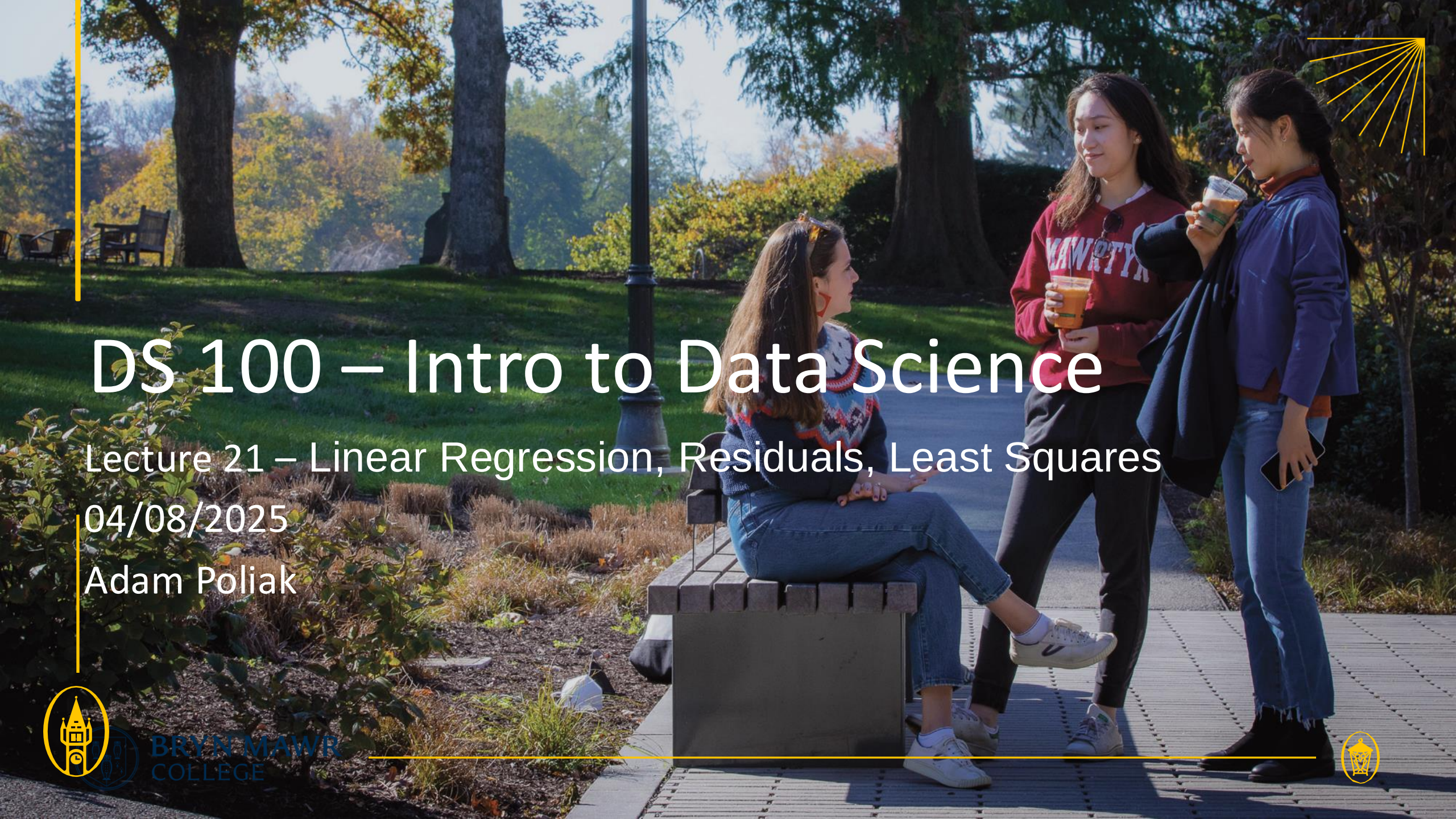


DS 100 – Intro to Data Science

Lecture 21 – Linear Regression, Residuals, Least Squares

04/08/2025

Adam Poliak

BRYN MAWR COLLEGE

Announcements

Midsemester feedback form - https://forms.gle/M4jVdGTDgQknWAwo6

Lab 08 (due Friday April 11th)

HW07 (due Wednesday April 9th), HW08 (due Wednesday April 16th)

Project 2 (due Friday April 11th)

No class: 04/10, 04/15, 04/17

Project 2 (due Monday April 14th)

Project 3 (due Friday May 2nd)

Prediction

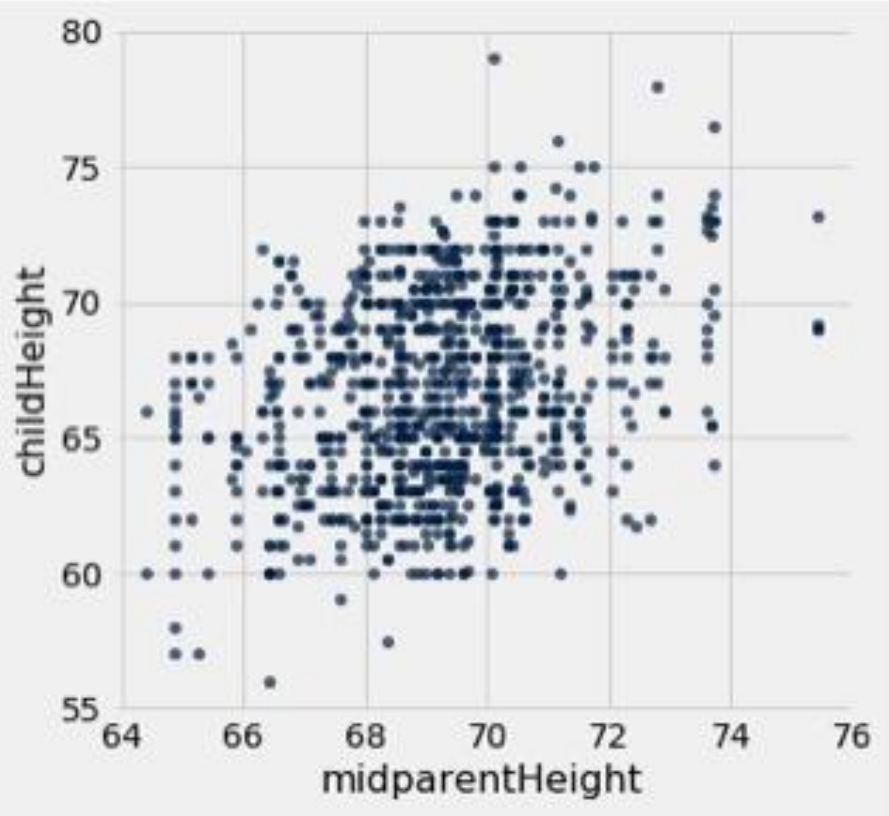
Guess the future

Based on incomplete information

One way of making predictions:

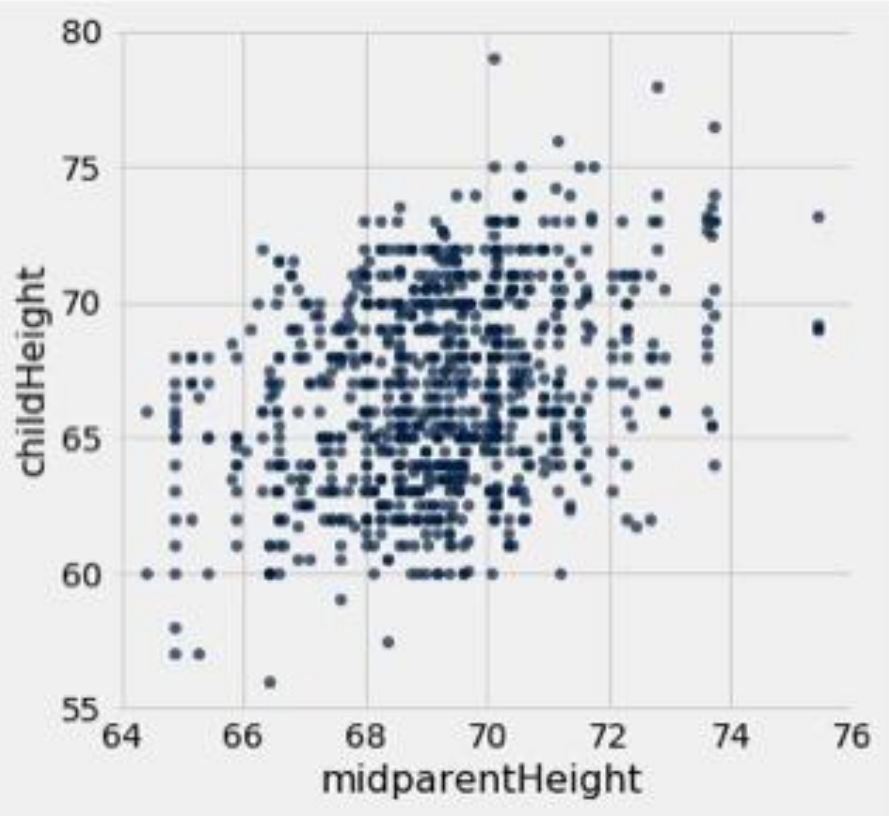
- To predict an outcome for an individual,
- find others who are like that individual
- and whose outcomes you know.
- Use those outcomes as the basis of your prediction.

Galton's Heights



Goal: Predict the height of a new child, based on that child's midparent height

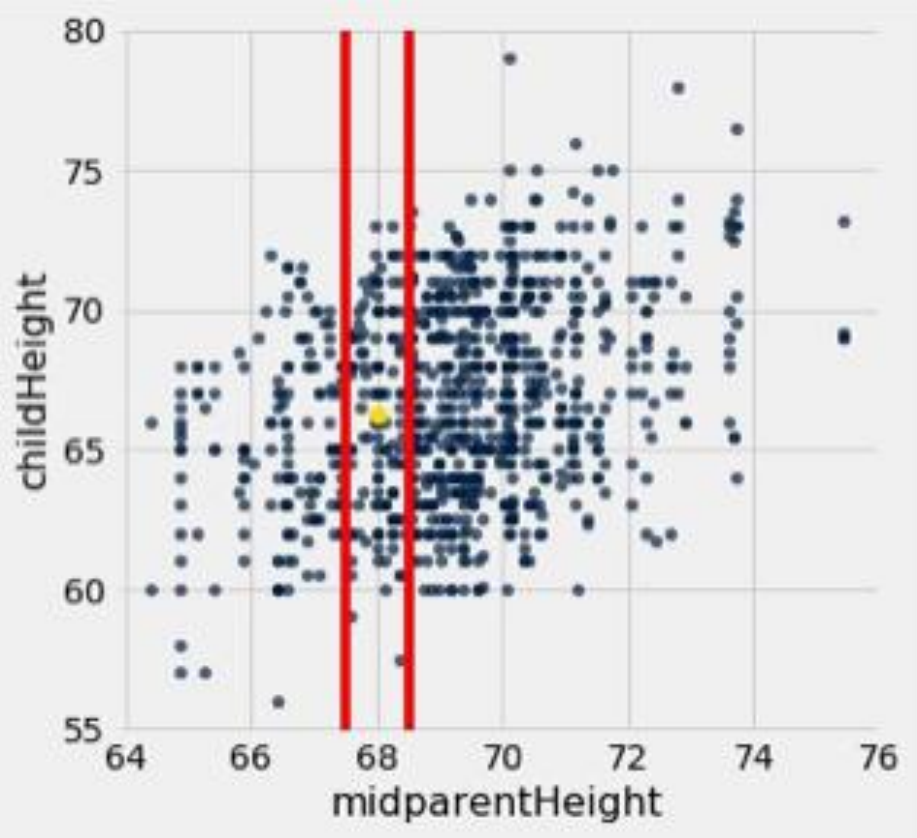
Galton's Heights



How can we predict a child's height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

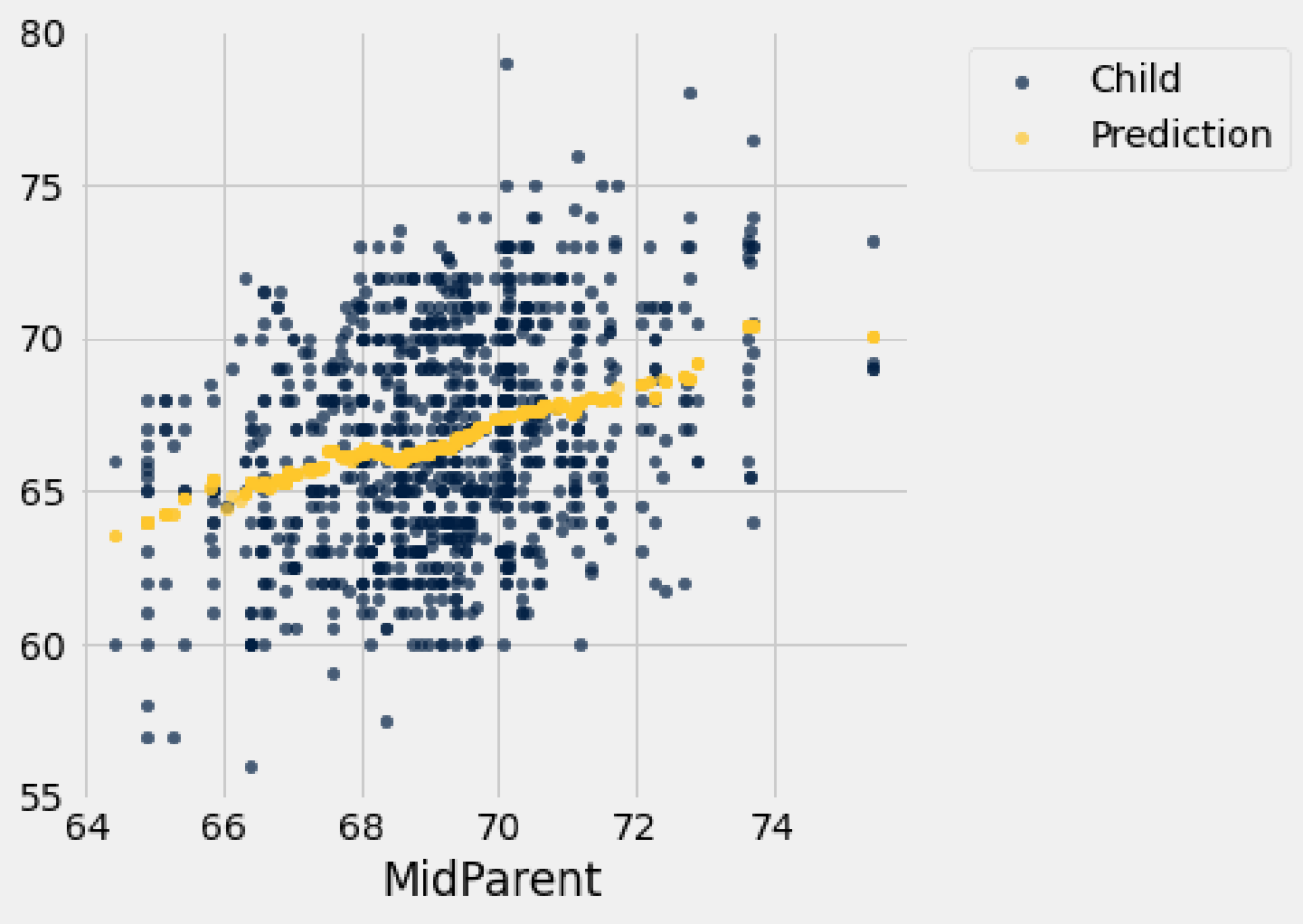
Galton's Heights



How can we predict a child's height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

Predicted Heights




Correlation
BRYN MAWR
COLLEGE

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*

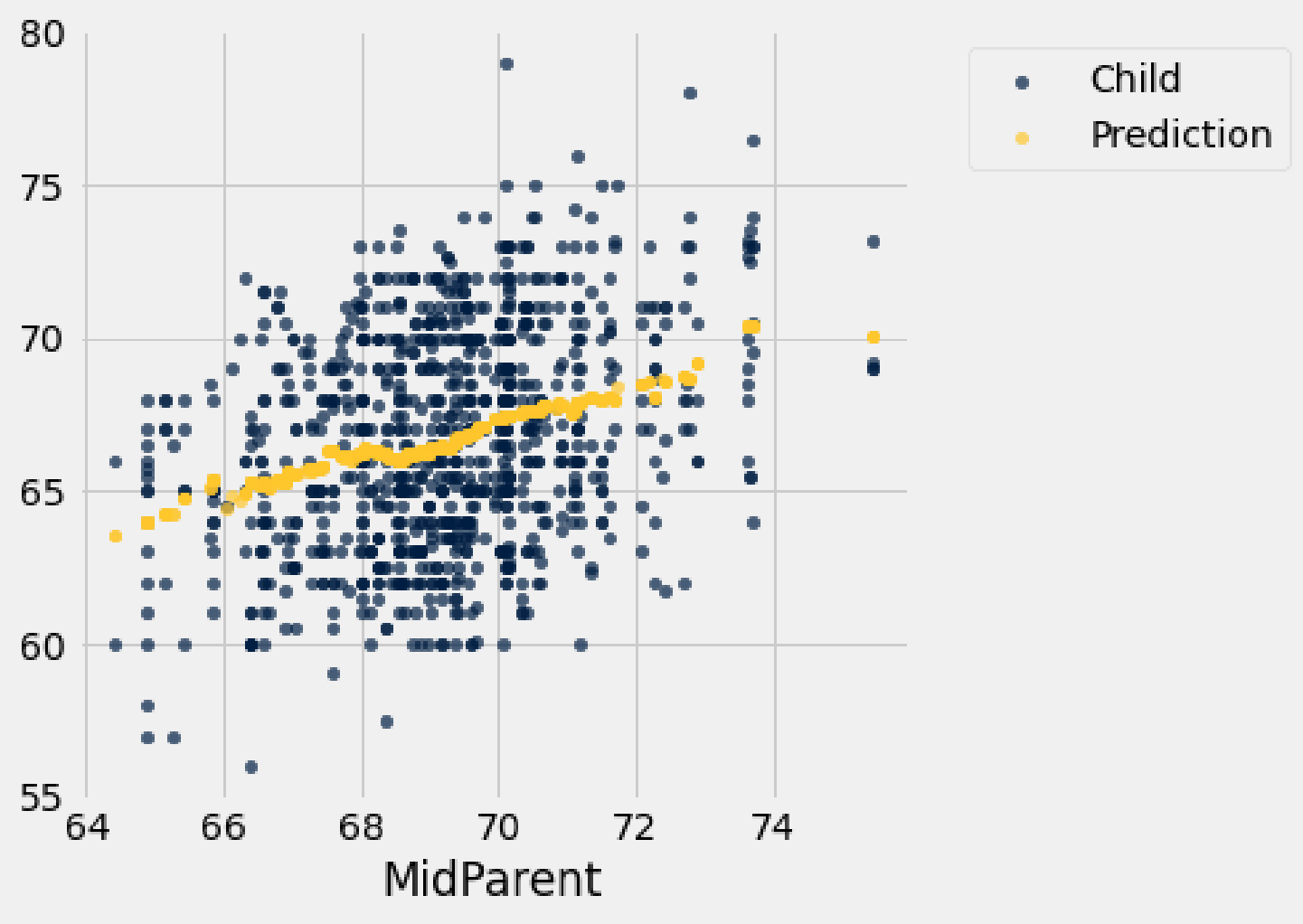
Correlation Coefficient (r) =

average of product of standard(x) and standard(y)

Steps: 4 3 2 1

Measures how clustered the scattered data are around a straight line

Predicted Heights



Graph of Averages

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.

The line is called the **regression line**

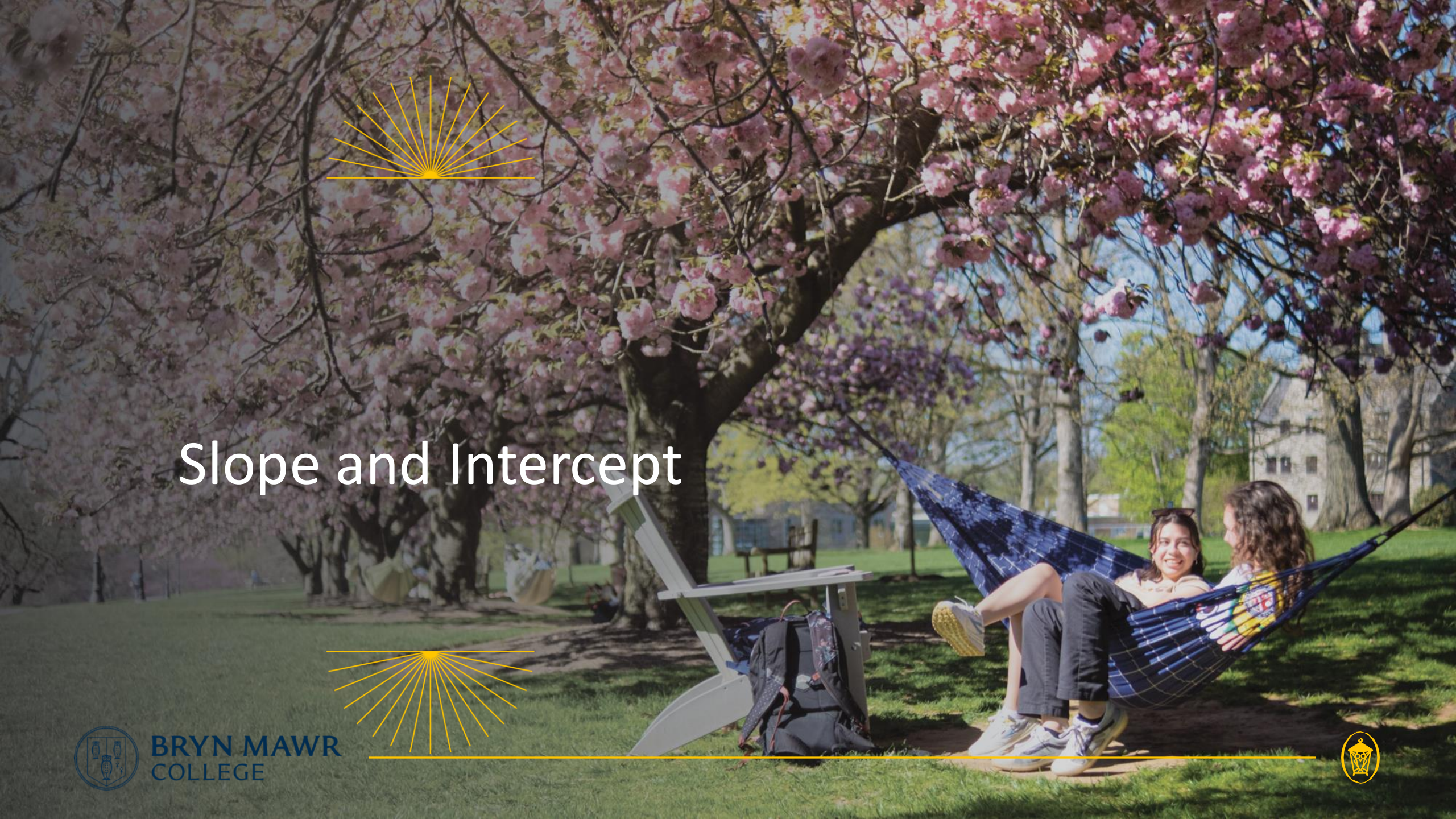

Linear Regression
BRYN MAWR
COLLEGE

Linear Regression

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)

$$y_{su} = r \times x_{su}$$



Slope and Intercept

BRYN MAWR COLLEGE

Regression Line Equation

$$y_{su} = r \times x_{su}$$

In original units, the regression line has this equation:

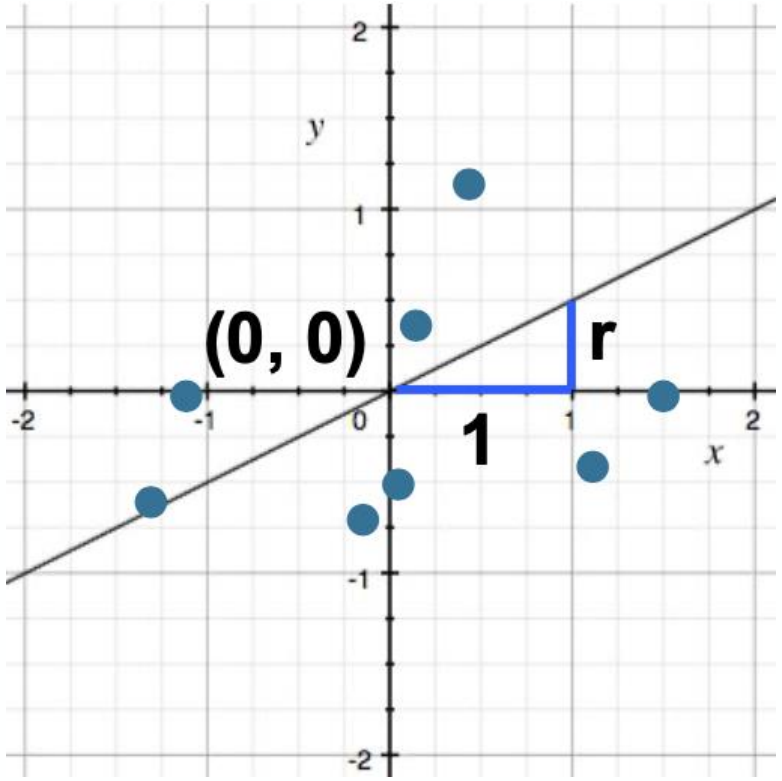
$$\frac{estimate\ of\ y\ - mean(y)}{SD\ of\ y} = r \times \frac{given\ x\ - mean(x)}{SD\ of\ x}$$

Lines can be expressed by *slope* & *intercept*

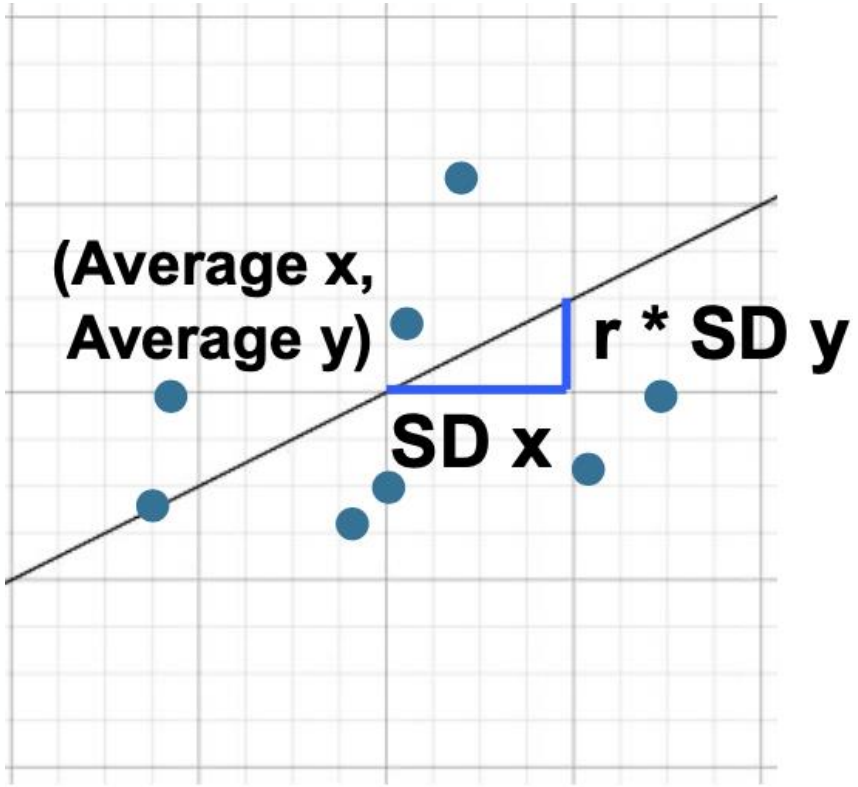
$$y = slope \times x + intercept$$

Regression Line

Standard Units



Original Units



Slope and Intercept

$$estimate\ of\ y = slope\ * x + intercept$$

$$\boldsymbol{slope\ of\ the\ regression\ line}$$

$$r\ * \frac{SD\ of\ y}{SD\ of\ x}$$

$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

$$mean(y) - slope\ \times\ mean(x)$$

Prediction with Linear Regression

Goal: Predict *y* using *x*

Examples:
Predict *# hospital beds available* using *air pollution*

Predict *house prices* using *house size*

Predict *# app users* using # app downloads

Regression Estimate

Goal: Predict *y* using *x*

To find the regression estimate oy *y*:

Convert the given *x* to standard units

Multiply by *r*

That's the regression estimate of *y,* but:

- It's in standard units
- So convert it back to the original units of *y*

Regression Line Estimate

In original units, the regression line has this equation:

$$y_{su} = r \times x_{su}$$

$$\frac{estimate\ of\ y\ - mean(y)}{SD\ of\ y} = r \times \frac{given\ x\ - mean(x)}{SD\ of\ x}$$

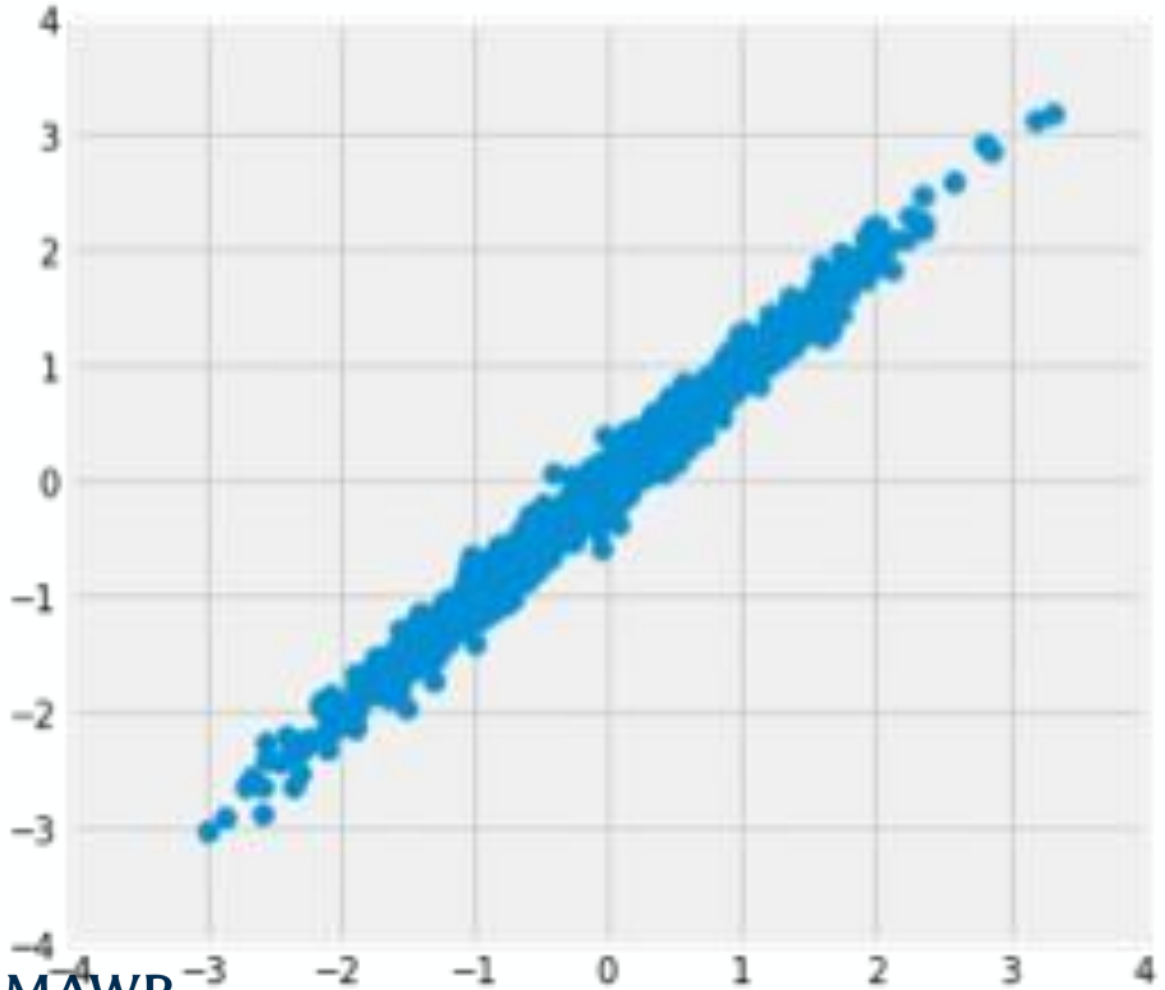
Lines can be expressed by *slope* & *intercept*

$$y = slope \times x + intercept$$

What we want

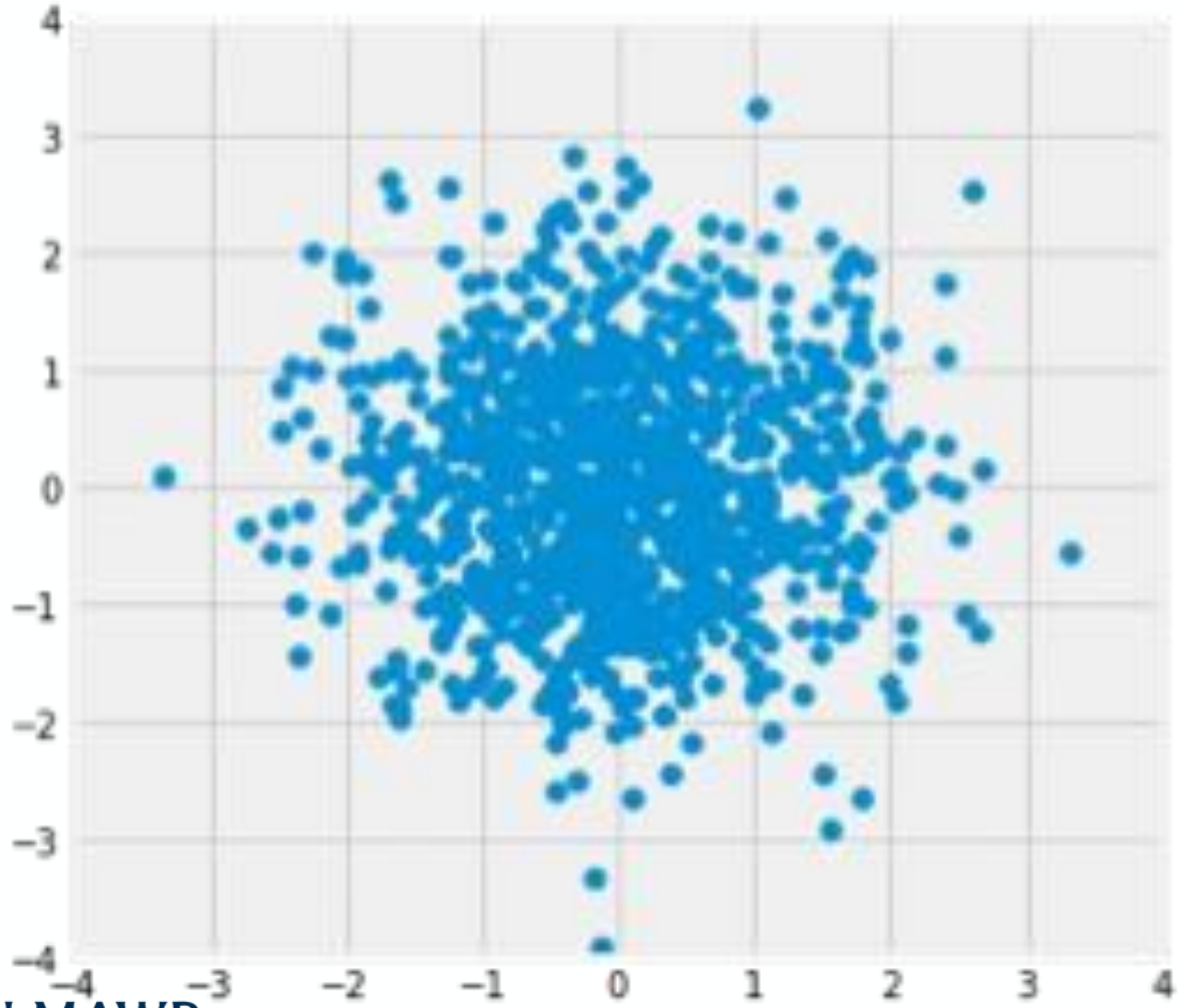
What we observe

Where is the prediction line?



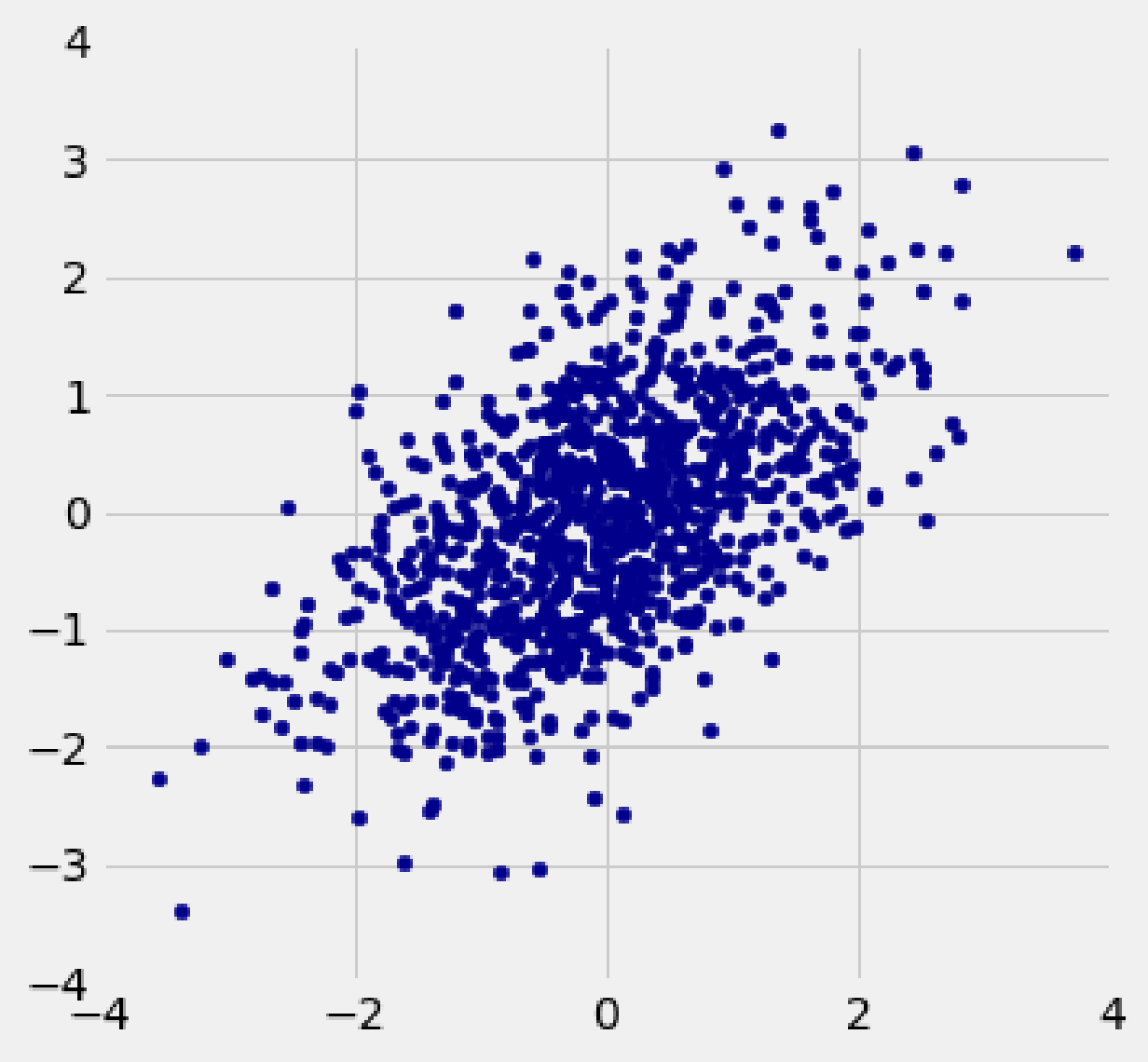
$r = 0.99$

Where is the prediction line?



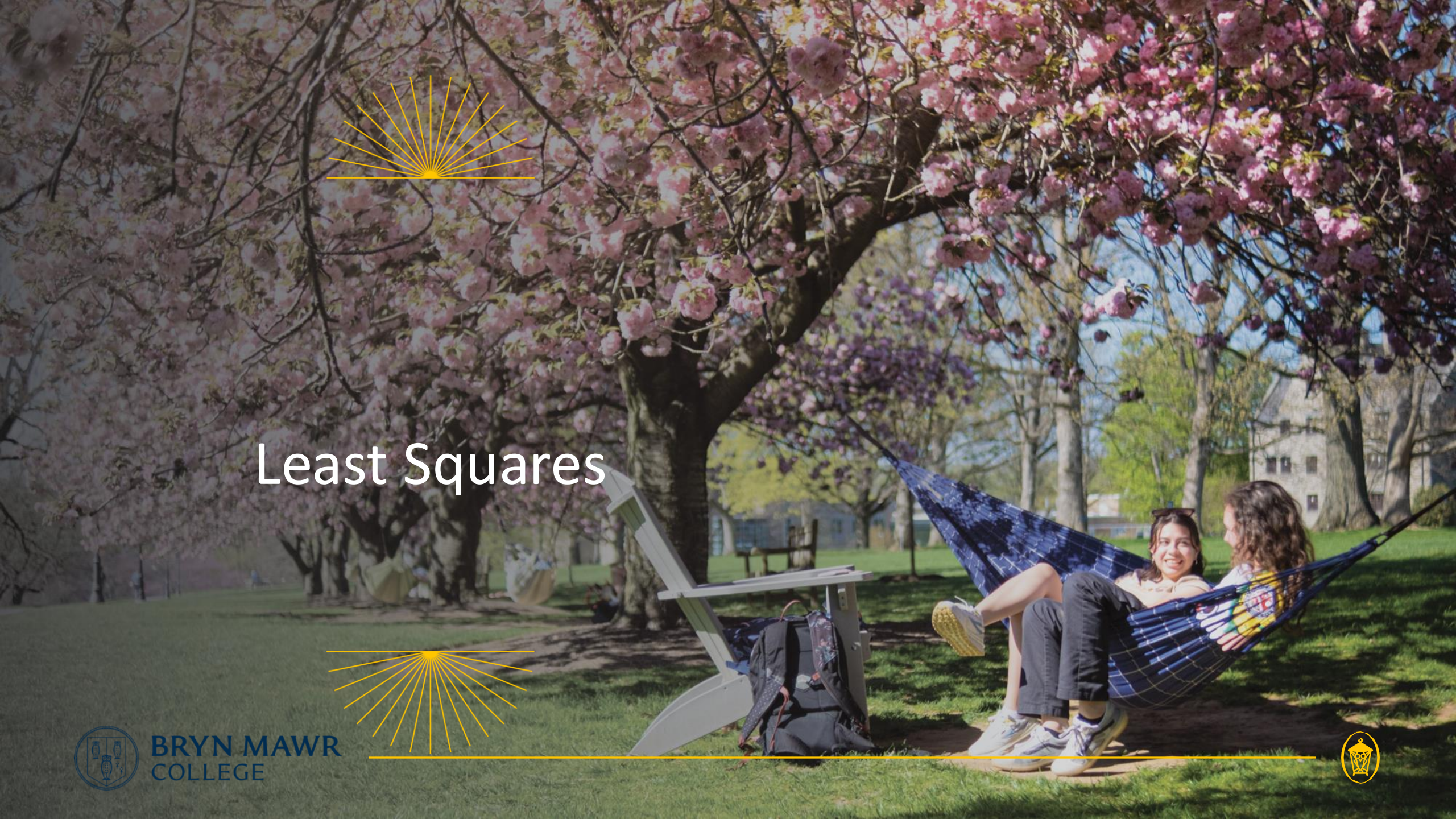
$r = 0$

Where is the prediction line?



$r = 0.5$

Least Squares



BRYN MAWR COLLEGE

Error in Estimation

error = actual value – estimate

Typically, some errors are positive and some are negative

To measure the rough size of the errors

- **square** the **errors** to eliminate cancellation
- Take the **mean** of the squared errors
- Take the square **root** to fix the units

Root mean square error (rmse)

Least Squares Line

Minimized the root mean squared error among all lines

Equivalently, minimizes the mean squared error among all lines

Names:

- “Best fit” line
- Least squares line
- Regression line

Numerical Optimization

Numerical minimization is approximate but effective

Lots of machine learning uses numerical minimization (demo)

If the function **mse(a, b)** returns the mse of estimation using the line "estimate = a*x* + b",

- then **minimize(mse)**returns array [a0, b0]
- a0 is the slope and b0 the intercept of the line that *minimizes* the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

Residuals

Error in regression estimate

One residual corresponding to each point (x, y)

residual
= observed *y* - regression estimate of *y*
= observed y - height of regression line at x
= vertical distance between the point and the best line

Residual Plot

A scatter diagram of residuals

Should look like an unassociated blob for linear relations

But will show patterns for non-linear relations

Used to check whether linear regression is appropriate

Look for curves, trends, changes in spread, outliers, or any other patterns

Properties of residuals

Residuals from a linear regression **always** have

- Zero mean
  - (so rmse = SD of residuals)
- **Zero** correlation with x
- **Zero** correlation with the fitted values

These are all true **no matter what the data look like**

- Just like deviations from mean are zero on average