Do I trust this finding?
❖ Proportionate to what?
❖ Is it cherry picking?
❖ Why the scare quotes?

Stop and search

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
<table>
<thead>
<tr>
<th>Force</th>
<th>Date</th>
<th>Lat/Lng</th>
<th>Object of search</th>
<th>Gender</th>
<th>Age range</th>
<th>Officer-defined ethnicity</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 15:44:04+00:00</td>
<td>(52.43,-0.142)</td>
<td>Controlled drugs</td>
<td></td>
<td></td>
<td></td>
<td>A no further action disposal</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 15:35:41+00:00</td>
<td>(52.43,-0.142)</td>
<td>Firearms</td>
<td>Male</td>
<td>25-34</td>
<td>White</td>
<td>Khat or Cannabis warning</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 14:44:04+00:00</td>
<td>(52.43,-0.142)</td>
<td>Firearms</td>
<td>Male</td>
<td>25-34</td>
<td>White</td>
<td>Khat or Cannabis warning</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 03:44:14+00:00</td>
<td>(52.58,-0.244)</td>
<td>Offensive weapons</td>
<td>Male</td>
<td></td>
<td>Other</td>
<td>A no further action disposal</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 02:34:16+00:00</td>
<td>(52.59,-0.247)</td>
<td>Controlled drugs</td>
<td>Male</td>
<td>25-34</td>
<td>White</td>
<td>Arrest</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-31 02:27:10+00:00</td>
<td>(52.21,0.124)</td>
<td>Controlled drugs</td>
<td>Male</td>
<td>18-24</td>
<td>White</td>
<td>A no further action disposal</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-30 22:28:13+00:00</td>
<td>(52.45,-0.117)</td>
<td>Controlled drugs</td>
<td>Female</td>
<td>over 34</td>
<td>White</td>
<td>A no further action disposal</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-30 20:24:13+00:00</td>
<td>(52.32,-0.0708)</td>
<td>Controlled drugs</td>
<td>Male</td>
<td>10-17</td>
<td>White</td>
<td>Summons / charged by post</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-30 14:26:58+00:00</td>
<td>(52.57,-0.24)</td>
<td>Controlled drugs</td>
<td>Male</td>
<td>over 34</td>
<td>Asian</td>
<td>A no further action disposal</td>
</tr>
<tr>
<td>cambridgeshire</td>
<td>2023-08-30 14:13:45+00:00</td>
<td>(52.57,-0.24)</td>
<td>Controlled drugs</td>
<td>Male</td>
<td>25-34</td>
<td>Black</td>
<td>Arrest</td>
</tr>
</tbody>
</table>

Log of England+Wales stop-and-search incidents, from the UK home office https://data.police.uk/
In a dataset of police stop-and-search records, is there evidence of ethnic bias? What about gender bias? If so, do these biases intersect, or is the net bias simply additive?

I can predict exactly what will happen to a person when they’re stopped by the police!

Just tell me their gender. And ethnicity. And location. And whether they’re left or right handed. And whether they have a pet cat or a dog. And what their pet is called. …

What was the question again?

**ML mindset**
We just want to make good predictions, we don’t care about the parameters

**Science mindset**
We have questions in mind, and we can answer them by looking at our model’s parameters

**ML**
building models from data

**data science**
Can I set up a model with a parameter that measures the quantity I’m interested in?
Let $y_i$ be the response in row $i$, $y_i = \begin{cases} 1 & \text{if the police found something} \\ 0 & \text{if the police found nothing} \end{cases}$

The average response in ethnic group $k$ is

$$\text{avg } y = \frac{\sum_{i: \text{eth } i = k} y_i}{|\{i: \text{eth } i = k\}|} = \frac{\#\text{finds}}{\#\text{stops}} = P(\text{find something})$$

Let's fit the model $y_i \approx \alpha + \beta_{\text{eth}_i}$

i.e. $P(\text{find something}) \approx \alpha + \beta_k$ for a person in ethnic group $k$

If $\beta_k < 0$, that means $P(\text{find something})$ is low compared to other ethnic groups, i.e. the police are stopping relatively more innocent people.
Let’s fit a model using officer-defined ethnicity as the predictor,

\[ y \approx \alpha + \beta_{\text{eth}} \]

Writing it as a linear model with one-hot coding,

\[ y \approx \alpha + \beta_{\text{Asian}} 1_{\text{eth} = \text{Asian}} + \beta_{\text{Black}} 1_{\text{eth} = \text{Black}} + \beta_{\text{Mixed}} 1_{\text{eth} = \text{Mixed}} + \beta_{\text{Other}} 1_{\text{eth} = \text{Other}} + \beta_{\text{White}} 1_{\text{eth} = \text{White}} \]

Ethnicity levels:

- Asian
- Black
- Mixed
- Other
- White

```python
ethnicity_levels = np.unique(eth)
eth_onehot = [np.where(eth==k,1,0) for k in ethnicity_levels]

model = sklearn.linear_model.LinearRegression()
model.fit(np.column_stack(eth_onehot), y)
α,βs = model.intercept_, model.coef_

print(f'α = {α}')
for k,β in zip(ethnicity_levels, βs):
    print(f'β[{k}] = {β}')
```

\[ \alpha = -34037792910.00365 \]
\[ \beta[\text{Asian}] = 34037792910.26522 \]
\[ \beta[\text{Black}] = 34037792910.265717 \]
\[ \beta[\text{Mixed}] = 34037792910.2939 \]
\[ \beta[\text{Other}] = 34037792910.26049 \]
\[ \beta[\text{White}] = 34037792910.261383 \]
§2.5 The geometry of linear models

NST Maths A, Michaelmas

1.7.2 Shortest distance of a point from a plane

Consider the plane that passes through point \( A \) (given by position vector \( \mathbf{a} \)) and that has unit normal \( \mathbf{h} \). From (1)), the equation for the plane is defined by the points \( r \) satisfying

\[
(r - a) \cdot h = 0.
\]

The closest point on the plane to a point \( P \), given by position vector \( \mathbf{p} \), is the point \( Q \), where \( \mathbf{QP} \) is normal to the plane and \( \mathbf{QP} \parallel \mathbf{h} \).

Example: Distance of point from plane

- What is distance of point \( A \) with position vector \( \mathbf{a} \) from plane \( \mathbf{r} \cdot \mathbf{h} = I \)?

- Line containing \( A \) and point of closest approach of plane to \( A \) must be \( \parallel \mathbf{h} \); has equation

\[
\mathbf{r} = \mathbf{a} + \lambda \mathbf{h}
\]

- Line meets plane where \( \mathbf{r} \cdot \mathbf{h} = I \), i.e. where

\[
I = \mathbf{a} \cdot \mathbf{h} + \lambda
\]

- \( \lambda \) is distance along line from \( a \) so required distance is \( |\mathbf{a} - \mathbf{h} - I| \)

NST Maths B, Michaelmas

NST Maths A, Easter

Orthogonality - 2/3

- The vectors \( \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \) are a basis of orthonormal vectors in \( \mathbb{R}^3 \):

\[
\mathbf{e}_1 \cdot \mathbf{e}_2 = \delta_{12}.
\]

- We use the orthogonality properties (2) to calculate the components of \( \mathbf{u} \):

\[
\mathbf{e}_1 \cdot \mathbf{u} = \mathbf{e}_1 \times \mathbf{e}_2 \times 0 + \mathbf{e}_3 \times 0 = \mathbf{e}_1.
\]

In general

\[
\mathbf{e}_i = \mathbf{e}_1 \cdot \mathbf{a}, \quad \text{for} \quad i = 1, 2, 3.
\]

- The above generalises to Euclidean space \( \mathbb{R}^n \) with

\[
\mathbf{a} = \mathbf{e}_1 \mathbf{e}_1, \quad \text{for} \quad i = 1, \ldots, n.
\]

The components \( \mathbf{e}_i \mathbf{e}_i \) are evaluated in the same way as in the case with \( n = 3 \) because (2) and (3) still hold, but with \( i, j \) now in the range 1 to \( n \).

---

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 \( \mathbf{u}, \mathbf{v}, \mathbf{w} \in V \), \( \mathbf{u} + \mathbf{v} + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}) \). (Associativity.)

A2. For any vectors \( \mathbf{u}, \mathbf{v} \in V \), \( \mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u} \). (Commutativity.)

A3. There is a vector in \( V \) denoted \( \mathbf{0} \), called the zero vector for which \( \mathbf{u} + \mathbf{0} = \mathbf{u} \) \( \forall \mathbf{u} \in V \).

A4. For each vector \( \mathbf{u} \in V \) there is a vector in \( V \) denoted \( -\mathbf{u} \) for which \( \mathbf{u} + (-\mathbf{u}) = \mathbf{0} \). (Inverse.)

A5. For any \( \mathbf{u} \in K \) and any \( \mathbf{u}, \mathbf{v} \in V \), \( a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v} \). (Distributivity.)

A6. For any \( \mathbf{u}, \mathbf{v} \in V \), \( a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v} \).

A7. For any \( a, b \in K \) and any \( \mathbf{u} \in V \), \( (ab)\mathbf{u} = a(b\mathbf{u}) \).

A8. For the unit scalar \( 1 \in K \) and any \( \mathbf{u} \in V \), \( 1\mathbf{u} = \mathbf{u} \).
The *subspace spanned* by a collection of vectors \( \{e_1, \ldots, e_K\} \) is the set of all linear combinations
\[
S = \{ \lambda_1 e_1 + \cdots + \lambda_K e_K : \lambda_k \in \mathbb{R} \text{ for all } k \}
\]

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 \( (\lambda_1, \ldots, \lambda_K) \) not all equal to zero such that
\[
\lambda_1 e_1 + \cdots + \lambda_K e_K = 0
\]
If not, they are *linearly independent*, and
\[
\lambda_1 e_1 + \cdots + \lambda_K e_K = 0 \Rightarrow \lambda_1 = \cdots = \lambda_K = 0
\]

\[\text{np.linalg.matrix_rank(np.column_stack([e_1, \ldots, e_K]))}\] is \(< K\) if linearly dependent
\[= K\] if linearly independent
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? \textbf{No}.

If we discarded \( e_2 \) ...
Are \( \{e_1, e_3, e_4\} \) linearly independent? What's the span? \textbf{They are linearly independent, span is } \mathbb{R}^3.

If we discarded \( e_1 \) ...
Are \( \{e_2, e_3, e_4\} \) linearly independent? What's the span? \textbf{They are linearly independent, span is } \mathbb{R}^3.
Exercise 2.5.2
Are the following five vectors linearly independent? If not, find a subset that is.

\[
\hat{e}_1 = [1,1,1,1] \\
\hat{e}_2 = [0,1,1,0] \\
\hat{e}_3 = [1,0,0,1] \\
\hat{e}_4 = [1,1,1,0] \\
\hat{e}_5 = [0,0,0,1]
\]

\[
e_1 = e_2 + e_3 \\
e_1 = e_4 + e_5
\]

So \(\text{span} \{ \hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4, \hat{e}_5 \} \)
\[
= \text{span} \{ \hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4 \} \quad \text{since} \quad \hat{e}_5 = \hat{e}_1 - \hat{e}_4 \\
= \text{span} \{ \hat{e}_1, \hat{e}_2, \hat{e}_4 \} \quad \text{since} \quad \hat{e}_3 = \hat{e}_1 - \hat{e}_2.
\]

Q. Are \(\{\hat{e}_1, \hat{e}_2, \hat{e}_4\}\) linearly independent?

Suppose \(\lambda_1 \hat{e}_1 + \lambda_2 \hat{e}_2 + \lambda_4 \hat{e}_4 = 0\)

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

\[
\Rightarrow \begin{cases} 
\lambda_1 + \lambda_4 = 0 \\
\lambda_1 + \lambda_2 + \lambda_4 = 0 \\
\lambda_1 + \lambda_2 + \lambda_4 = 0 \\
\lambda_1 = 0
\end{cases} \Rightarrow \begin{cases} 
\lambda_1 = 0 \\
\lambda_2 = 0 \\
\lambda_4 = 0
\end{cases} \Rightarrow \text{they are linearly independent.}
GEOMETRY
Let $S$ be the span of $\{e_1, \ldots, e_K\}$. What's the closest we can get to $y$, while staying in $S$?

LINEAR MODELLING / LEAST SQUARES ESTIMATION
Given features $\{e_1, \ldots, e_K\}$ let's approximate $y \approx \beta_1 e_1 + \cdots + \beta_K e_K$. What parameters give us the best approximation?

What does "closest" mean? It means: find $\hat{y} \in S$ to minimize $\|\hat{y} - y\|$. In other words, to minimize $\sqrt{\sum \epsilon^2}$ where $\epsilon = y - \hat{y}$.

The minimization is over all $\hat{y} \in S$. Every $\hat{y} \in S$ can be written as a linear combination of $e_1, \ldots, e_K$.

What does "best approx." mean? It means: find $\beta_1, \ldots, \beta_K$ to minimize $\frac{1}{\text{#data points}} \sum \epsilon^2$ where $\epsilon = y - (\beta_1 \hat{e}_1 + \cdots + \beta_K \hat{e}_K)$. The span of $\{e_1, \ldots, e_K\}$ is called the feature space.
Q. Is there a unique way to write $\hat{y}$ as a linear combination of $\{e_1, \ldots, e_K\}$?
A. If they are linearly independent, yes. If they’re not linearly independent, maybe not.

Q. When we run least squares estimation, does it always return the same parameter estimates?
A. If the features are NOT linearly independent, different runs might give different parameters (and if some other data scientist reports different parameter estimates, our audience will be confused!)
§2.6 Interpreting parameters

- Write out the predicted response for a few typical / representative datapoints. 
  This helps see what the parameters mean.

- Write out the features. 
  If two models have different features but the same feature space, then (once fitted) they make the same predictions on the dataset.

- Check if the features are linearly dependent. 
  If so, the parameters have no intrinsic meaning. 
  We say the features are *confounded*, and the parameters are *non-identifiable*. 
These three models yielded very different estimates for $\alpha$. Why?

Model 0: $\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi \ t) + \beta_2 \cos(2\pi \ t) \quad \Rightarrow \quad \hat{\alpha} = 10.6^\circ C$

Model A: $\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi \ t) + \beta_2 \cos(2\pi \ t) + \gamma \ t \quad \Rightarrow \quad \hat{\alpha} = -60.2^\circ C$

Model B: $\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi \ t) + \beta_2 \cos(2\pi \ t) + \gamma (t - 2000) \quad \Rightarrow \quad \hat{\alpha} = 10.5^\circ C$

**EXERCISE.** Write out the predicted response for a few typical / representative datapoints.

- **Model 0:** at $t = 0$, pred. temp. $= \alpha + \beta_2$ over a single sinusoid, $\alpha$ is avg. temperature.

- **Model A:** at $t = 0$, pred. temp. $= \alpha + \beta_2 + 2000\gamma$
  
  at $t = 2000$, $= \alpha + \beta_2 + 2000\gamma$

  over a single sinusoid in 1 BC, $\alpha$ is avg. temperature.

- **Model B:** at $t = 0$, pred temp $= \alpha + \beta_2 - 2000\gamma$

  at $t = 2000$, $= \alpha + \beta_2$

  over a single sinusoid in 2000 AD, $\alpha$ is avg. temperature.
These three models yielded very different estimates for $\alpha$. Why?

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
<th>$\hat{\alpha}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 0</td>
<td>$\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t)$</td>
<td>$10.6 \degree C$</td>
</tr>
<tr>
<td>Model A</td>
<td>$\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t$</td>
<td>$-60.2 \degree C$</td>
</tr>
<tr>
<td>Model B</td>
<td>$\text{temp} \approx \alpha_1 + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma (t-2000)$</td>
<td>$10.5 \degree C$</td>
</tr>
</tbody>
</table>

**EXERCISE.** Write out the features.

Model A: features \{ $\hat{t}$, $\sin(2\pi t)$, $\cos(2\pi t)$, $\hat{t}$ \}

Model B: features \{ $\hat{t}$, $\sin(2\pi t)$, $\cos(2\pi t)$, $\hat{t} - 2000 \times \hat{t}$ \}

Model A and B have the same feature space.

They're essentially the same model, just with different parameterization.

Model A asks: tell me the avg. temp in 1BC
Model B asks: tell me the avg. temp in 2000BC

This can be written as a linear combination of $\hat{t}$ and $\hat{t}$.
These three models yielded very different estimates for \(\alpha\). Why?

**Model 0:**
\[
\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) \quad \Rightarrow \quad \hat{\alpha} = 10.6 \, ^\circ C
\]

**Model A:**
\[
\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t \quad \Rightarrow \quad \hat{\alpha} = -60.2 \, ^\circ C
\]

**Model B:**
\[
\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma (t-2000) \quad \Rightarrow \quad \hat{\alpha} = 10.5 \, ^\circ C
\]

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

But because they use different representations of the feature space, they report different readouts.
Can I set up a model with a parameter that measures the quantity I’m interested in?
Example 2.6.4

The UK Home Office makes available a dataset of police stop-and-search incidents. We wish to investigate whether there is racial bias in police decisions to stop-and-search. Consider the linear model

\[ y_i \approx \alpha + \beta_{\text{eth}_i} \]

where \( \text{eth}_i \) is the officer-defined ethnicity for record \( i \), and \( y_i \) records the outcome: \( y_i = 1 \) if the police found something, 0 otherwise.

(a) Write this as a linear equation using one-hot coding.
(b) Are the parameters identifiable? If not, rewrite the model so that they are.
(c) Does the model suggest there is racial bias in policing actions?

(a)

\[ y \approx \alpha \cdot 1 + \beta_{\text{As}} e_{\text{As}} + \beta_{\text{Bl}} e_{\text{Bl}} + \beta_{\text{Mi}} e_{\text{Mi}} + \beta_{\text{Om}} e_{\text{Om}} + \beta_{\text{Ow}} e_{\text{Ow}} \quad \text{where} \quad e_k = 1_{\text{eth}_i = k} \]

(b) These are linearly dependent: \( 1 = e_{\text{As}} + e_{\text{Bl}} + e_{\text{Mi}} + e_{\text{Om}} + e_{\text{Ow}} \)

So the parameters are not identifiable, i.e. we’re likely to get silly answers out of linear model fitting.
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.

Upload your answers to Moodle by Sunday for presentation / discussion next week