Lecture 27
Linear Regression
Announcements
Correlation (Review)
The Correlation Coefficient $r$

- Measures **linear** association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*
## Definition of $r$

### Correlation Coefficient ($r$)

$\text{Correlation Coefficient } (r) = \frac{\text{average of product of } x \text{ in standard units and } y \text{ in standard units}}{}$

Measures how clustered the scatter is around a straight line.
Prediction
Galton's Heights
Galton's Heights
Galton's Heights
Where is the prediction line?

\[ r = 0.99 \]
Where is the prediction line?

\[ r = 0.0 \]
Where is the prediction line?

$r = 0.5$
Where is the prediction line?

\[ r = 0.2 \]
Nearest Neighbor Regression

A method for prediction:

- Group each x with a representative x value (rounding)
- Average the corresponding y values for each group

For each representative x value, the corresponding prediction is the average of the y values in the group.

Graph these predictions.

If the association between x and y is linear, then points in the graph of averages tend to fall on the regression line.
Linear Regression

(Demo)
Regression to the Mean

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y's)

*On average*, y deviates from 0 less than x deviates from 0.

\[ y(\text{su}) = r \times x(\text{su}) \]

Not true for all points — a statement about averages.
Slope & Intercept
Regression Line Equation

In original units, the regression line has this equation:

\[
\frac{\text{estimate of } y - \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x - \text{average of } x}{\text{SD of } x}
\]

- estimated \( y \) in standard units
- \( x \) in standard units

Lines can be expressed by slope & intercept

\[
y = \text{slope} \times x + \text{intercept}
\]
Regression Line

Standard Units

Original Units

$(0, 0)$

$1$

$r$

$(\text{Average x, Average y})$

$r \times \text{SD y}$

$\text{SD x}$
Slope and Intercept

estimate of \( y = \text{slope} \times x + \text{intercept} \)

\[
\text{slope of the regression line} = r \times \frac{\text{SD of } y}{\text{SD of } x}
\]

\[
\text{intercept of the regression line} = \text{average of } y - \text{slope} \times \text{average of } x
\]

(Demo)
Scenario Question

You use a regression line to predict height based on weight and get a slope of .52 inches per pound.

I eat a lot of ice cream and gain 1 pound.

**True or False:**
My regression line predicts I will gain .52 inch on my height.
The regression line is a **statement about averages**. Given two groups of people with 1 lb difference, we expect the average height of the heavier group to be .52 inches greater than the average height of the lighter group.

The regression line is based on a snapshot of time and looking at holistic trends. We are **not** following 1 person and intending to make a prediction about them.