§2.4 Least squares estimation & probability

Carl Friedrich Gauss
1777–1855

Least squares estimation

Given observed data $y_1, \dots, y_n$, fit the linear model

$$y \approx \beta_1 e_1 + \cdots + \beta_K e_K$$

i.e.

$$y_i \approx \beta_1 e_{1,i} + \cdots + \beta_K e_{K,i}$$

by choosing the parameters $\beta_1, \dots, \beta_K$ so as to minimize the mean square error

$$\text{mse} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \text{pred}_i)^2$$

where $\text{pred}_i = \beta_1 e_{1,i} + \cdots + \beta_K e_{K,i}$

Maximum likelihood estimation

Given observed data $y_1, \dots, y_n$, fit the probability model

$$Y_i \sim \cdots$$

by choosing the model parameters so as to maximize the log likelihood of the observed data

$$\log \Pr(y_1, \dots, y_n) = \sum_{i=1}^{n} \log \Pr_Y(y_i ; \dots)$$

Example 2.1.1

The Iris dataset has 50 records of iris measurements, from three species.

How does `Petal.Length (PL)` depend on `Sepal.Length (SL)`?

We fitted the linear model

$$\text{PL} \approx \alpha + \beta\ \text{SL} + \gamma\ \text{SL}^2$$

Example

Let's fit the probability model

$$\text{PL}_i \sim \alpha + \beta\ \text{SL}_i + \gamma\ \text{SL}_i^2 + \text{Normal}(0, \sigma^2)$$

Model for a single observation: $\mathrm{PL}_i \sim \alpha + \beta\ \mathrm{SL}_i + \gamma\ \mathrm{SL}_i^2 + N(0, \sigma^2)$

$y_i \sim \alpha + \beta e_i + \gamma f_i + N(0, \sigma^2)$

$\sim N(\alpha + \beta e_i + \gamma f_i, \sigma^2)$

Likelihood of a single observation:

$$\Pr_{y_i}(y_i ; \alpha, \beta, \gamma, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - (\alpha + \beta e_i + \gamma f_i))^2 / 2\sigma^2}$$

Log likelihood of the dataset:

$$\log \Pr(y_1 \cdots y_n ; \alpha, \beta, \gamma, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - (\alpha + \beta e_i + \gamma f_i))^2$$

We then seek $\alpha, \beta, \gamma, \sigma$ to maximize this.

Maximize over the unknown parameters,
α, β, γ, and σ:

$$\max_{\alpha,\beta,\gamma,\sigma} \left\{ -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha+\beta e_i+\gamma f_i))^2 \right\}$$

$$= \max_{\sigma}\left[\max_{\alpha,\beta,\gamma} \left\{ -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}(y_i - (\alpha+\beta e_i+\gamma f_i))^2 \right\} \right]$$

$$= \max_{\sigma}\left[-\frac{n}{2}\log(2\pi\sigma^2) + \max_{\alpha,\beta,\gamma} \left\{ -\frac{1}{2\sigma^2}\sum_{i}(y_i - (\alpha+\beta e_i+\gamma f_i))^2 \right\} \right]$$

$$= \max_{\sigma}\left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\{ \min_{\alpha,\beta,\gamma} \sum_{i}(y_i - (\alpha+\beta e_i+\gamma f_i))^2 \right\} \right]$$

$$= \max_{\sigma}\left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}(y_i - \hat{y}_i)^2 \right] \quad \text{where } \hat{y}_i = \underbrace{\hat{\alpha}+\hat{\beta}e_i+\hat{\gamma}f_i}$$

obtained by least squares estimation

$$\Rightarrow \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i}(y_i - \hat{y}_i)}$$

Maximize over the unknown parameters, α, β, γ, and σ:

Least squares estimation *derives* from a Gaussian probability model.

If that model doesn't fit the data, then don't use least squares estimation!

Sign in

The Guardian

News website of the year

Stop and search

This article is more than 3 years old

Met police 'disproportionately' use stop and search powers on black people

London's minority black population targeted more than white population in 2018 - official figures

Let $y_i \in \{0,1\}$ be the outcome for stop-and-search incident i.

$y_i \approx \alpha + \beta_{\text{eth}_i}$ i.e. $Y_i \sim \alpha + \beta_{\text{eth}_i} + N(0, \sigma^2)$

Fit α and $\beta_{\text{Bl}}, \beta_{\text{Mi}}$, ... using least squares estimation
or, equivalently, fit using maximum likelihood estimation

$Y_i \sim \text{Bin}(1, \alpha + \beta_{\text{eth}_i})$
Fit the parameters using maximum likelihood estimation

There's a more advanced version called *Logistic Regression*, for $\text{Bin}(1, \theta_i)$ where θ_i depends on multiple features. It uses softmax. See the code in `[stop-and-search.ipynb]`, or Part II Advanced Data Science.

climate

cl.cam.ac.uk/teaching/2223/DataSci/datasci/ex/climate.html

In [1]:
```
import numpy as np
import pandas
import matplotlib.pyplot as plt
import sklearn.linear_model
π = np.pi
```

Climate dataset challenge

- What is the rate of temperature increase in Cambridge?
- Are temperatures increasing at a constant rate, or has the increase accelerated?
- How do results compare across the whole of the UK?

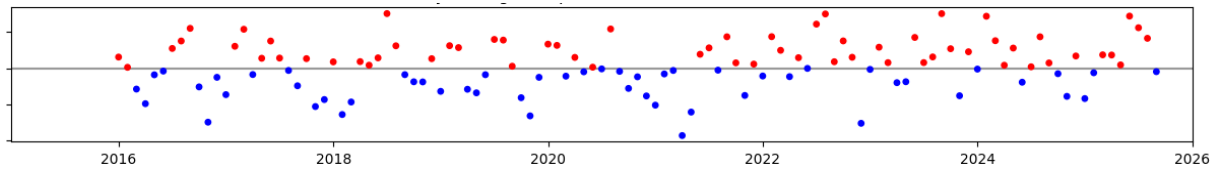
Your task is to answer these questions using appropriate linear models, and to produce elegant plots to communicate your findings. Please submit a Jupyter notebook, or a pdf. Include explanations of what your models are, and of what your plots show.

The dataset is from https://www.metoffice.gov.uk/pub/data/weather/uk/climate/. Code for retrieving the dataset is given at the bottom.

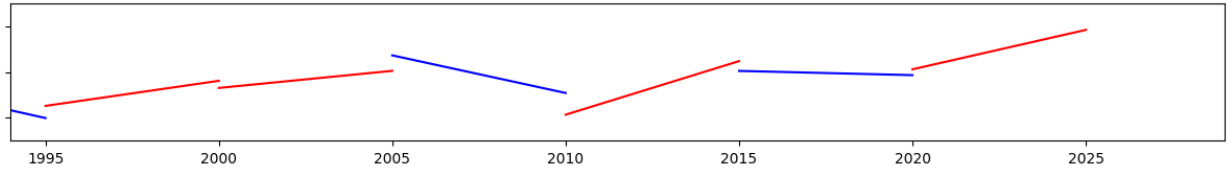
Arthur Sun (Trinity)

```
dt = y_val - model.predict(X_val)
binomtest((dt > 0).sum(), len(dt), alternative='greater')
```

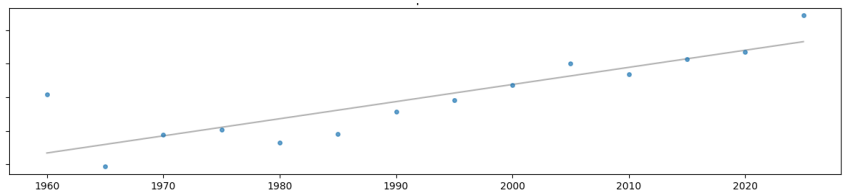
Oliver Gibson (Corpus)



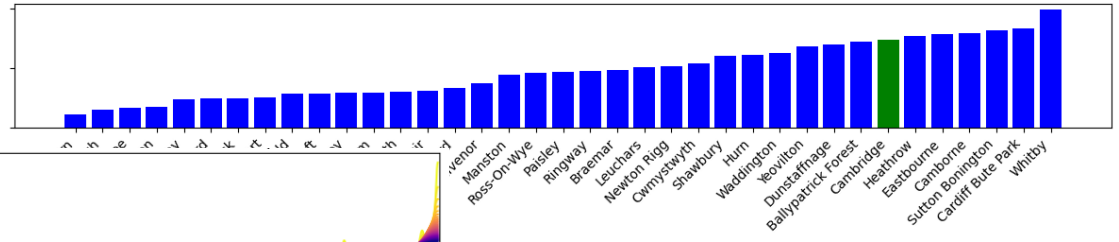
Penny Li (St John's)



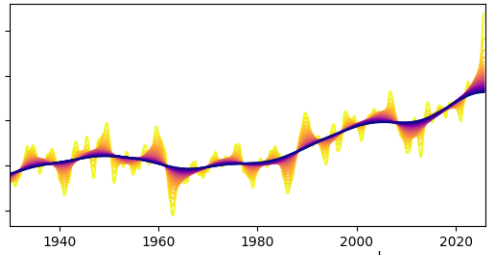
Tosin Bea Jokosenumi (Churchill)



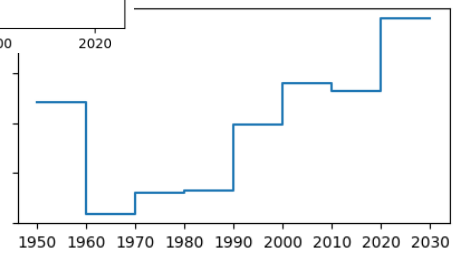
Ben Tinsley (Fitzwilliam)



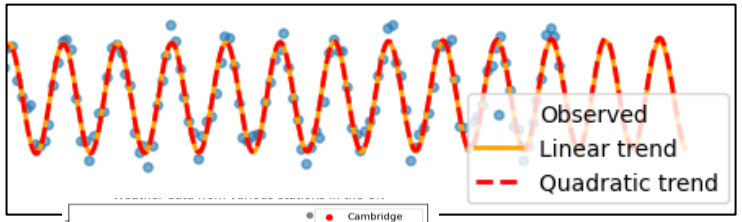
Adam Whittome (Caius)

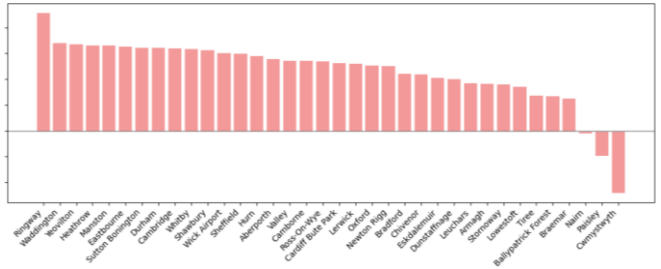


Arya Golkari (Trinity)

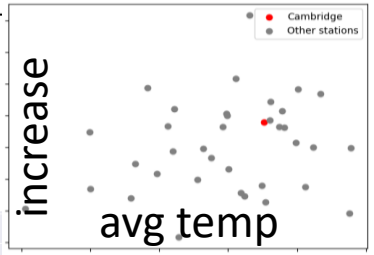


Ed Read (Churchill)

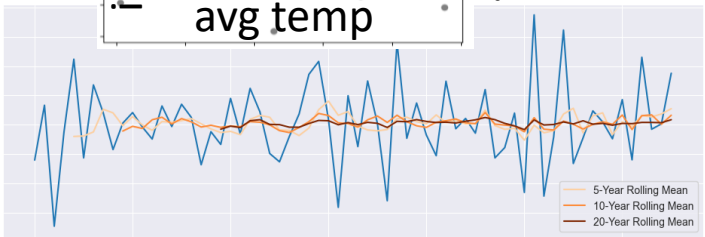




Justin Leung (Churchill)



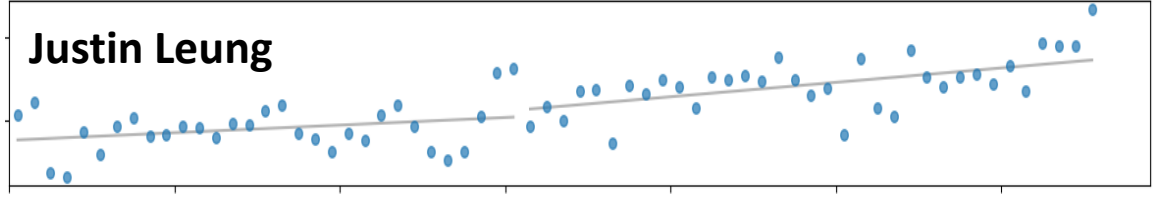
Zerui Chen (Wolfson)



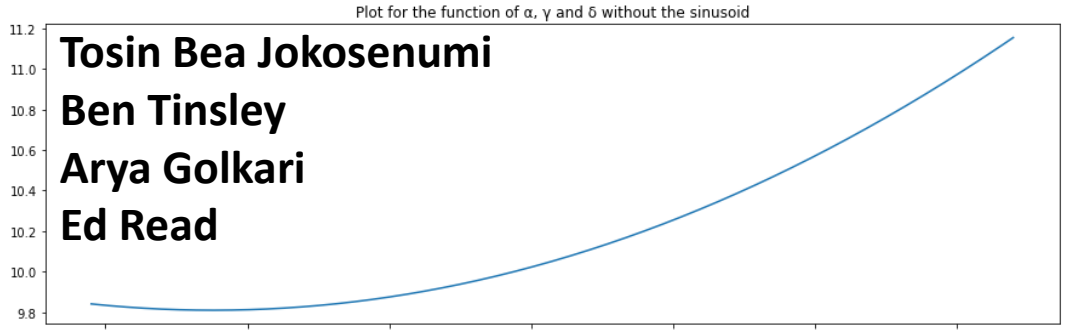
Isaac Chan (Churchill)

$$\gamma_{1990} = -1.019, \quad \gamma_{20000} = -0.649, \quad \gamma_{2010} = -0.737, \quad \gamma_{2020} = 0$$

This model assumes a simple functional form, which makes it easy to summarize the nonlinearity – but which may be wrong.

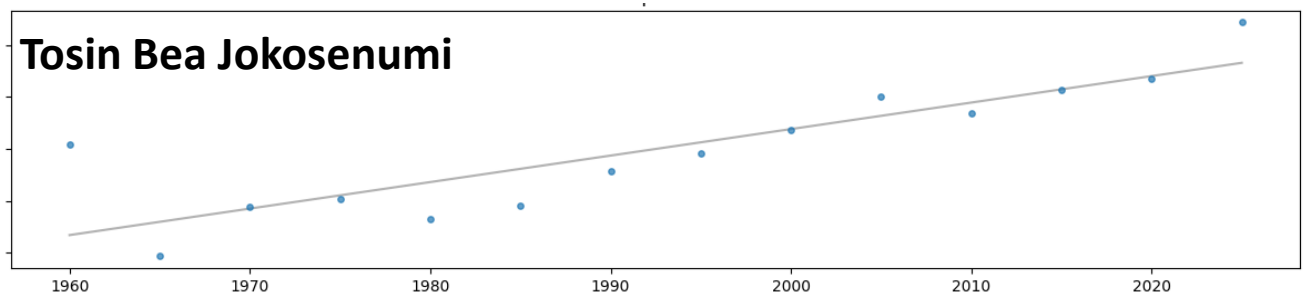


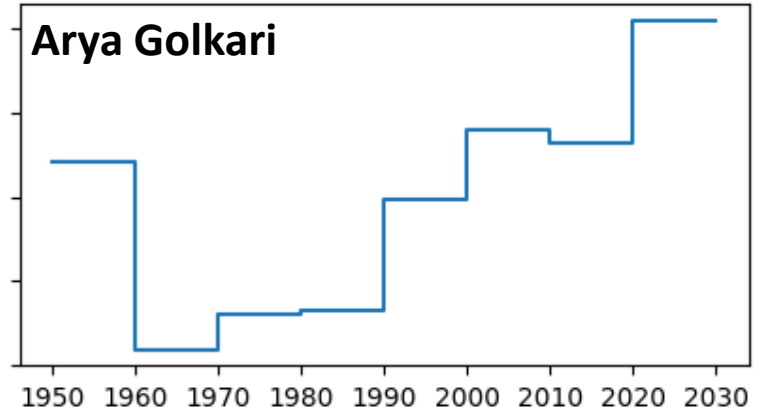
$$\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t + \delta t^2$$

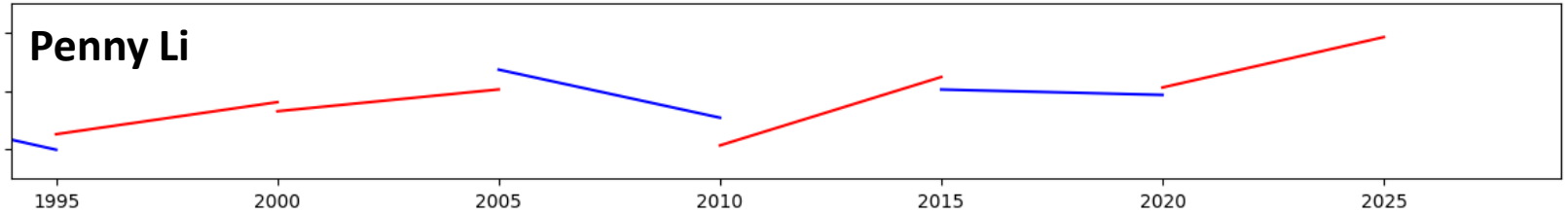


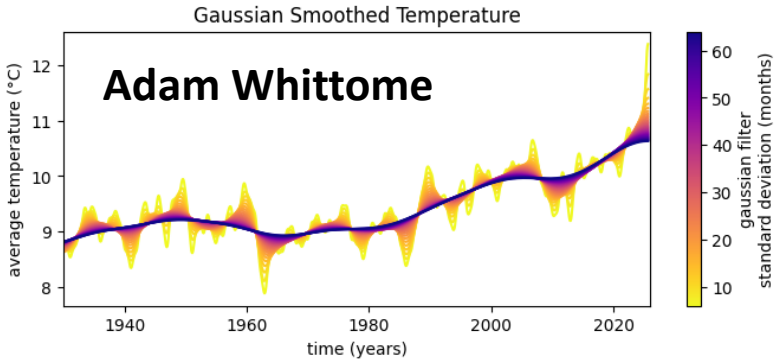
These models don't presume the shape of the response; they're open to anything. But we still need to find a way to summarize all the coefficients.

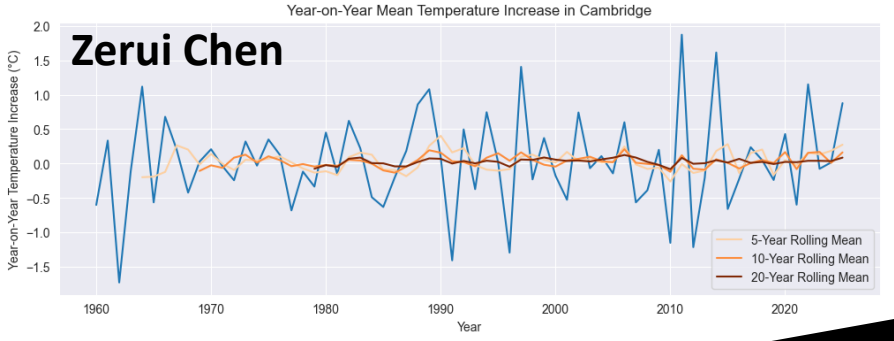
$$\text{temp} \approx \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma_{\text{decade}}$$





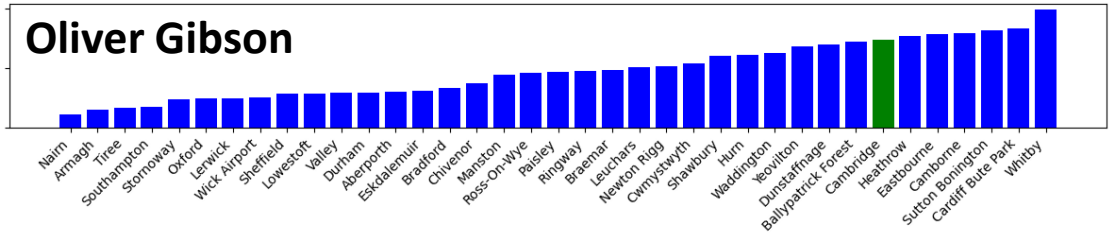






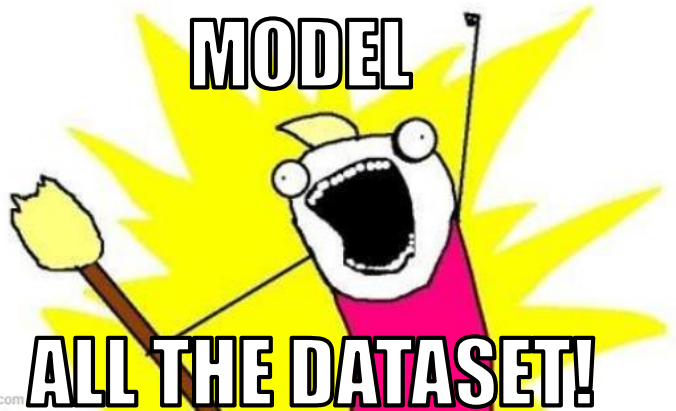
"Don't trust models, just invent algorithms to process the plain honest data"

"I lowkey kind of trust models ..."



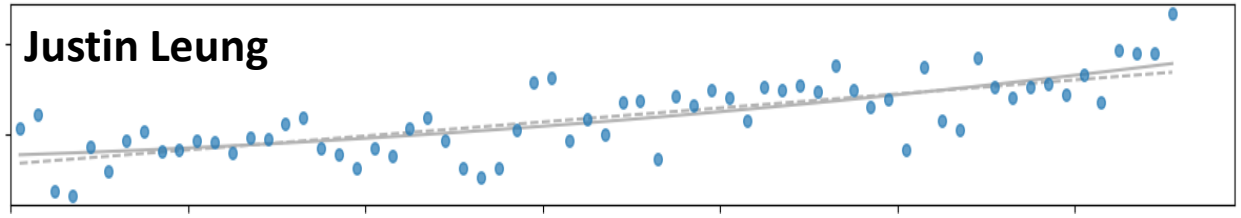
for s in stations:

model data from station s as $\text{Temp} \sim \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t + N(0, \sigma^2)$



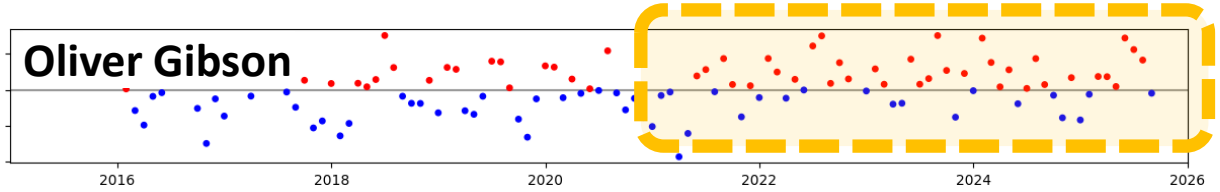
$$\text{Temp} \approx \alpha_{\text{station}} + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma_{\text{station}} t$$

$$\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t + \delta t^2$$



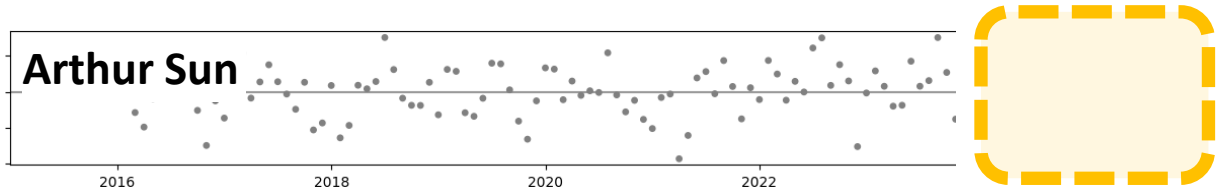
"I found the mle to be $\hat{\delta}$=0.000757, which is positive, indicating that temperature increase is accelerating"

$$\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t$$



"I fitted the model, then examined the signs of the later residuals, and did a hypothesis test to see if they're consistent with Bin(1,½)."

$$\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t$$



"I did the hypothesis test on a holdout set, to see if the signs of the residuals are consistent with Bin(1,½)."

model choice ≈ parameter estimation

generalization

hypothesis testing

holdout set evaluation