

This week we're building up our skills at inventing useful probability models.

MONDAY

- ❖ linear models & feature design
(to quickly turn our ideas into easy-to-fit models)

WEDNESDAY

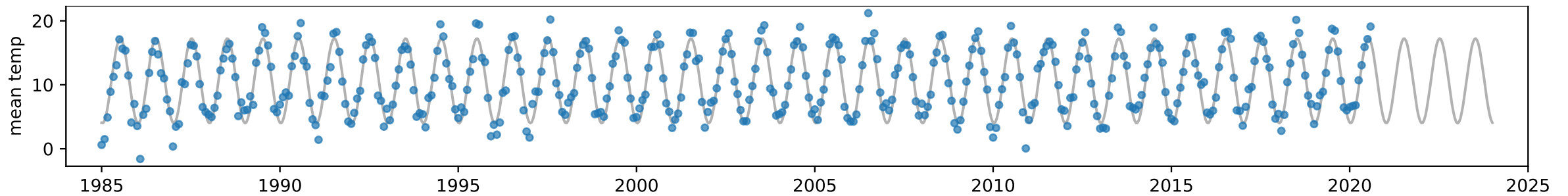
- ❖ “debugging” models

FRIDAY

- ❖ identifiability of parameters
- ❖ the link between least squares and likelihood

With our periodic model ...

$$\text{temp} \approx \alpha + \beta \sin(2\pi(\text{t} + \phi))$$



... how do we *discover* we should add a secular term, such as $+\gamma\text{t}$?

Machine learning
models don't fail with
nice friendly
exceptions. They fail
by giving us fishy
answers.

What does
“debugging” look like?

§2.3. Diagnosing a linear model

After fitting a model

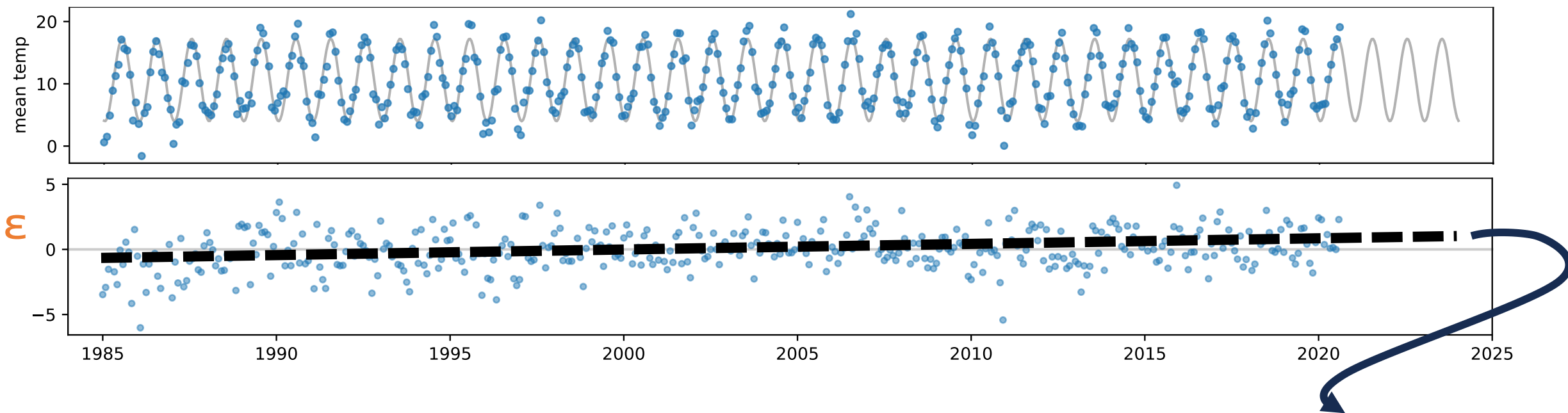
```
model.fit(...,  $y$ )
```

1. Compute the prediction errors
a.k.a. the residuals

```
 $\epsilon$  =  $y$  - model.predict()
```

2. Plot ϵ against anything we can think of. It tells us where our model is poor.

$$\text{temp} \approx \alpha + \beta \sin(2\pi(\mathbf{t} + \phi)) \quad \text{i.e. } \text{temp} = \alpha + \beta \sin(2\pi(\mathbf{t} + \phi)) + \varepsilon$$



$$\varepsilon \approx \delta + \gamma t$$

This suggests a revised model ...

$$\text{temp} \approx \underbrace{(\alpha + \delta)}_{\alpha'} + \beta \sin(2\pi(\mathbf{t} + \phi)) + \gamma \mathbf{t}$$

Example sheet 1

Question 8. For the climate data from section 2.2.5 of lecture notes, we proposed the model

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

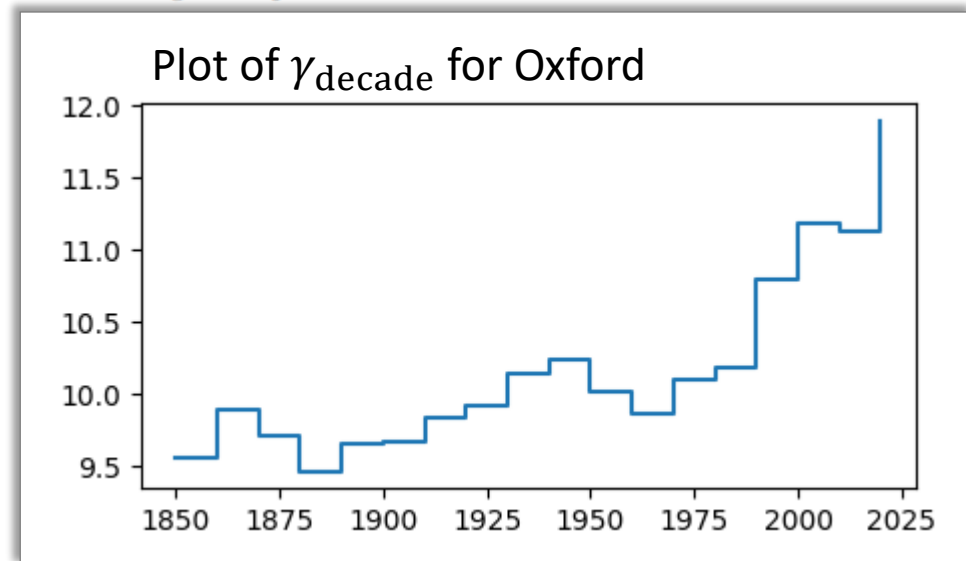
in which the $+\gamma \mathbf{t}$ term asserts that temperatures are increasing at a constant rate. We might suspect though that temperatures are increasing non-linearly. To test this, we can create a non-numerical feature out of $\mathbf{t}$ by

$$\mathbf{u} = \text{'decade_'} + \text{str}(\text{math.floor}(\mathbf{t}/10)) + \text{'0s'}$$

(which gives us values like 'decade_1980s', 'decade_1990s', etc.) and fit the model

$$\text{temp} \approx \alpha + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t}) + \gamma_{\mathbf{u}}.$$

Write this as a linear model, and give code to fit it. *[Note. You should explain what your feature vectors are, then give a one-line command to estimate the parameters.]*



QUESTION. This code doesn't pass the Moodle tester. What's the bug?

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

```
class StepPeriodicModel():
    def __init__(self):
        self.mindec = np.nan
        self.maxdec = np.nan

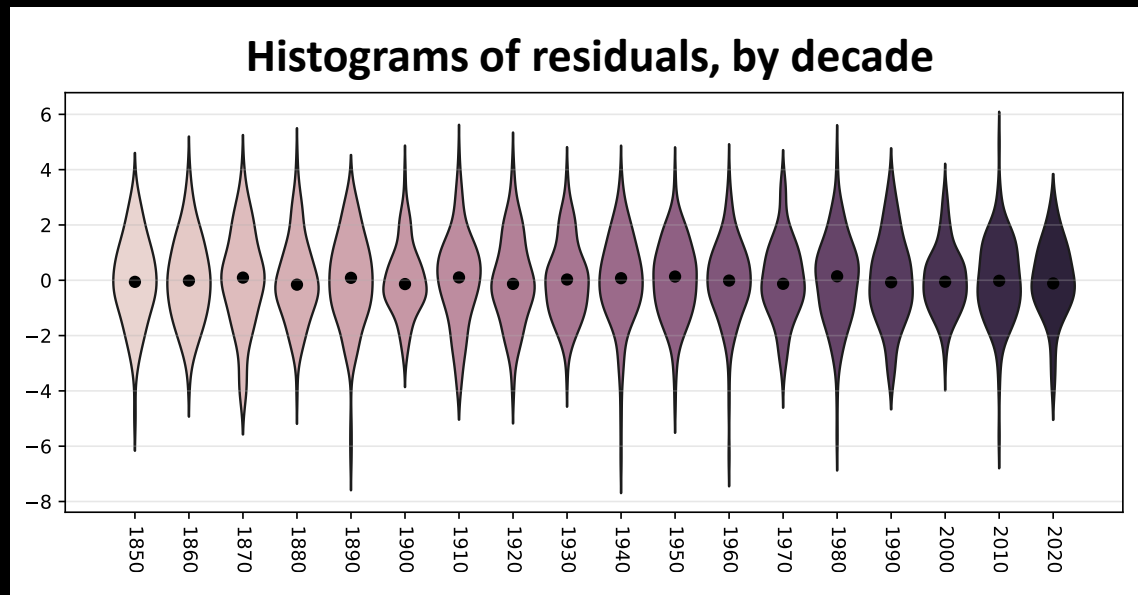
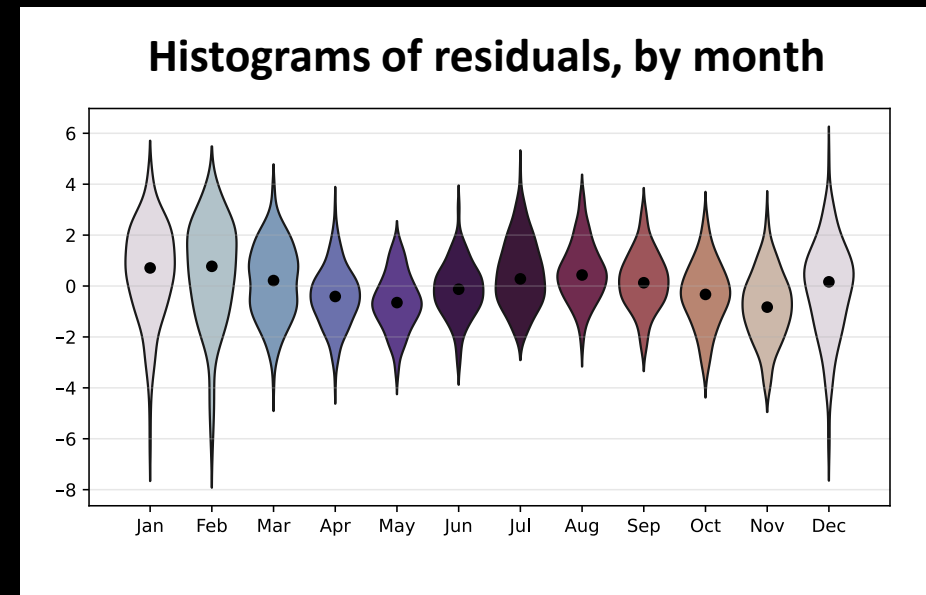
    def fit(self, t, temp):
        self.mindec = np.floor(min(t)/10)*10
        self.maxdec = np.floor(max(t) / 10) * 10
        indicators = [np.where(np.floor(t/10)*10 == year*10 + self.mindec, 1, 0)
                       for year in range(int((self.maxdec - self.mindec)/10) + 1)]
        X = np.column_stack([np.sin(2 * np.pi * np.mod(t,1)), np.cos(2 * np.mod(t,1)), *indicators])
        model = sklearn.linear_model.LinearRegression(fit_intercept=False)
        model.fit(X, temp)
        (_,_,*γ) = model.coef_
        self.γ = np.append(γ, np.nan)

    def predict_step(self, t):
        t = np.array(t).astype(float)
        ℓ = ((np.floor(t/10)*10-self.mindec)/10).astype(int)
        replace_mask = np.where((ℓ<0) | (ℓ>=len(self.γ)-1))
        ℓ[replace_mask] = len(self.γ) - 1
        return np.take(self.γ, ℓ)
```

cos(2πt)

I computed the residuals.

I plotted them against the variables in the question, looking for any systematic pattern.



This tells me there's something dodgy with how the model captures month-by-month variation, i.e. with the features for the sinusoid shape.

To diagnose fishy coefficients:

When we propose a model, write down the predicted response at some “useful” datapoints.

This should help us interpret what the parameters mean.

To find places where my model’s fit is fishy:

After fitting a model

```
model.fit(...,  $y$ )
```

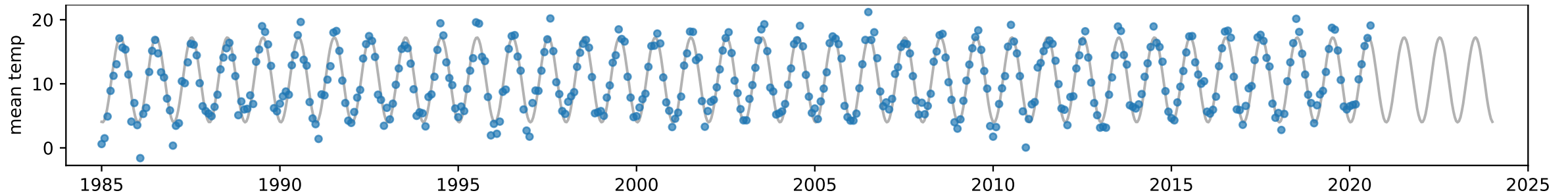
1. Compute the prediction errors a.k.a. the residuals

```
 $\epsilon$  =  $y$  - model.predict()
```

2. Plot ϵ against anything we can think of.

When we fit the model with a constant-rate climate change term ...

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



the fitted parameters are surprising:

α	β_1	β_2	γ
-67.5	-1.1	-6.6	0.039

Why do these three models lead to such different estimates for α ?

Model 0:	$\text{temp} \approx \alpha \mathbf{1} + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t})$	$\Rightarrow \hat{\alpha} = 10.6^\circ\text{C}$
Model A:	$\text{temp} \approx \alpha \mathbf{1} + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t}) + \gamma \mathbf{t}$	$\Rightarrow \hat{\alpha} = -67.5^\circ\text{C}$
Model B:	$\text{temp} \approx \alpha \mathbf{1} + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t}) + \gamma (\mathbf{t} - 2000)$	$\Rightarrow \hat{\alpha} = 10.5^\circ\text{C}$

EXERCISE. Interpret α , by writing out the predicted response temp at several different timepoints t .

Model 0: at $t=0$, pred. temp $= \alpha + \beta_2 \rightarrow \alpha$ is avg. temp over a period $\checkmark$

Model A: at $t=0$, $\alpha + \beta_2$
at $t=2000$, $\alpha + \beta_2 + 2000\gamma$ } $\rightarrow \alpha$ is avg. temp. in ~~AD~~ 1BC

Model B: at $t=0$, $\alpha + \beta_2 - 2000\gamma$
at $t=2000$, $\alpha + \beta_2$ } $\rightarrow \alpha$ is avg. temp. in 2000 AD.

Model A's α asks: what was the average temperature in 1BC?

Model B's α asks: what was the average temperature in 2000AD?

If the coefficients look fishy, check you're reading your model right.

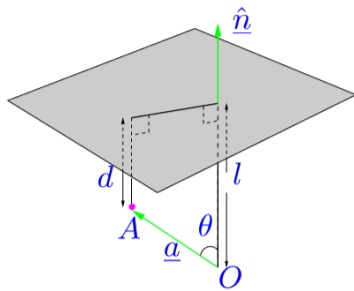
§2.5 The geometry of linear models

To really understand what's going on in a linear model, we need to understand the linear algebra behind them.

NST Maths B, Michaelmas

Example: Distance of point from plane

- What is distance of point A with position vector $\underline{a}$ from plane $\underline{r} \cdot \hat{n} = l$?



- Line containing A and point of closest approach of plane to A must be $\parallel \hat{n}$; has equation

$$\underline{r} = \underline{a} + \lambda \hat{n}$$

- Line meets plane where $\underline{r} \cdot \hat{n} = l$, i.e. where

$$l = \underline{a} \cdot \hat{n} + \lambda$$

- λ is distance along line from $\underline{a}$ so required distance is $|\underline{a} \cdot \hat{n} - l|$

NST Maths B, Easter

Definition. V is called a **vector space over K** , and the elements of V are called **vectors**, if the following **axioms** hold:

- A1** For any vectors $u, v, w \in V$, $(u + v) + w = u + (v + w)$. (**Associativity.**)
- A2** For any vectors $u, v \in V$, $u + v = v + u$. (**Commutativity.**)
- A3** There is a vector in V denoted 0 , called the **zero vector** for which $u + 0 = u$ $\forall u \in V$.
- A4** For each vector $u \in V$ there is a vector in V denoted $-u$ for which $u + (-u) = 0$. (**Inverse.**)
- A5** For any $a \in K$ and any $u, v \in V$, $a(u + v) = au + av$.
- A6** For any $a, b \in K$ and any $u \in V$, $(a + b)u = au + bu$.
- A7** For any $a, b \in K$ and any $u \in V$, $(ab)u = a(bu)$.
- A8** For the unit scalar $1 \in K$ and any $u \in V$, $1u = u$.

The *subspace spanned* by a collection of vectors $\{e_1, \dots, e_K\}$ is the set of all linear combinations

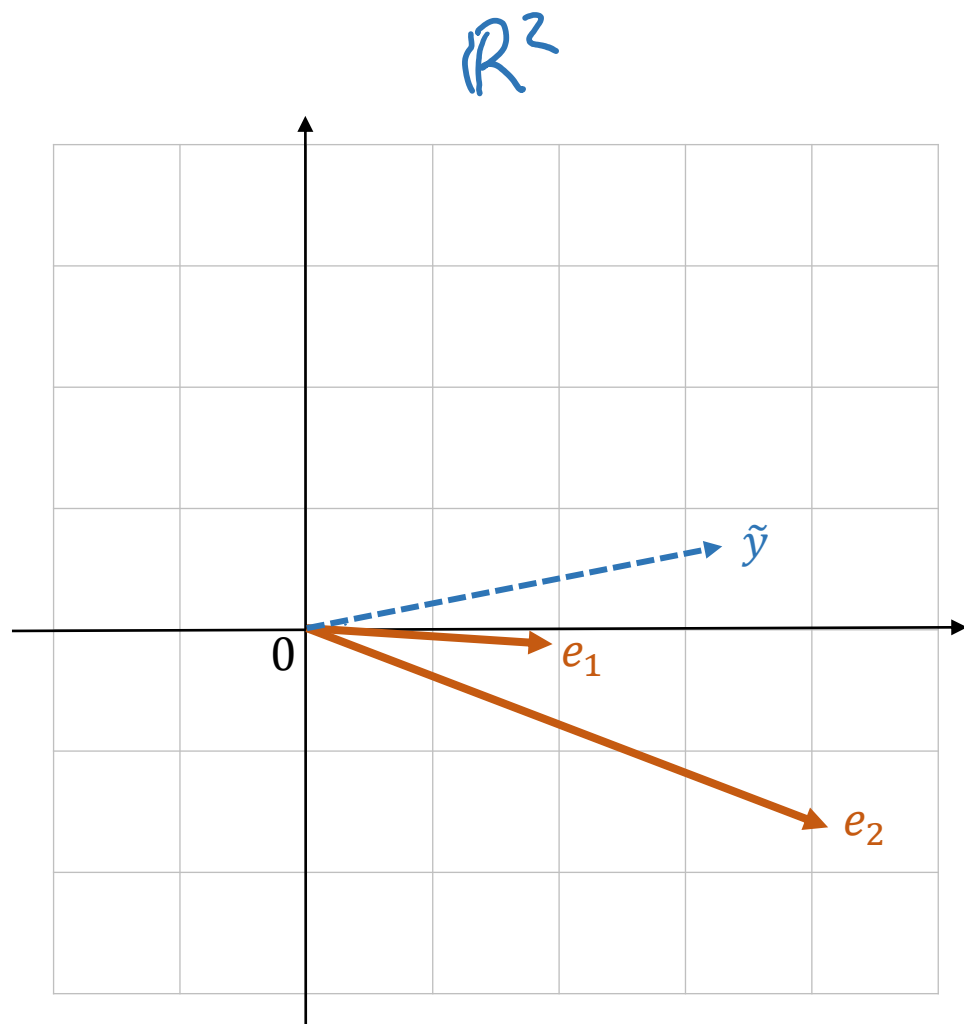
$$\mathcal{S} = \{\lambda_1 e_1 + \dots + \lambda_K e_K : \lambda_k \in \mathbb{R} \text{ for all } k\}$$

The vectors are *linearly dependent* if at least one of the e_k can be written as a linear combination of the others, i.e. there is some set of real numbers $(\lambda_1, \dots, \lambda_K)$ not all equal to zero such that

$$\lambda_1 e_1 + \dots + \lambda_K e_K = 0$$

If not, they are *linearly independent*, and

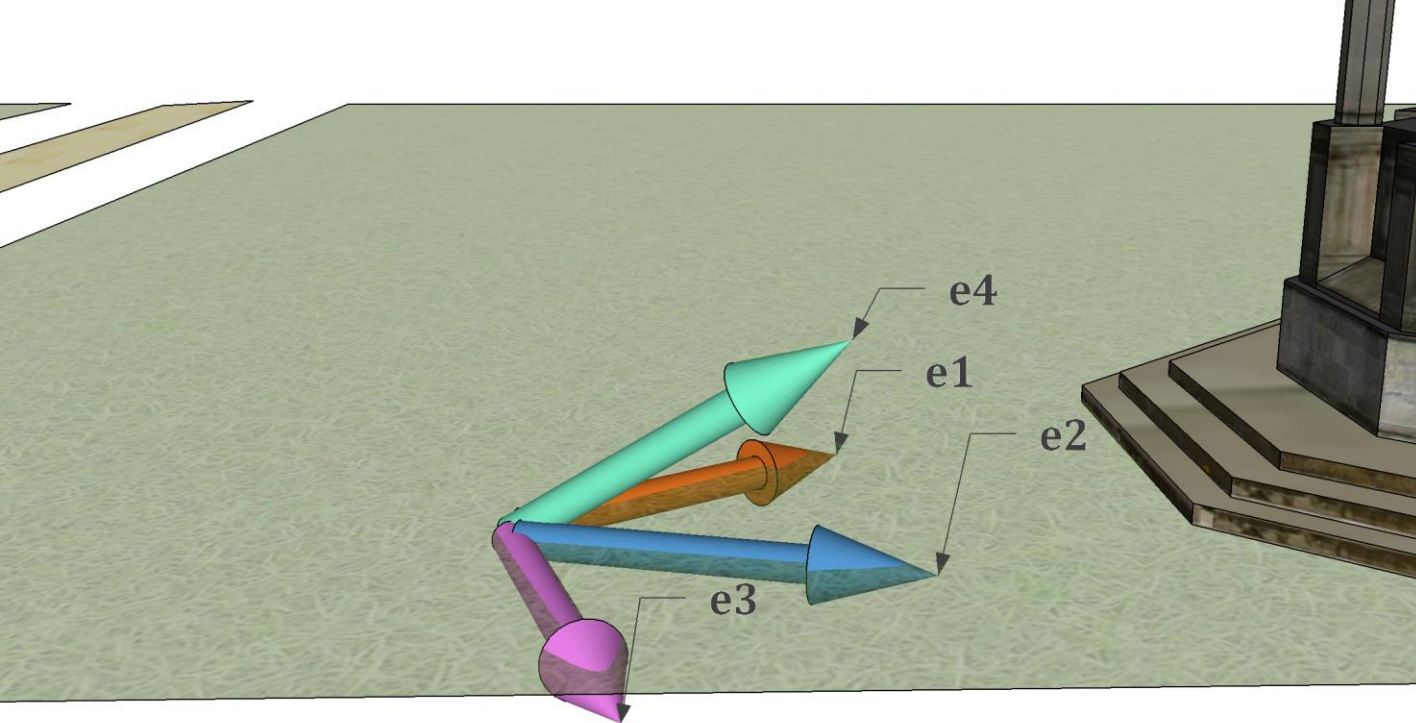
$$\lambda_1 e_1 + \dots + \lambda_K e_K = 0 \quad \Rightarrow \quad \lambda_1 = \dots = \lambda_K = 0$$



The subspace spanned by $\{e_1, e_2\}$ is $\mathbb{R}^2$

Any $\tilde{y} \in \mathbb{R}^2$ can be written as a linear combination of e_1 and e_2

❖ by eye, $\tilde{y} = 2.5e_1 - 0.3e_2$



The subspace spanned by $\{e_1, e_2, e_3, e_4\}$ is $\mathbb{R}^3$

Are $\{e_1, e_2, e_3, e_4\}$ linearly independent? **No.**

$$e_2 \approx 0.5e_1 + 0.5e_3$$

$$0.5e_1 - e_2 + 0.5e_3 = 0$$

If we discarded e_2 ...

Yes.

Are $\{e_1, e_3, e_4\}$ linearly independent? What's the span?

$\mathbb{R}^3$

If we discarded e_1 ...

Are $\{e_2, e_3, e_4\}$ linearly independent? What's the span?

Exercise 2.5.2

Are the following five vectors linearly independent? If not, find a subset that is.

$$e_1 = [1, 1, 1, 1]$$

$$e_2 = [0, 1, 1, 0]$$

$$e_3 = [1, 0, 0, 1]$$

$$e_4 = [1, 1, 1, 0]$$

$$e_5 = [0, 0, 0, 1]$$

Are these 5 vectors linearly independent?

No! $e_1 = e_2 + e_3$

$$e_1 = e_4 + e_5$$

$$y = \beta_1 e_1 + \dots + \beta_5 e_5 \\ = \beta_1' e_1 + \dots + \beta_4' e_4$$

So $\text{span}(\{e_1, e_2, e_3, e_4, e_5\})$

$= \text{span}(\{e_1, e_2, e_3, e_4\})$ since $e_5 = e_1 - e_4$

$= \text{span}(\{e_1, e_2, e_4\})$ since $e_3 = e_1 - e_2$.

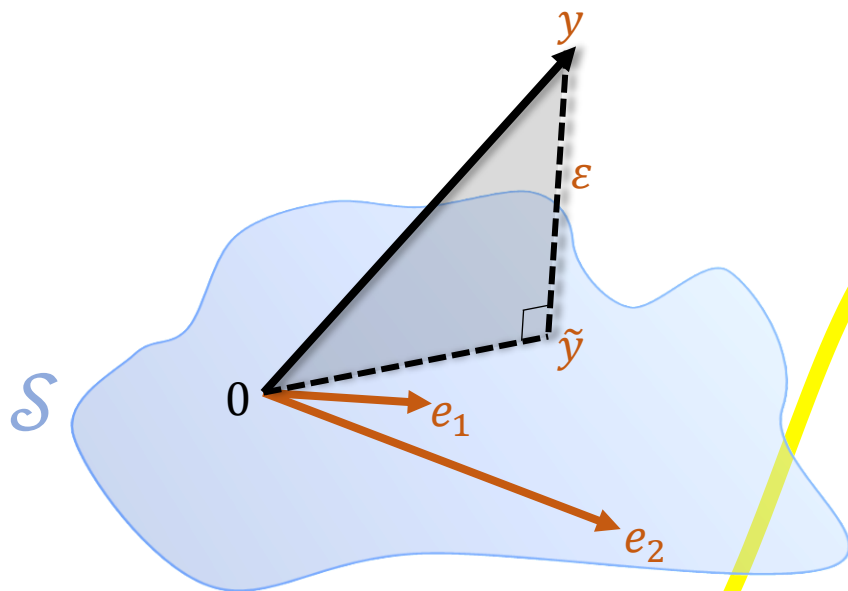
Are e_1, e_2, e_4 linearly independent?

Suppose $\lambda_1 e_1 + \lambda_2 e_2 + \lambda_4 e_4 = 0$

$$\Rightarrow \lambda_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \lambda_4 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = 0$$

$$\Rightarrow \left. \begin{array}{l} \lambda_1 + \lambda_4 = 0 \\ \lambda_1 + \lambda_2 + \lambda_4 = 0 \\ \lambda_1 = 0 \end{array} \right\} \Rightarrow \left. \begin{array}{l} \lambda_1 = 0 \\ \lambda_4 = 0 \\ \lambda_2 = 0 \end{array} \right\} \begin{array}{l} \text{ie coefficients are} \\ \text{all } 0. \\ \therefore \text{they are} \\ \text{linearly independent.} \end{array}$$

`np.linalg.matrix_rank(np.column_stack([ $e_1, \dots, e_K$ ]))` is $\begin{cases} < K & \text{if linearly dependent} \\ = K & \text{if linearly independent} \end{cases}$



What does “closest” even mean?

It means: find $\tilde{y} \in \mathcal{S}$ to minimize $\|y - \tilde{y}\|$

i.e. to minimize $\sqrt{\sum_i \varepsilon_i^2}$ where $\varepsilon = y - \tilde{y}$.

The minimization is over all $\tilde{y} \in \mathcal{S}$,
i.e. over all linear combinations of $\{e_1, \dots, e_K\}$.

GEOMETRY OF PROJECTIONS

Let $\mathcal{S}$ be the span of $\{e_1, \dots, e_K\}$.

What's the **closest** we can get to y , while staying in $\mathcal{S}$?

LINEAR MODELLING / LEAST SQUARES ESTIMATION

Given features $\{e_1, \dots, e_K\}$ let's approximate $y \approx \beta_1 e_1 + \dots + \beta_K e_K$.

What parameters give us the **best approximation**?

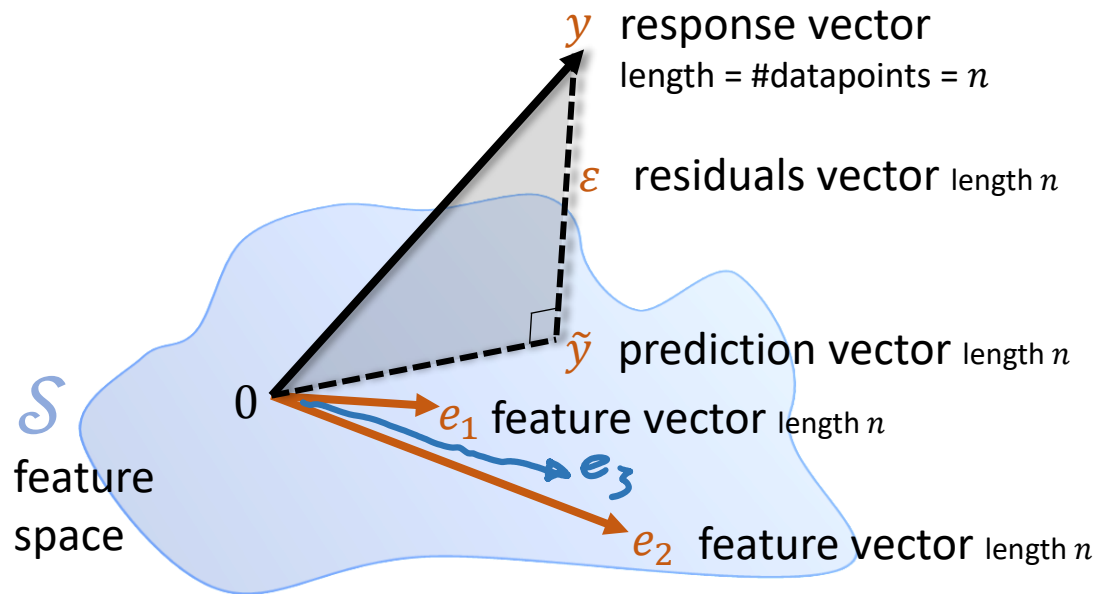
What does “best approximation” mean?

It means: find $\beta_1, \dots, \beta_K$ to minimize the mean square error,

$$\frac{1}{\text{\#datapoints}} \sum_i \varepsilon_i^2$$

where $\varepsilon = y - (\beta_1 e_1 + \dots + \beta_K e_K)$

These are exactly the same thing!



We can reach the same predictions $\tilde{y}$ in different ways — i.e. with different coefficients.

The span of $\{e_1, \dots, e_K\}$ is called the *feature space*

- ❖ Fitting a linear model $\equiv$ projecting the data y onto the feature space $\mathcal{S}$
- ❖ The residuals from the fitted model are perpendicular to the features: $\epsilon \cdot e_i = 0$ for all i
- ❖ The fitted parameters β_k tell us the coordinates of the predicted response, $\tilde{y} = \sum_k \beta_k e_k$
- ❖ If two models have different features but the same feature space, they'll make exactly the same prediction $\tilde{y}$ (but of course have different β_k)
- ❖ If the feature vectors are linearly independent, then the β_k are uniquely identified.
- ❖ Otherwise, the parameters have no intrinsic meaning, and we say they're *confounded*.

Why do these three models lead to such different estimates for α ?

Model A: $\text{temp} \approx \alpha \mathbf{1} + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t}) + \gamma \mathbf{t}$

Model B: $\text{temp} \approx \alpha \mathbf{1} + \beta_1 \sin(2\pi \mathbf{t}) + \beta_2 \cos(2\pi \mathbf{t}) + \gamma (\mathbf{t} - 2000)$

$$\Rightarrow \hat{\alpha} = -67.5^\circ\text{C}$$

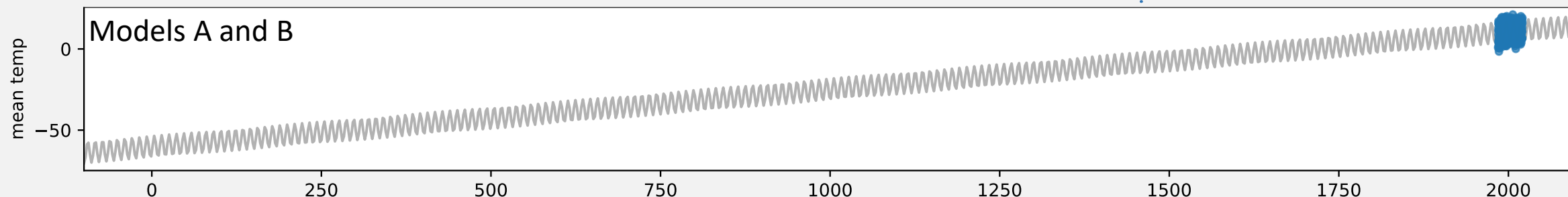
$$\Rightarrow \hat{\alpha} = 10.5^\circ\text{C}$$

EXERCISE. Write out the feature spaces. What do you notice?

Model A: feature space = $\text{span} \left(\left\{ \underline{\mathbf{1}}, \sin(2\pi \underline{\mathbf{t}}), \cos(2\pi \underline{\mathbf{t}}), \underline{\mathbf{t}} \right\} \right)$

B: $\text{span} \left(\left\{ \underline{\mathbf{1}}, \sin(2\pi \underline{\mathbf{t}}), \cos(2\pi \underline{\mathbf{t}}), \underline{\mathbf{t} - 2000} \right\} \right)$

These are the same feature space: $\underline{\mathbf{t} - 2000}$ is a linear combination of $\underline{\mathbf{1}}$ and $\underline{\mathbf{t}}$
and $\underline{\mathbf{t}}$ is a linear combination of $\underline{\mathbf{1}}$ and $\underline{\mathbf{t} - 2000}$



Models A and B are essentially the same, because they have the same feature space.

But because they use different bases for the feature space, they report different readouts.

 Sign in



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**

The Guardian

News website of the year

Let's build a model to
study this ourselves.