Prediction

Class 17
Predicting heights

Oval-shaped

Moderate positive correlation

How can we predict child height from the parents’ average height?
Approach to prediction

To predict an outcome for an individual, find others who are like that individual, and whose outcome you know.

Use those outcomes as the basis for your prediction.
Predicted heights

Average of parents’ heights
Nearest neighbor regression

A method for prediction:

Group each $x$ with similar (nearby) $x$ values
Average the corresponding $y$ values for each group

For each $x$ value, the prediction is the average of the $y$ values in its nearby group.

The graph of these predictions is the “graph of averages”.

If the association between $x$ and $y$ is linear, then points in the graph of averages tend to fall on a line.
Correlation
When we have two numeric variables, we can look for a correlation between them, which is measured with a correlation coefficient $r$.
Watch out for...

Claims of causation due to confounders

New England Journal of Medicine, 2012
Watch out for…

Claims of causation due to confounders

Non-linear relationships
Watch out for...

Claims of causation due to confounders

Non-linear relationships
Watch out for...

Claims of causation due to confounders

Non-linear relationships

Outliers
Watch out for…

Claims of causation due to confounders

Non-linear relationships

Outliers

Potentially misleading correlations based on aggregated data

Averages for each state
Misleading correlation on (fake) aggregated data

$r = 0.98$

$r = -0.09$
Misleading correlation on (fake) aggregated data

$r = 0.98$

$r = -0.09$
Linear regression
Linear regression models the relationship between two variables by fitting a straight line to the data.

Linear regression can be seen as a way to quantify and predict the relationship between the variables, with the slope of the line related to the strength and direction of the correlation coefficient $r$, but also taking into account the actual values rather than just the strength of the linear relationship.
We've already been using linear regression:

```python
feb.scatter(
    "Year",
    "Air temperature (C)",
    fit_line=True
)
```
To use linear regression for prediction, given an $x$ value, you just predict the $y$ value that’s on the line, based on the data you “trained” the model on (that is, the data you fit the line to).

This prediction is called the *regression estimate*. 
Acknowledgments

This class incorporates material from:

Stephen Freund and Katie Keith, Williams College
Data 8, University of California, Berkeley (CC BY-NC-SA)
