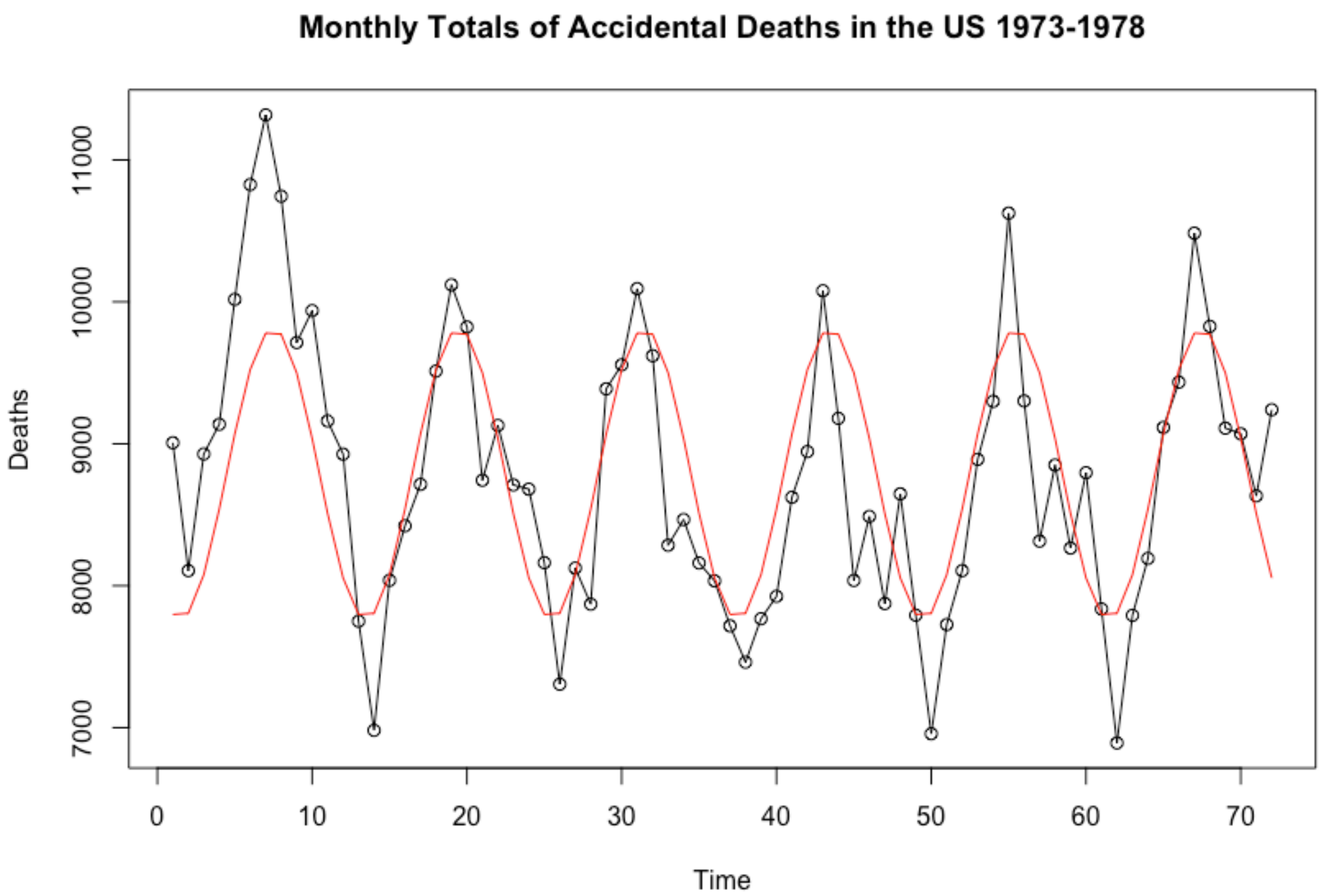
The first time series technique we will study is regression over the time variable t

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

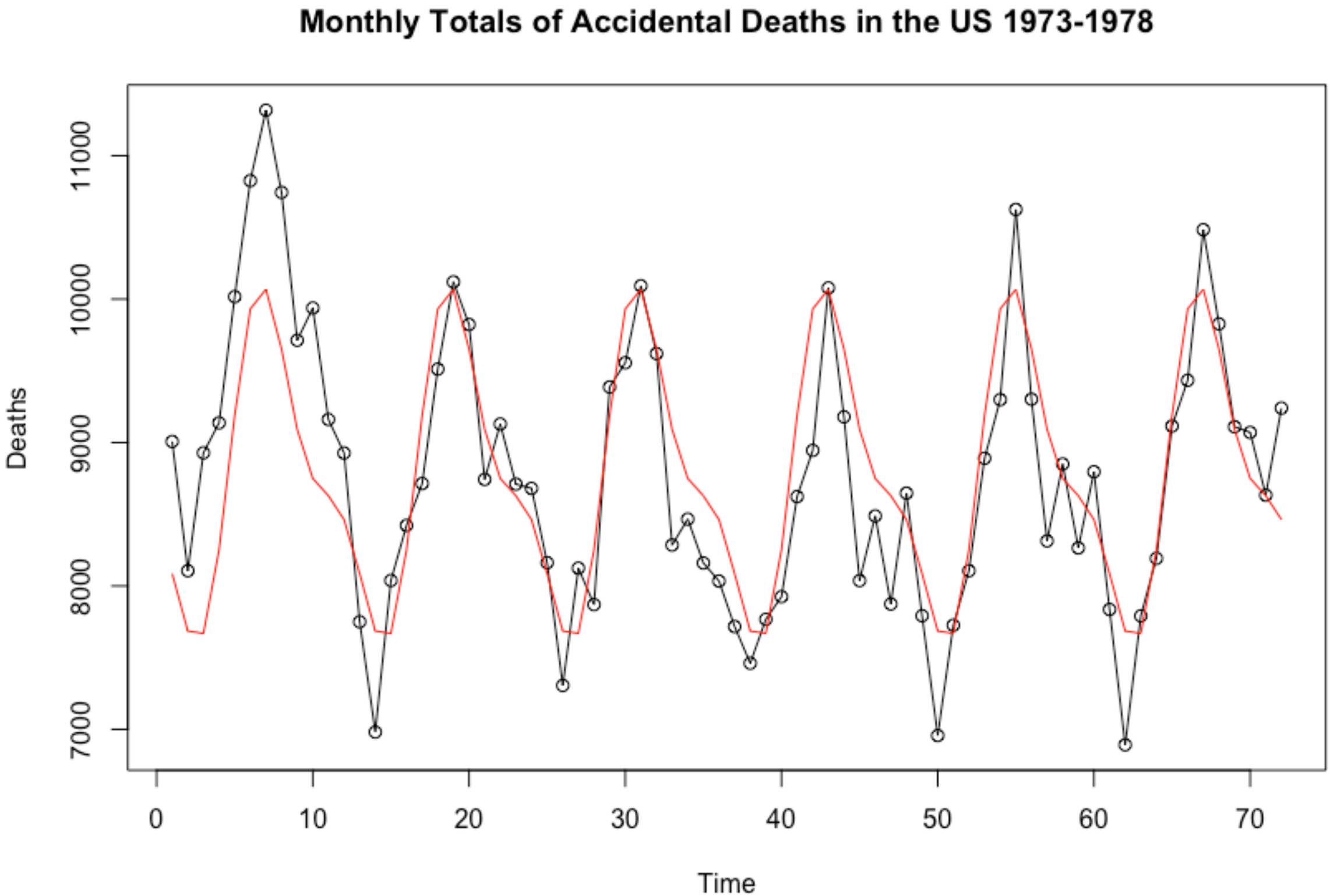
Multiple linear regression of Y_t on t and other functions of t (such as powers, sinusoids etc) will fit more general functions to the data


Monthly Totals of Accidental Deaths in the US 1973-1978
Deaths
7000
8000
9000
10000
11000
1973
1974
1975
1976
1977
1978
1979
Time

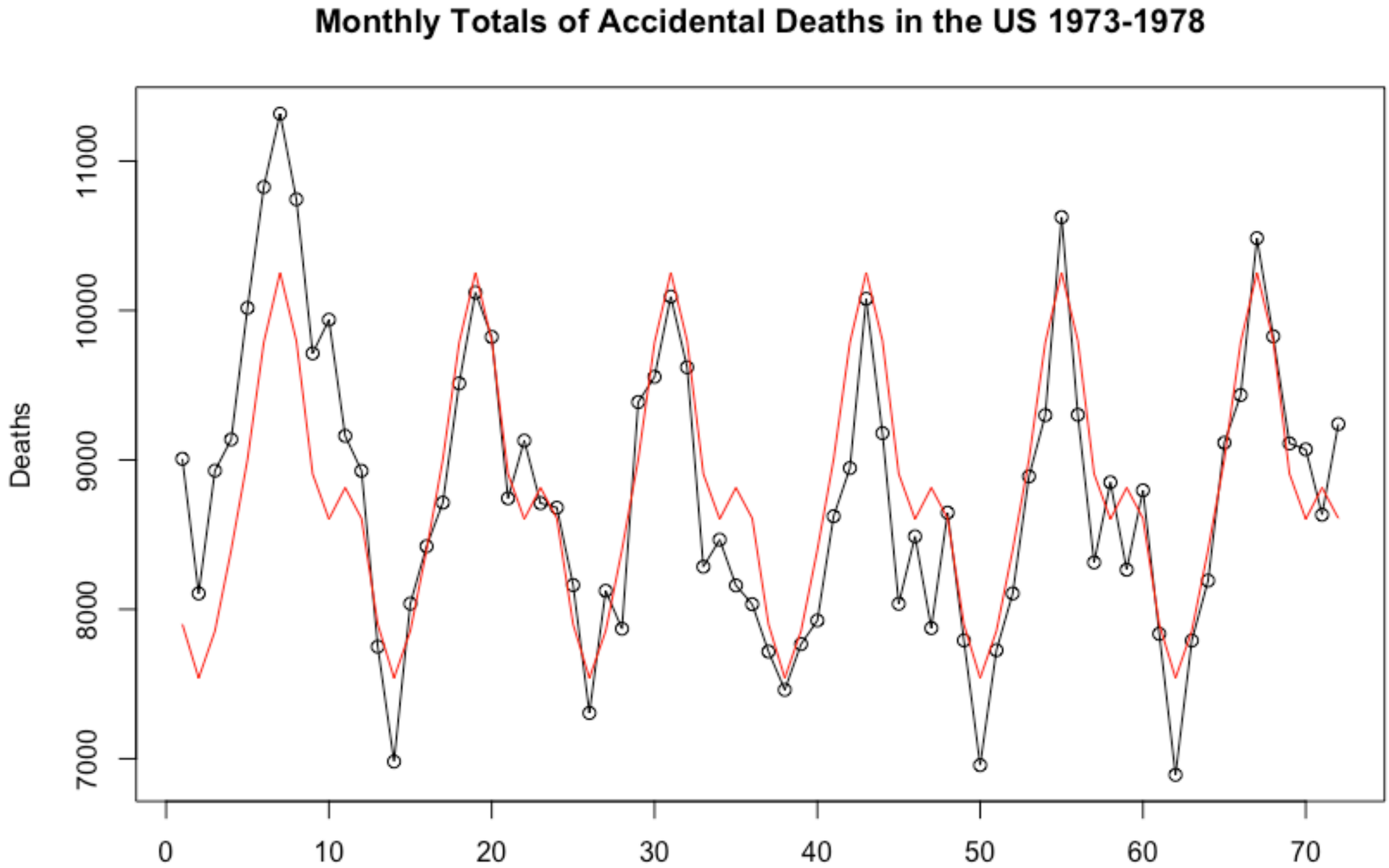
Linear regression of Y_t on $1, \cos(\pi t/6), \sin(\pi t/6)$:



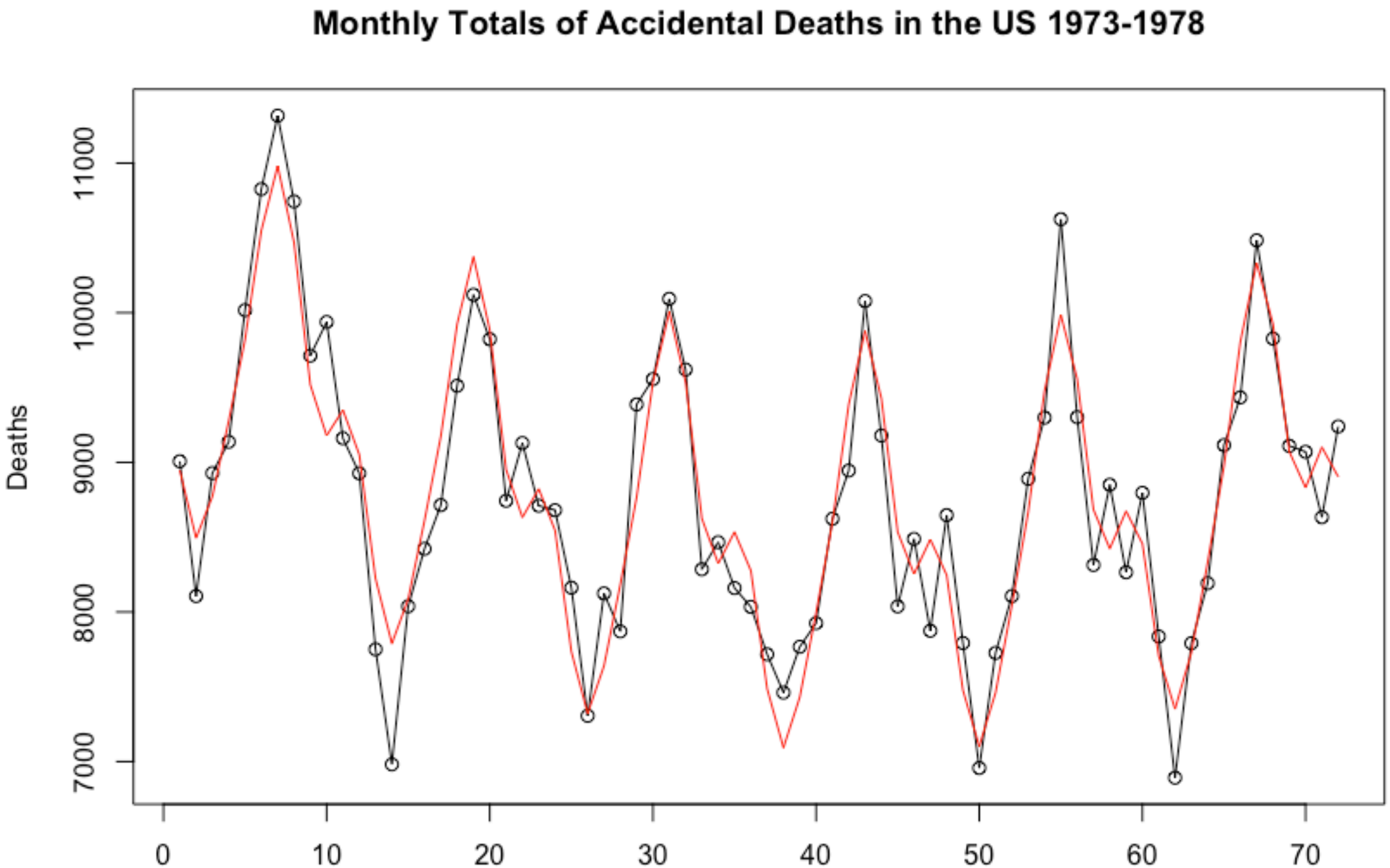
$$Y_t \sim 1, \cos(\pi t/6), \sin(\pi t/6), \cos(\pi t/3), \sin(\pi t/3)$$



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



$$Y_t \sim \text{sinusoids} \; + \; \text{quadratic}$$

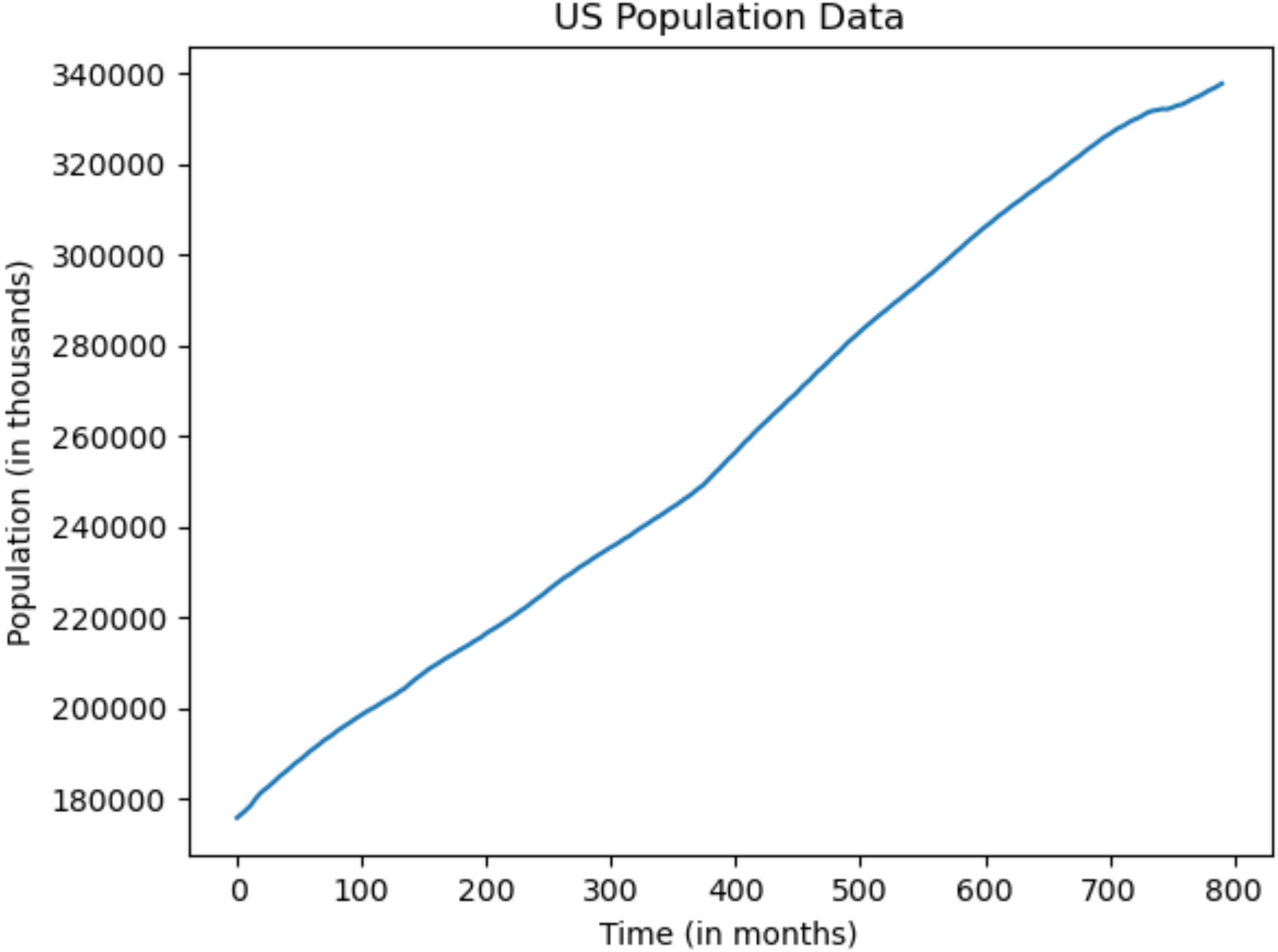


Topic One: Multiple Linear Regression

- 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

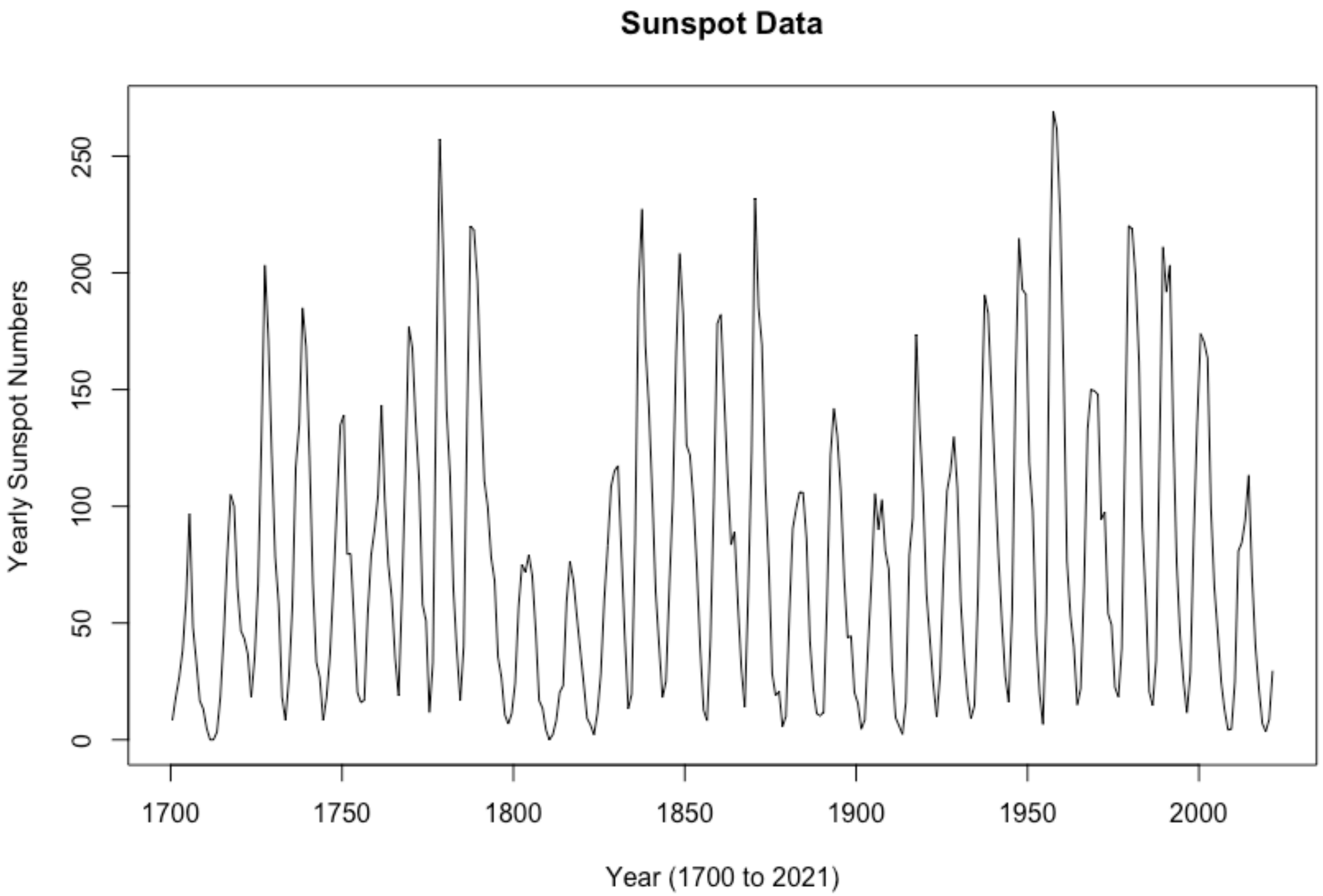
Topic Two: Nonlinear Regression

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


US Population Data
340000
320000
300000
280000
260000
240000
220000
200000
180000
Population (in thousands)
0
100
200
300
400
500
600
700
800
Time (in months)

- 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



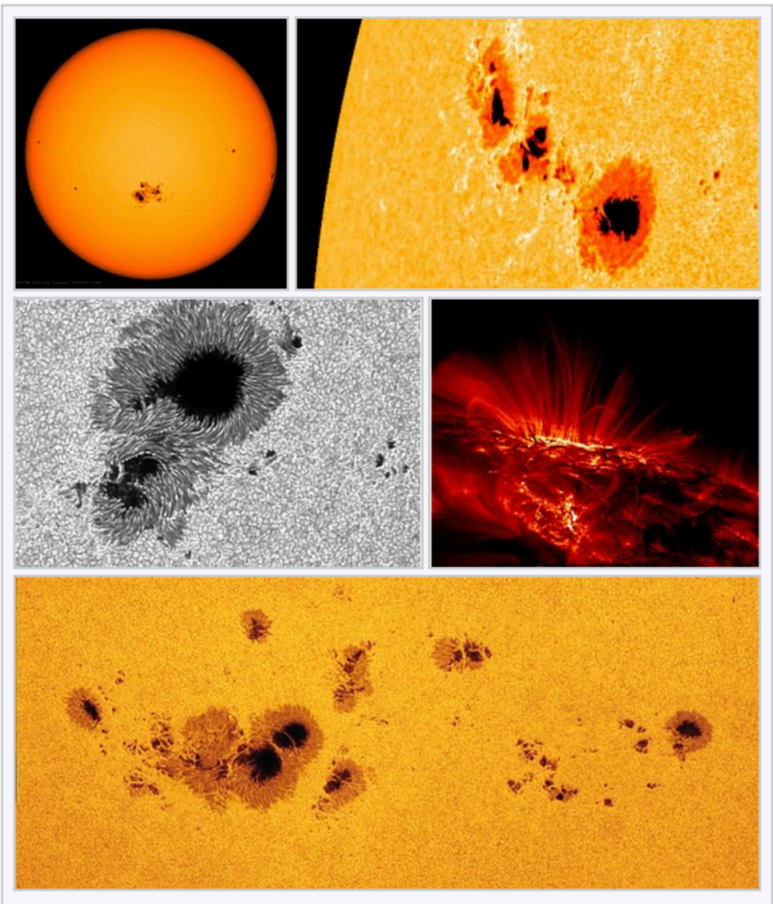
From Wikipedia, the free encyclopedia

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.

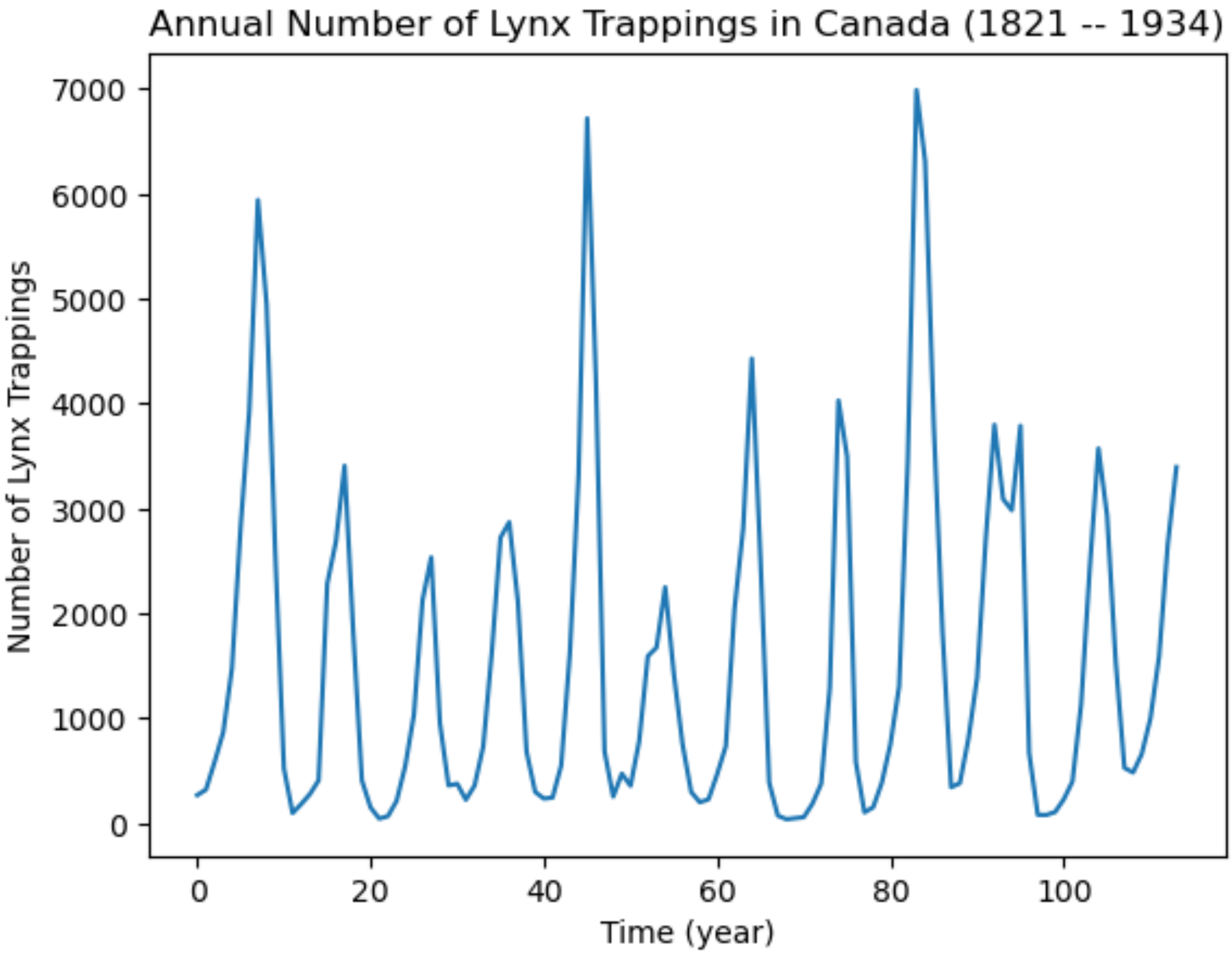
Indicating intense magnetic activity, sunspots accompany other active region phenomena such as coronal loops, prominences, and reconnection events. Most solar flares and coronal mass ejections originate in these magnetically active regions around visible sunspot groupings. Similar phenomena indirectly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



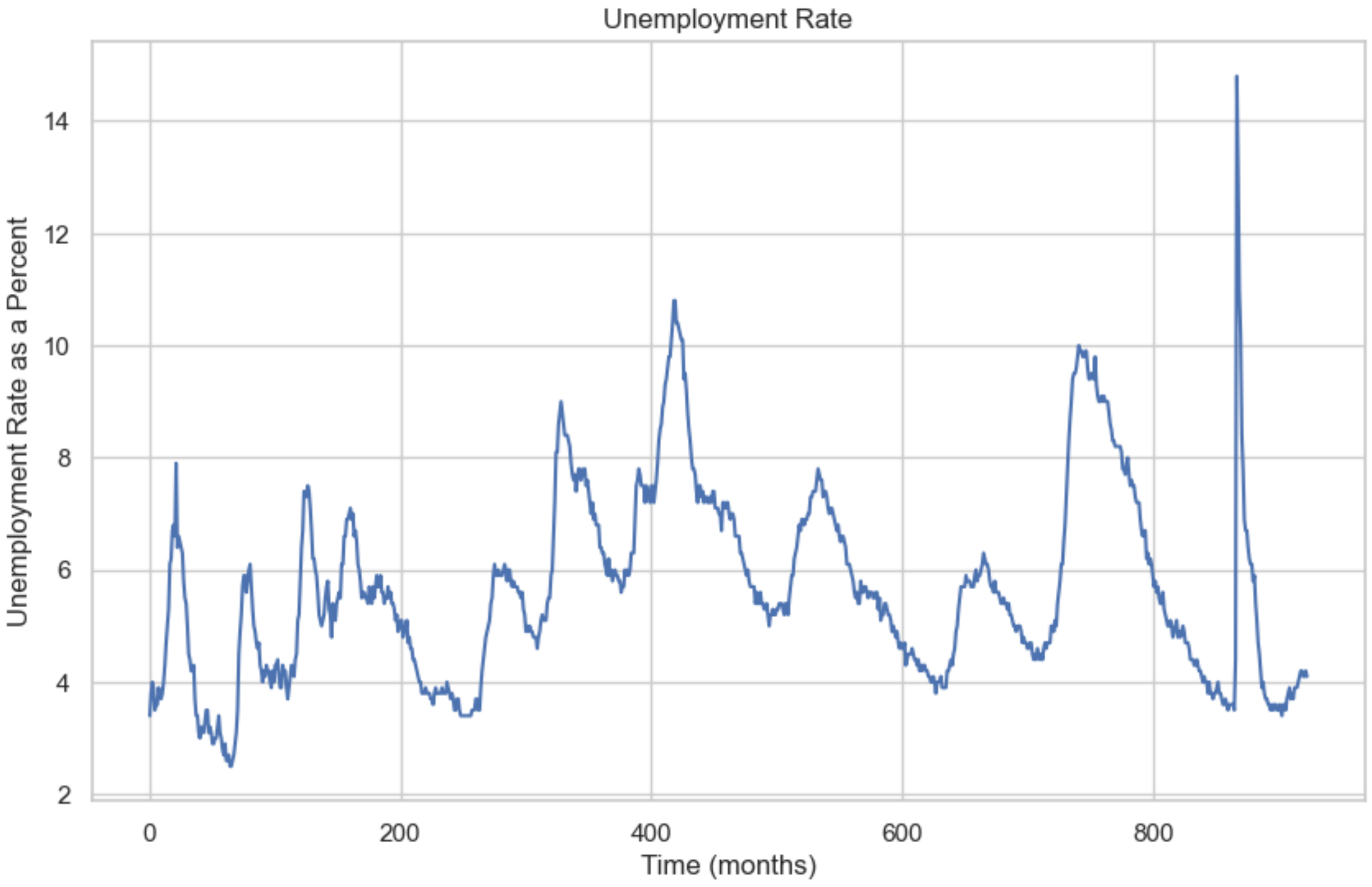
- Top: active region 2192 in 2014 containing the largest sunspot of solar cycle 24[1] and active region 1302 in September 2011.
- Middle: sunspot close-up in the visible spectrum (left) and another sunspot in UV, taken by the TRACE observatory.
- Bottom: a large group of sunspots

- 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



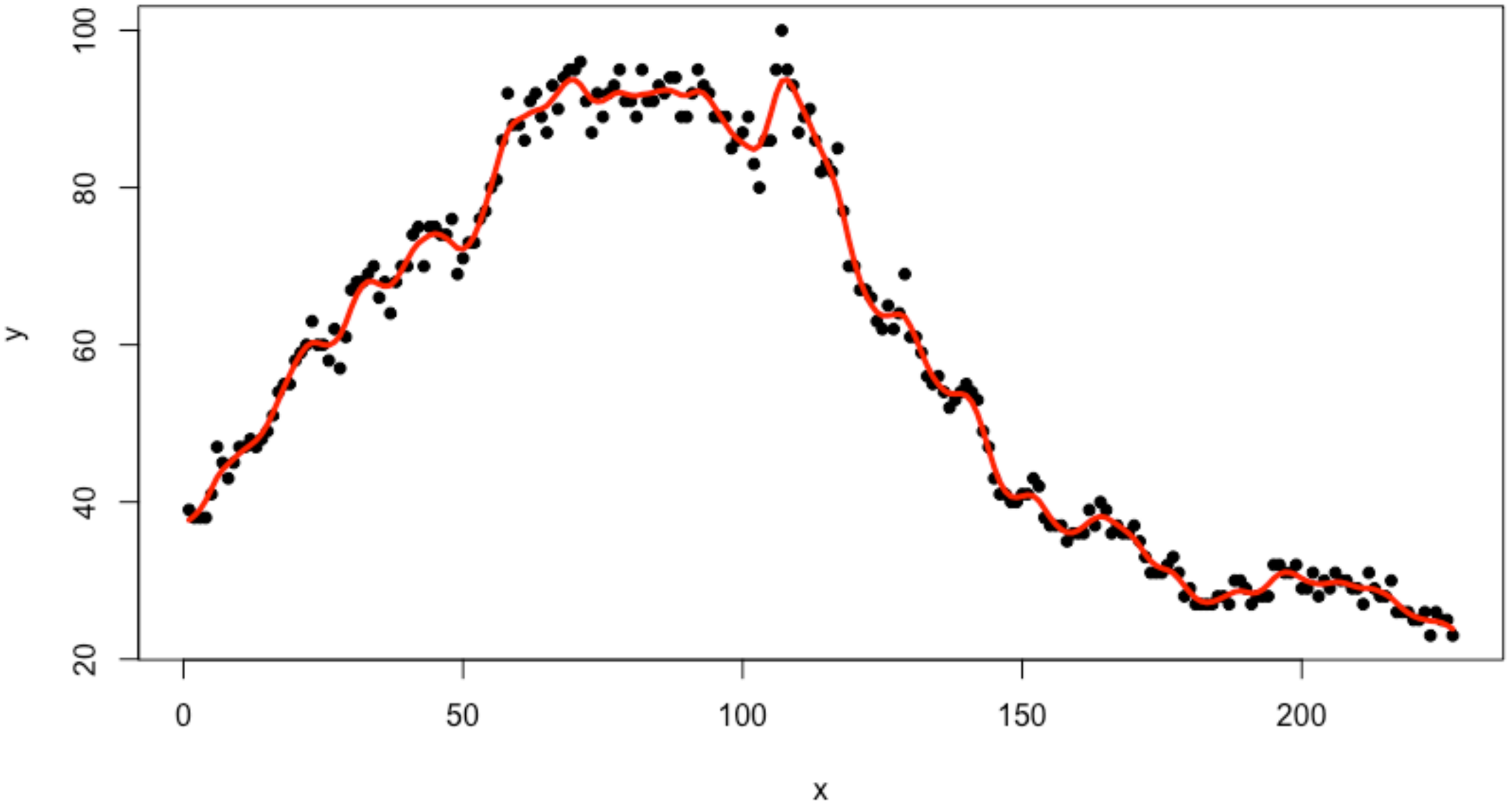
Unemployment Rate from FRED



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:
  $Y_t = \beta_0 + \beta_1 t + \beta_2 (t-2)_+ + \beta_3 (t-3)_+ + \ldots + \beta_n (t-n)_+ + \epsilon$
- 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

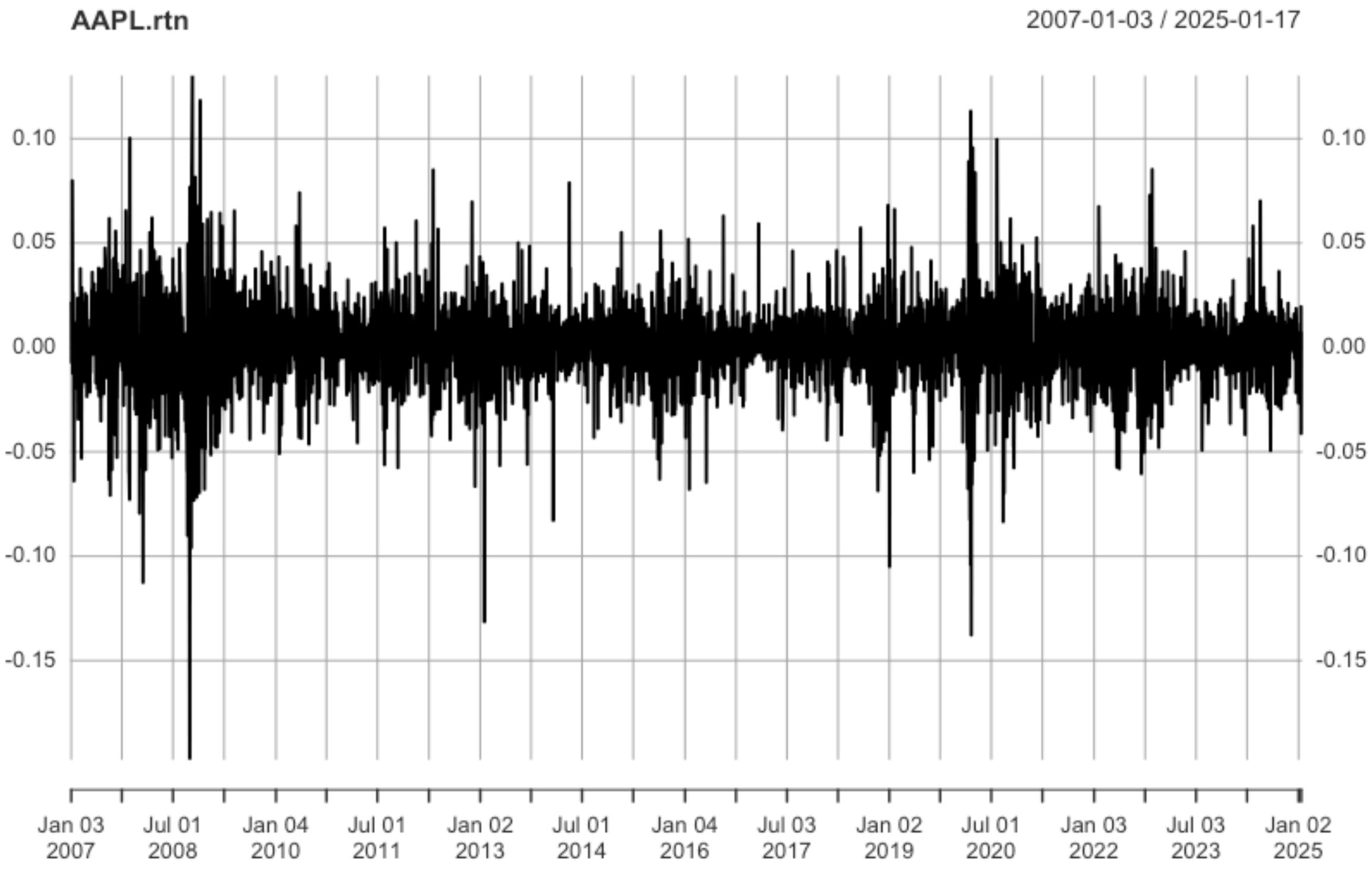
Here is this model (with Ridge regularization) applied to the Google trends data for “yahoo”



Topic Four: Variance Modeling

Daily closing prices of Apple Stock from Yahoo Finance

Stock Returns



- Financial analysts also study the volatility of stock price returns
- For estimating volatility, it is common to use the model: $Y_t \sim N(0,\sigma_t^2)$ 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
- Spectral Analysis converts the observed time series to the Fourier basis and then uses a variance model on the coefficients

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
- We shall study these models in Topic Five

Topic Six: Recurrent Neural Networks

- Recurrent Neural Networks (RNNs) are usually formulated in the framework of regression: $(x_t, y_t), t = 1, \ldots, n$
- This means that at each time point t, we observe a response value y_t as well as a covariate vector x_t
- Usually in regression, one uses models $y_t = f(x_t)$. But RNNs use $y_t = f_t(x_t, x_{t-1}, \ldots, x_1)$
- We shall go over these models (including LSTMs) and some of their applications

- 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

Different kinds of time series data

- Univariate Time Series
- Vector Time Series
- Time Series Regression data
- Sequential Data

Univariate Time Series

- $y_1, \ldots, y_T$ where each y_t is a real number
- This is the simplest time series dataset
- We shall be mostly working with these in this course

Vector Time Series

- $y_1, \ldots, y_T$ where each y_t is vector-valued
- For example, consider $y_t = (y_{t1}, y_{t2})^T$ where y_{t1} is the unemployment rate and y_{t2} is the GDP growth rate for the t^{th} quarter. This is a bivariate time series
- When the dimension of the vectors is large, these are referred to as high-dimensional time series
- We will not spend that much time on vector time series

Time Series Regression

- $(x_1, y_1), \ldots, (x_T, y_T)$ where each x_t is vector-valued and y_t is real-valued
- For each time point t, we observe a covariate vector x_t as well as a response value y_t
- The goal is to predict y_{T+1} given x_{T+1} and the current data set (and then y_{T+2} given x_{T+2} etc.)
- This is a more general setting compared to both regression over time ($x_t = t$) and lagged regression ($x_t = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})$)
- We shall study RNNs in this setting

Sequential Data

- In Machine Learning, sequential data mostly refers to $(x_i, y_i), i = 1,\ldots,n$ where each x_i is a time series and/or each y_i is a time series. There is often no dependence across i

- E.g., consider the problem of determining whether a review is positive or not based on the text of the review. x_i denotes the i^{th} review and y_i is binary

- Each review x_i is a sequence of words which can be viewed as a time series.

- RNNs have heavily been used here