Regression: the line

Topic 12 · Thursday, March 4

EDUC 15: Statistics for Education Research
UCI School of Education

Where are we?

Description

Summarize data

Visualize data

Prediction

Does one score go with another?

Use one score to predict another

Explanation

We are here



Peter the Anteater at a UC Irvine basketball game.

Plan for today

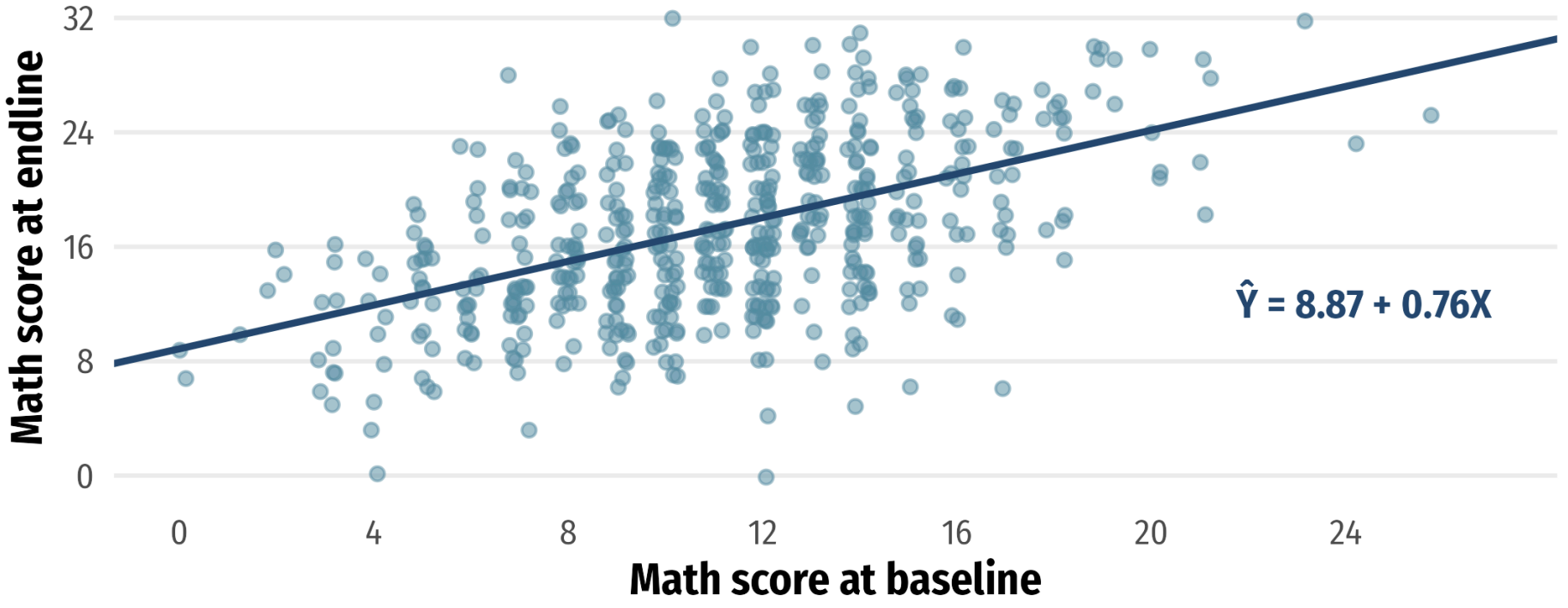
A line for prediction

What the line minimizes

Finding a and b

A line for prediction

A line sums up the cloud



The line shows the pattern, describes it in two numbers, and lets us predict

Data: Muralidharan, Singh, and Ganimian (2019). 539 students. Dots are nudged so equal scores don't hide each other.

A line needs two numbers

$$\hat{Y} = a + bX$$

$\hat{Y}$

The predicted Y. The hat says it's an estimate

a, the intercept

The value of $\hat{Y}$ when X = 0

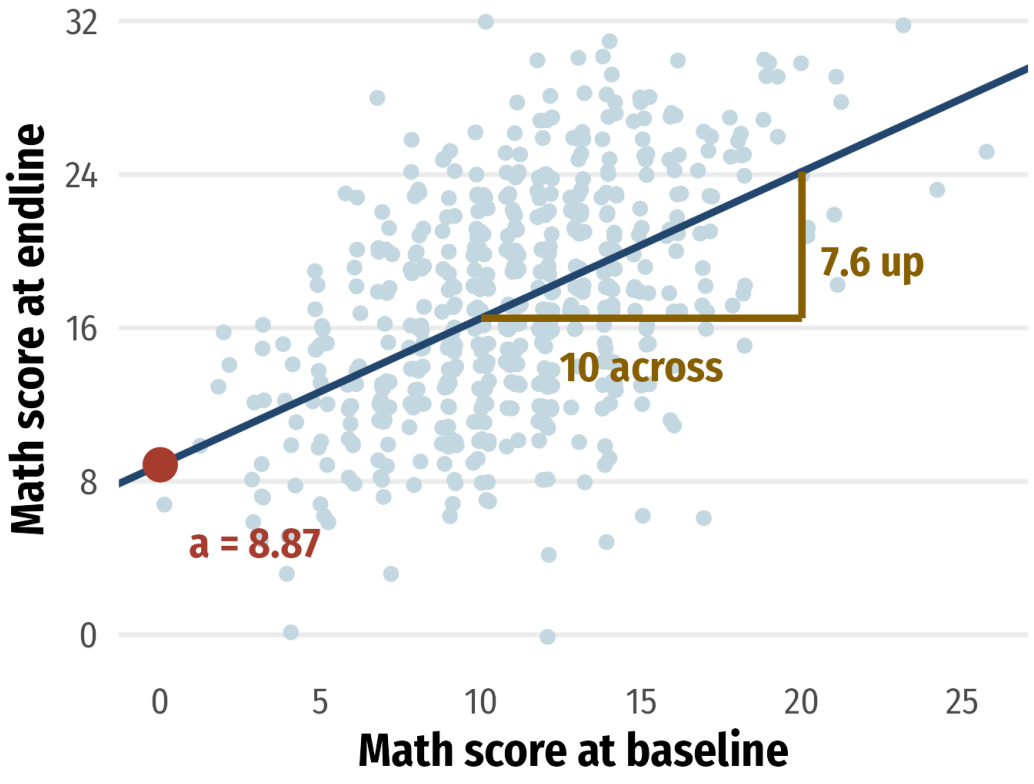
b, the slope

How much $\hat{Y}$ changes when X goes up by 1

Greek means population: β is the slope for everyone, and b is our estimate of it

In the book: $\hat{Y} = bX + a$. It's the same line, in a different order

Our line, read in points



$\hat{Y} = 8.87 + 0.76X$

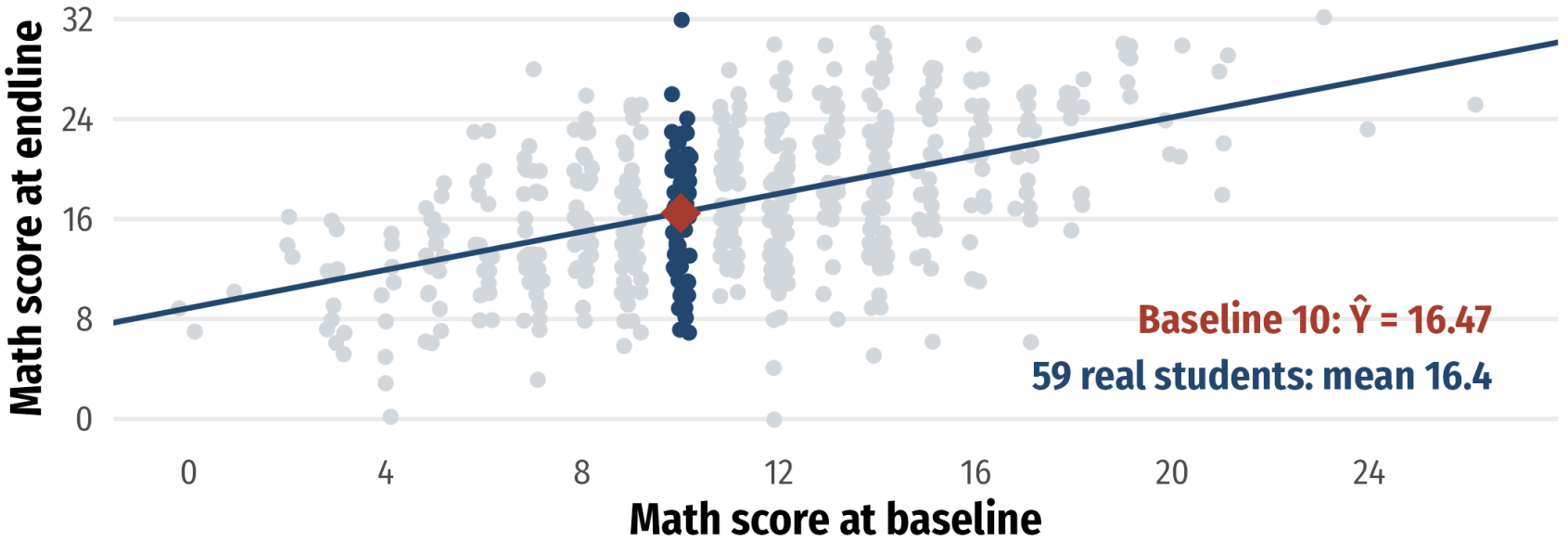
a = 8.87: a student with 0 at baseline is predicted to score 8.87 at endline

b = 0.76: one more point at baseline goes with 0.76 more points at endline, on average

Say "goes with", not "causes"

Data: Muralidharan, Singh, and Ganimian (2019).

It predicts an average, not a student



8.87 + 0.76 × 10 = 16.47. Yet these 59 students ended anywhere from 7 to 32

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Use the line

$$\hat{Y} = 8.87 + 0.76X$$

Two students scored 6 and 16 at baseline. Predict each one's endline score

How far apart are the students at baseline? How far apart are the predictions?

Work alone, compare with a neighbor, then vote with hands: is the gap bigger, smaller or the same?

Try it: Answer

X = 6: 8.87 + 0.76 × 6 = **13.43**

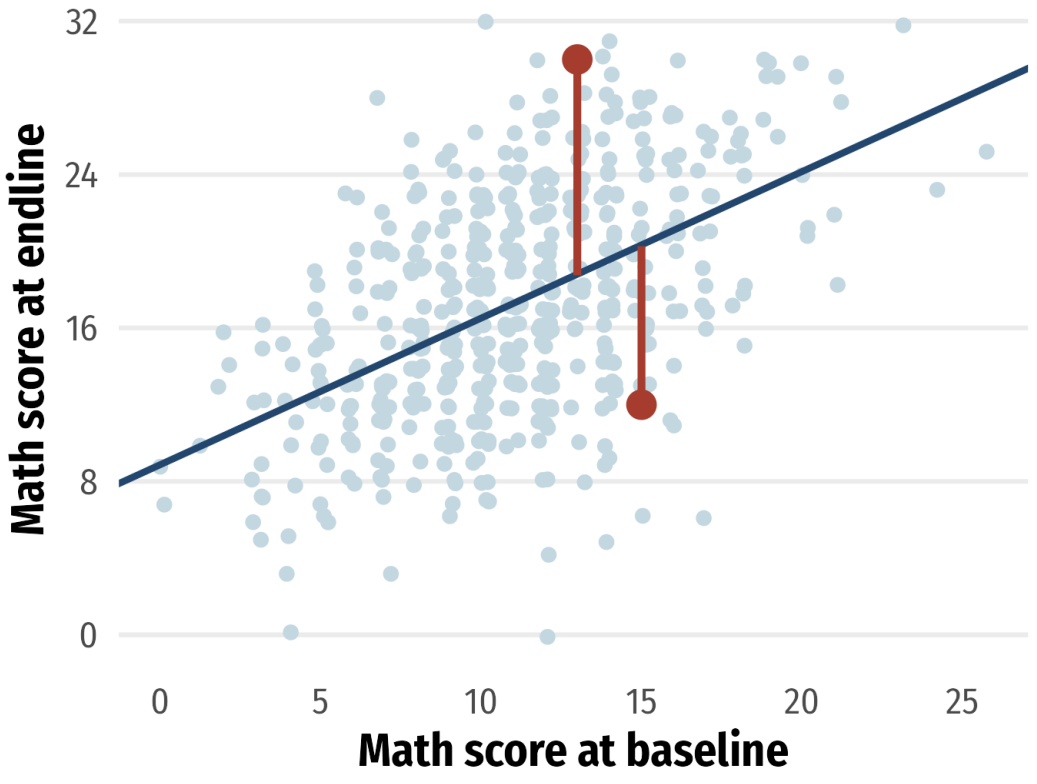
X = 16: 8.87 + 0.76 × 16 = **21.03**

10 points apart at baseline, 7.60 apart in prediction. The gap is 10 × b

The slope is all you need to compare two students. a only sets the height

What the line minimizes

A residual is how far the line missed



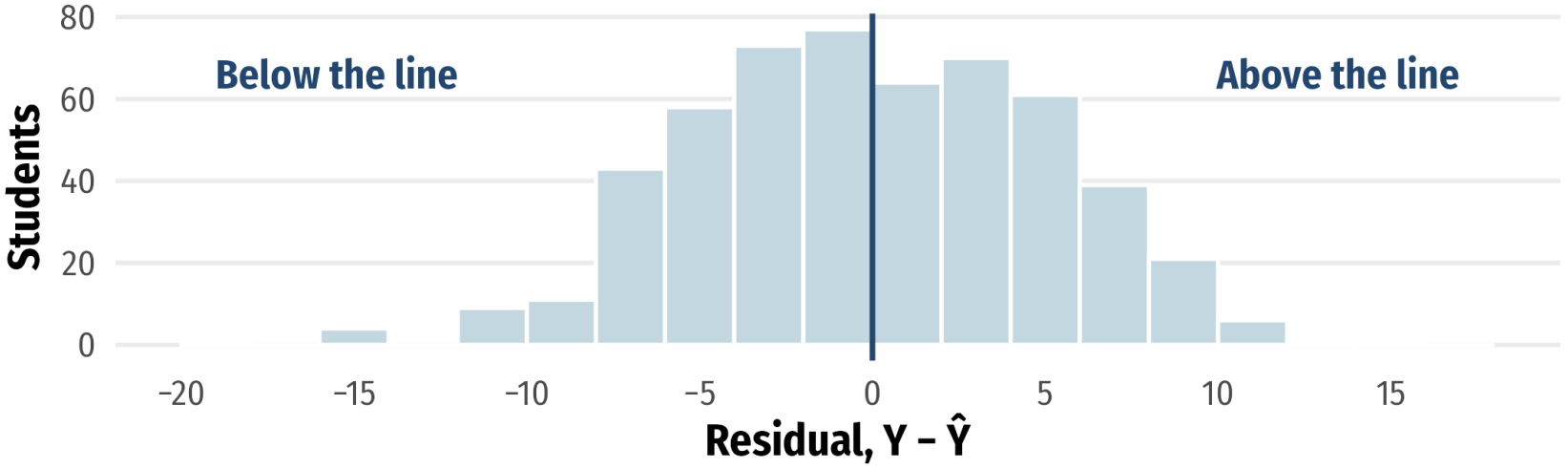
Residual = Y – Ŷ, what happened minus what we predicted

X	Y	Ŷ	Y – Ŷ
13	30	18.75	+11.25
15	12	20.27	−8.27

Above the line, it's positive. Below the line, it's negative

Data: Muralidharan, Singh, and Ganimian (2019).

Residuals add to 0, so we square them

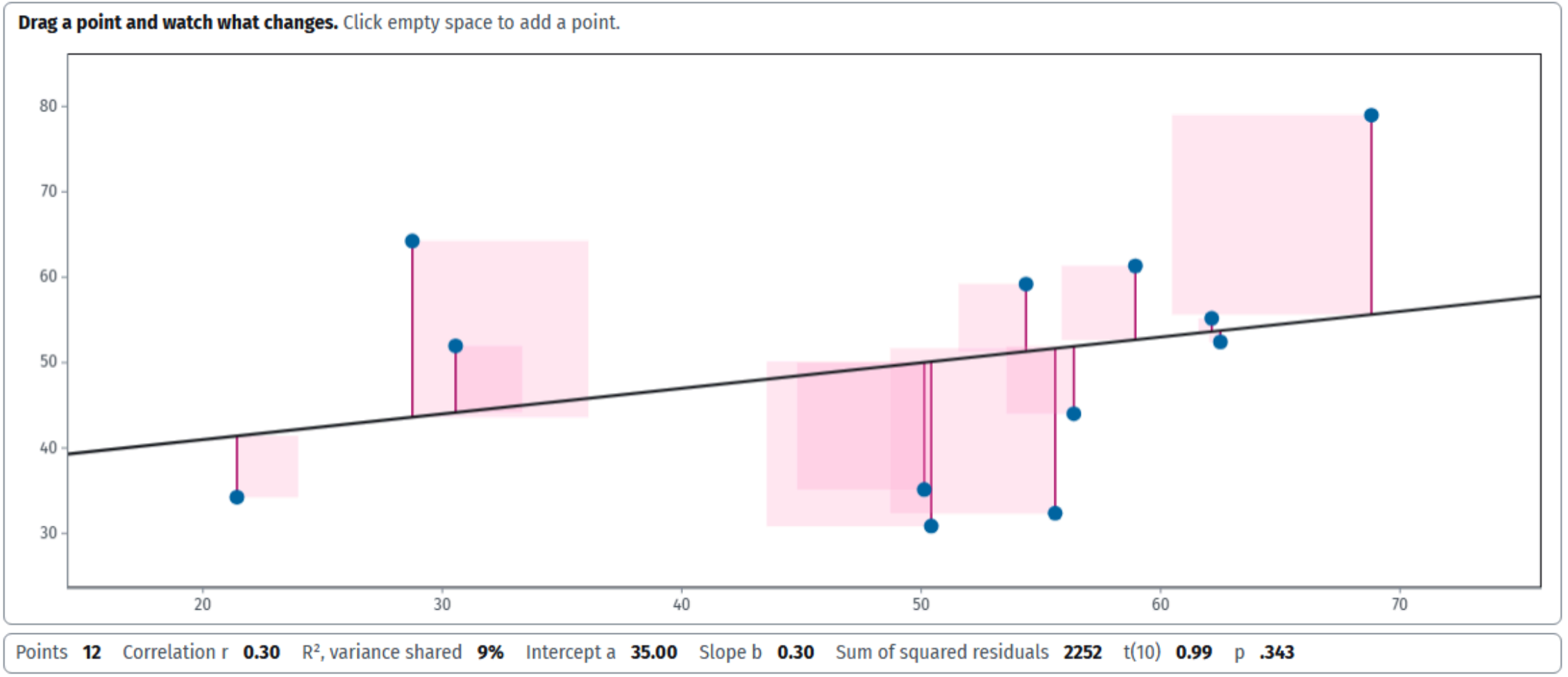


All 539 residuals add up to 0, like deviations from a mean

So we square them, as in Topic 3: $\Sigma(Y - \hat{Y})^2 = 14{,}314$, the $SS_{residual}$

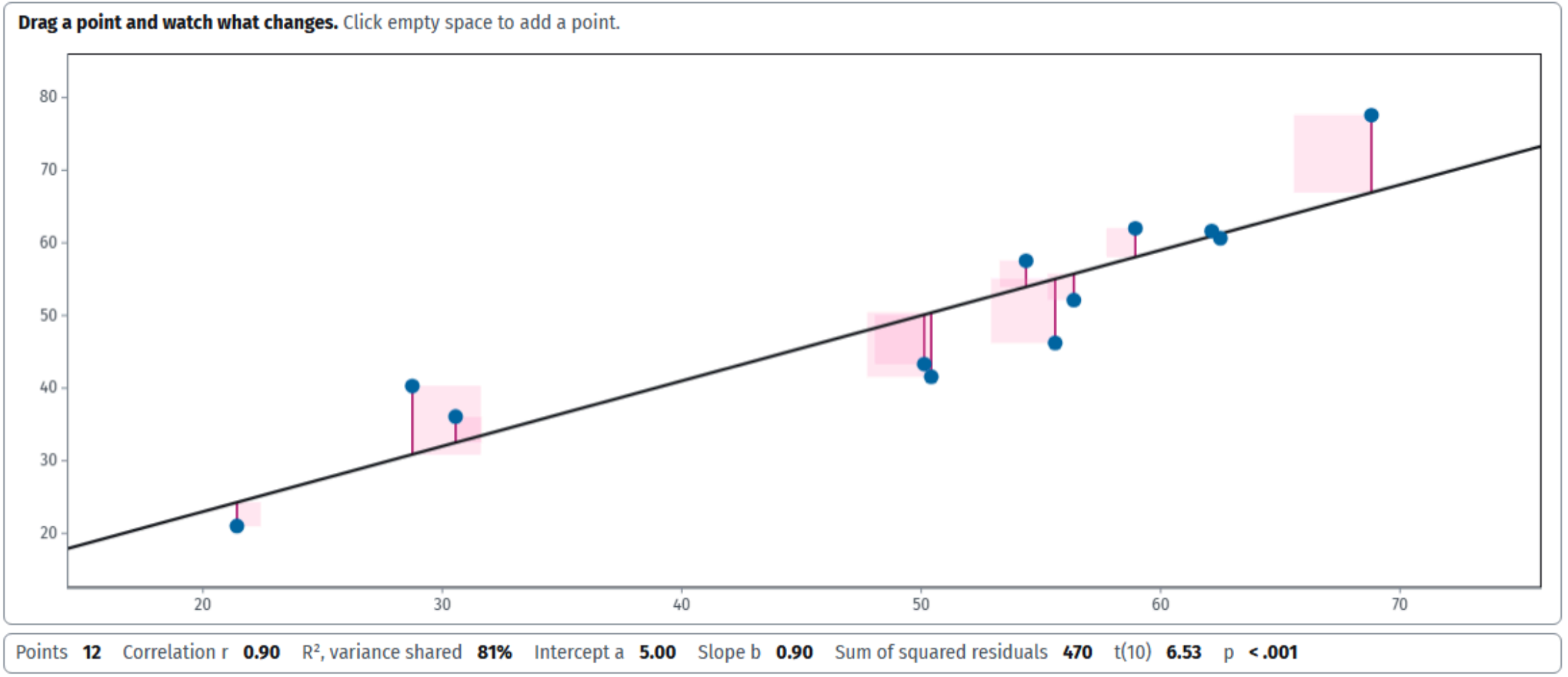
Data: Muralidharan, Singh, and Ganimian (2019).

Loose dots leave big squares



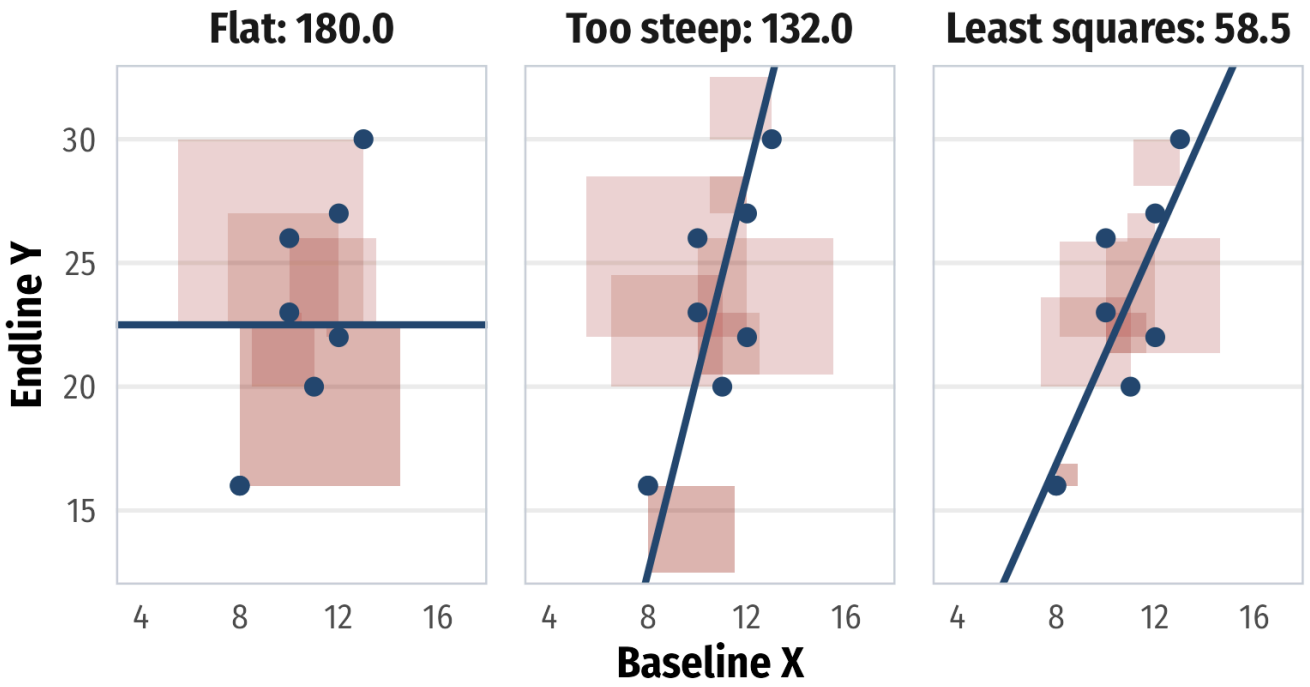
Simulator after Kristoffer Magnusson, rpsychologist.com. On your phone: the simulator page.

Tight dots leave small squares



Simulator after Kristoffer Magnusson, rpsychologist.com.

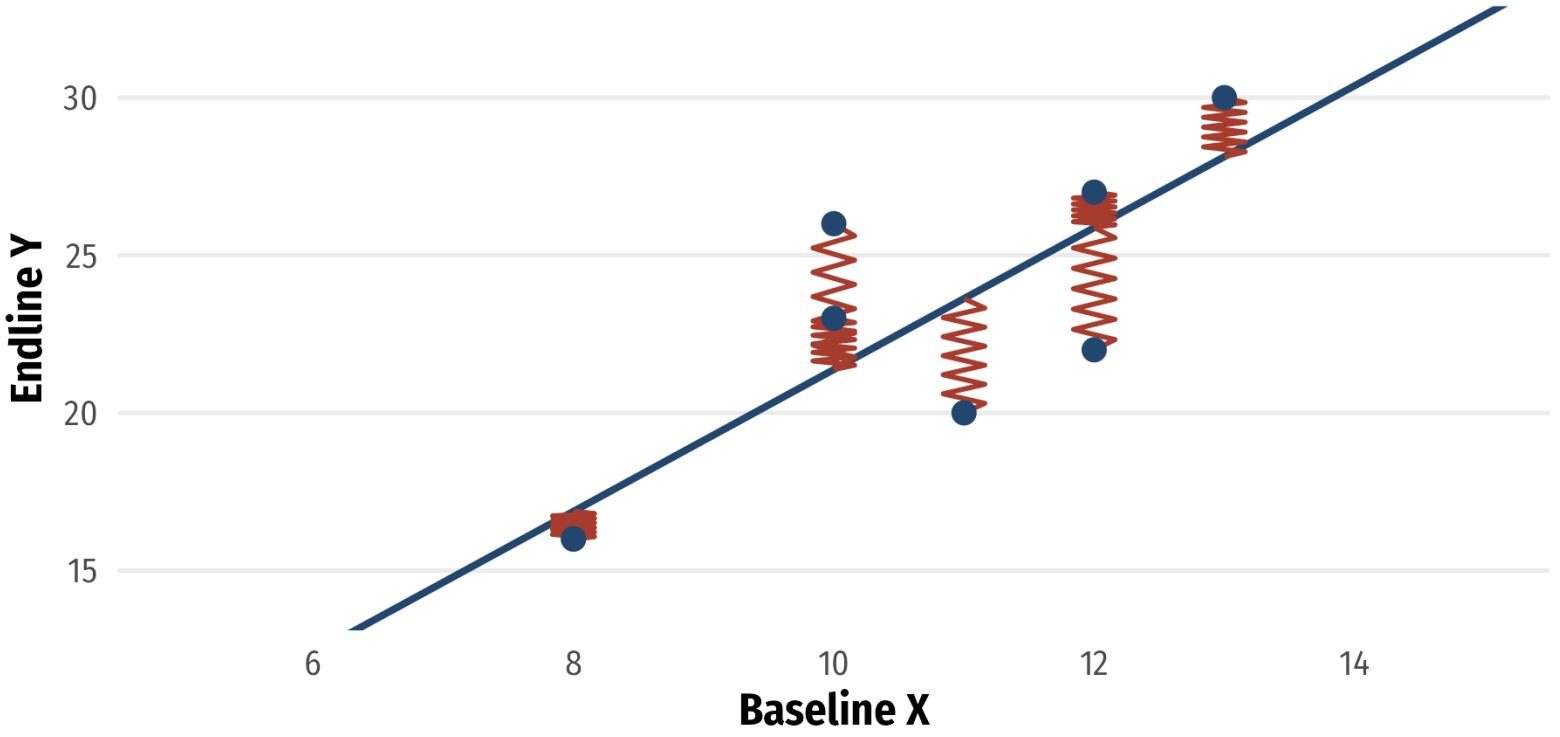
Least squares picks the smallest total



Regression chooses the a and b that make $\Sigma(Y - \hat{Y})^2$ as small as it can be

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32 from Topic 11.

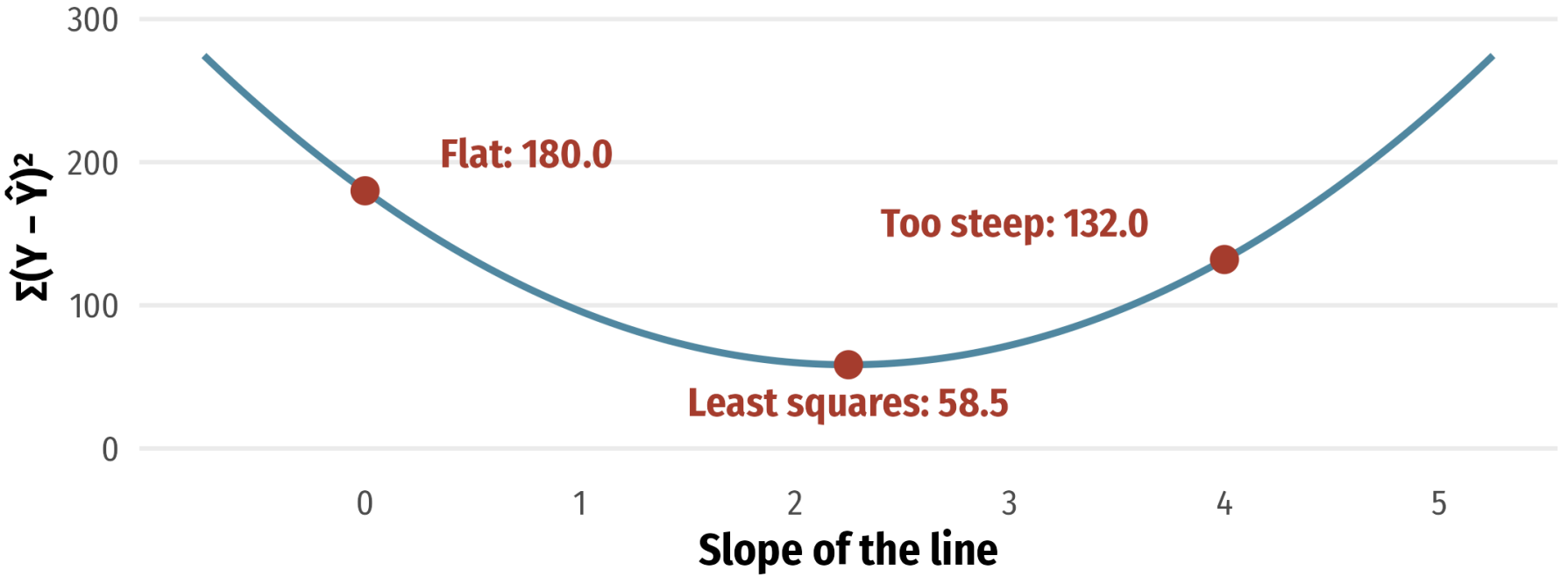
Think of the line as held by springs



Every student's spring pulls on the line. Least squares is where they balance

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32 from Topic 11.

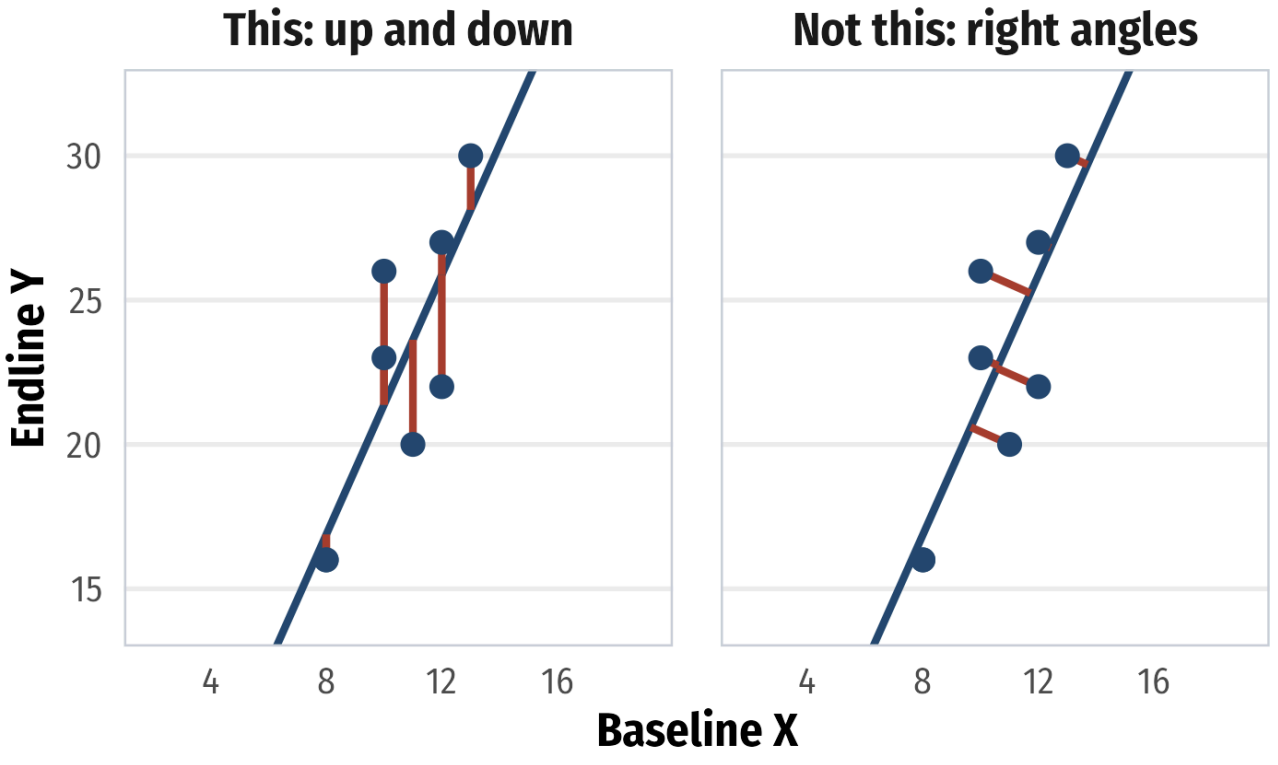
The best slope sits at the bottom



Every slope has a total. b = 2.25 is the only one at the bottom of the curve

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32. Each line goes through the two means.

We measure misses up and down



We're predicting Y, so a miss is a distance in Y

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32.

Predict it: Which line wins?

Respond at **PollEv.com/adb**

Line A misses four students by 0, 0, 0 and 8 points. Line B misses them by 2, 2, 2 and 2. Each line misses by 8 points in all. Which one does least squares prefer?

A Line A

B Line B

C It's a tie

D We can't tell without the slope

Finding a and b

We already have what we need

From Topic 11, our 8 students: SP = 54, SS_X = 24, M_X = 10.5, M_Y = 22.5

$$b = \frac{SP}{SS_X} = \frac{54}{24} = 2.25$$

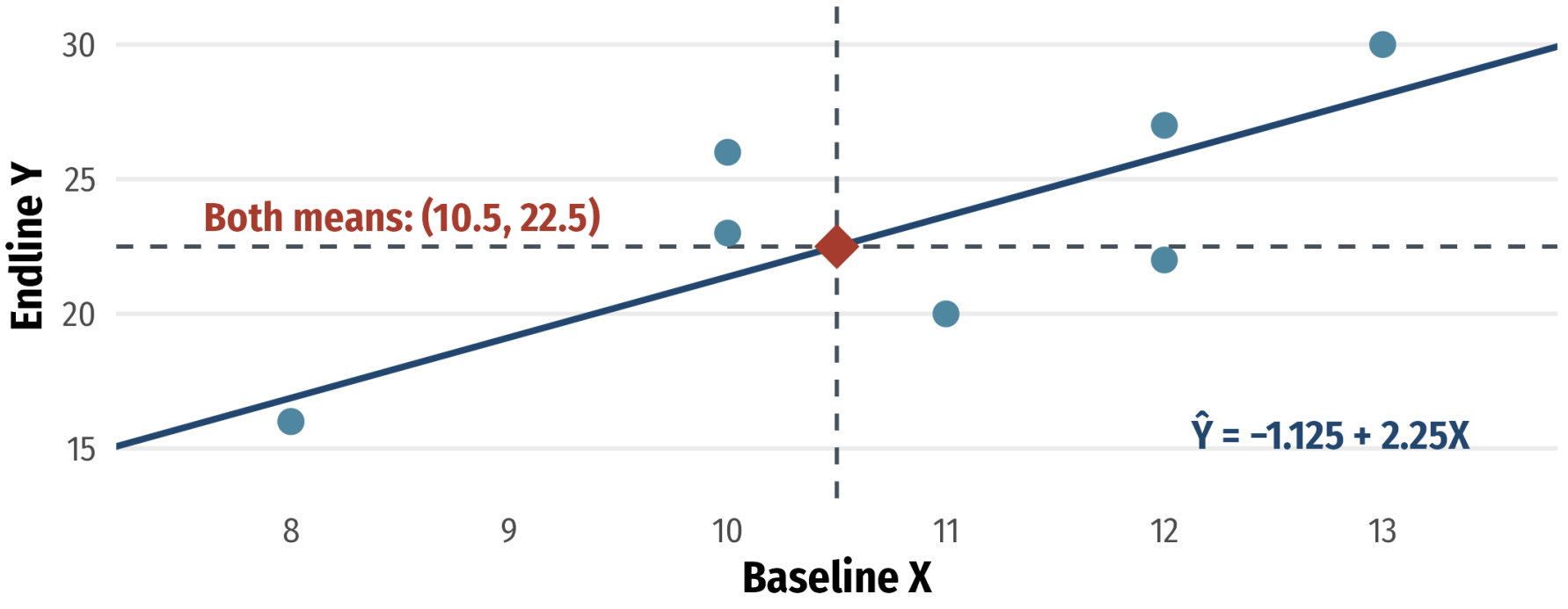
$$a = M_Y - bM_X = 22.5 - 2.25 \times 10.5 = -1.125$$

SP: how X and Y move together. SS_X: how X moves alone

Their line: $\hat{Y} = -1.125 + 2.25X$

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32.

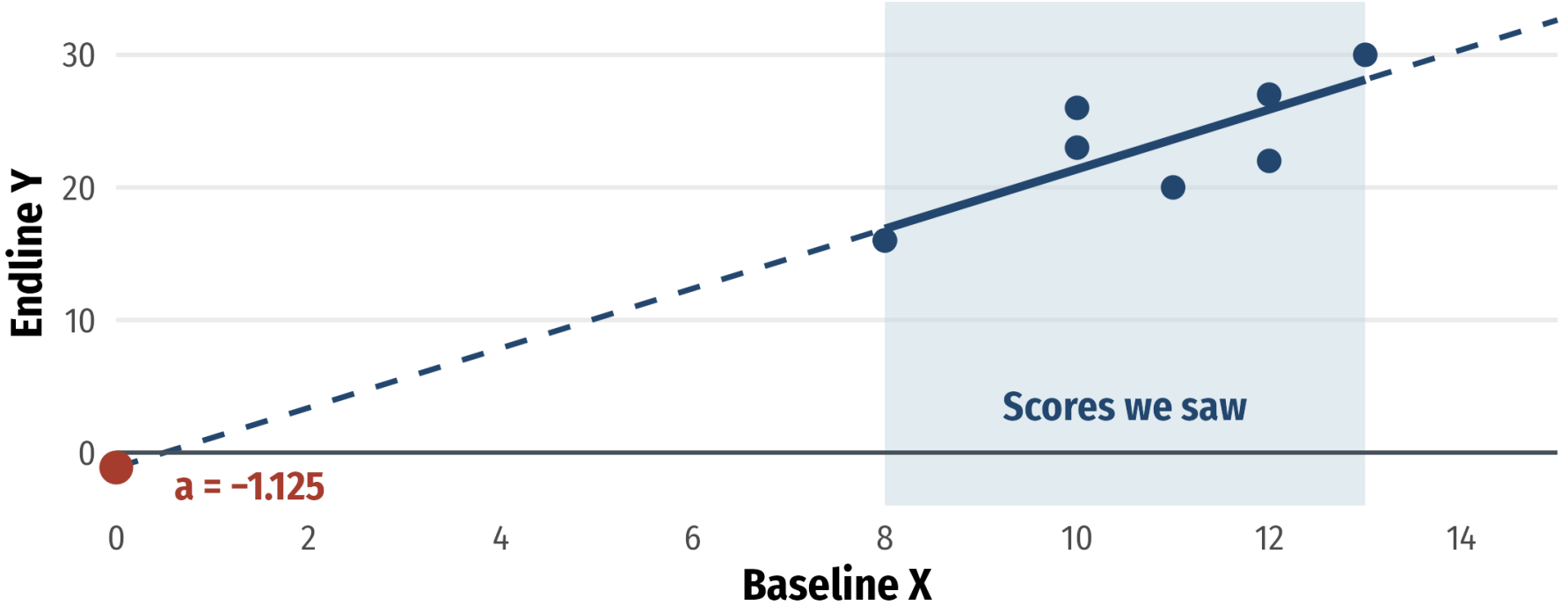
The line runs through the two means



A student at M_X is predicted to land at M_Y. That's where a's formula comes from

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32.

Don't trust the line past the data



A score below 0 is impossible. Predict only inside the scores you saw

Data: Muralidharan, Singh, and Ganimian (2019). Students 25 to 32.

b is r, rescaled into points

$$b = r \times \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$$

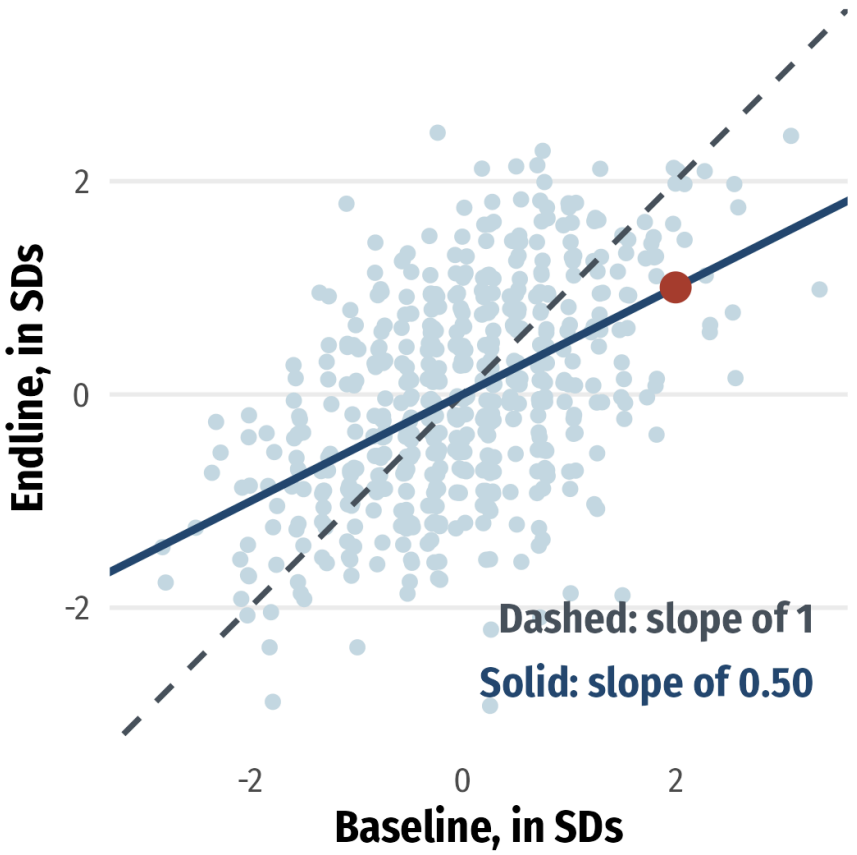
Students	r	$\hat{\sigma}_Y$	$\hat{\sigma}_X$	$r \times \hat{\sigma}_Y / \hat{\sigma}_X$
All 539	0.502	5.97	3.92	**0.76**
Our 8	0.822	5.07	1.85	**2.25**

r has no units. b is in points of Y per point of X. They always share a sign

In the book: $\hat{\sigma}_Y$ and $\hat{\sigma}_X$ are written s_Y and s_X

Data: Muralidharan, Singh, and Ganimian (2019).

Regression to the mean is in the slope



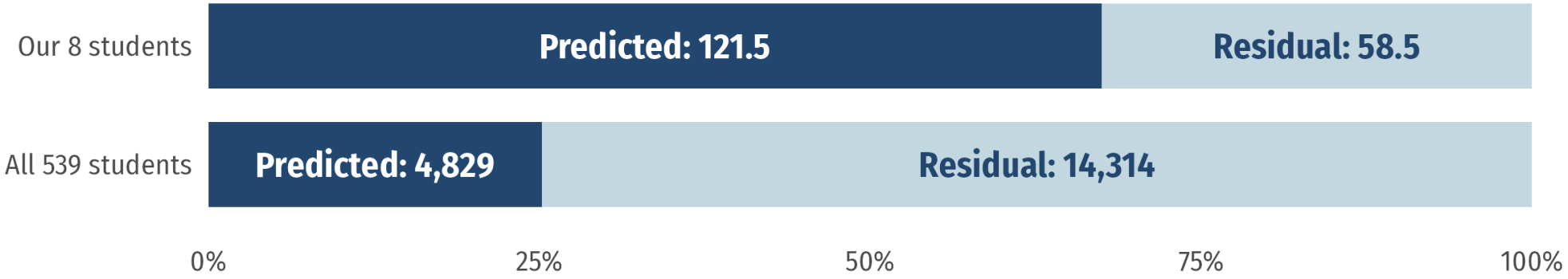
In SDs, the slope is r, and r is below 1 unless the dots fall on a perfect line

So a student 2 SDs above the mean is predicted only 1.0 SD above it

The 42 students with 17 or more at baseline were 1.96 SDs above the mean. At endline: 0.98

Data: Muralidharan, Singh, and Ganimian (2019).

r^2 is how much the squares shrink



$$r^2 = \frac{SS_Y - SS_{residual}}{SS_Y} = \frac{180 - 58.5}{180} = 0.675$$

A flat line ignores X, and its squares total SS_Y. For our 8, the best line cuts that by 67.5%

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Five students from Topic 11

Baseline X	Endline Y
13	14
10	21
11	24
11	14
10	22

$M_X = 11, M_Y = 19$

$SP = -15, SS_X = 6$

Find b and a. Then find the residual for the student with X = 11 and Y = 24

Work alone, compare with a neighbor, then vote with hands: is the residual positive or negative?

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Answer

$b = -15 / 6 = \mathbf{-2.5}$

$\hat{Y} = 46.5 + (-2.5 \times 11) = \mathbf{19}$

$a = 19 - (-2.5 \times 11) = \mathbf{46.5}$

$Y - \hat{Y} = 24 - 19 = \mathbf{+5}$

r was negative here, so b is too. And X = 11 is M_X, so $\hat{Y}$ is M_Y: no multiplying needed

Data: Muralidharan, Singh, and Ganimian (2019).

Each b estimates β

Students	n	r	a	b
Five from the try-it	5	−0.65	46.50	−2.50
Students 25 to 32	8	0.82	−1.125	2.25
Everyone with both scores	539	0.50	8.87	0.76

Same test, three groups, three slopes. Like M and r, b bounces from sample to sample

Tuesday's question: could we see a slope like this if β were 0?

Data: Muralidharan, Singh, and Ganimian (2019).

What you can do now

- Read a and b from a line, in the units of the scores
- Predict Y from X, and say why that's an average, not a promise
- Find a residual, and say what least squares minimizes
- Compute b and a by hand from SP, SS_X and the means
- Link b to r, r^2 to the squares, and r to regression to the mean

Next steps

Friday, March 5: Lab 9

Correlation and regression.
Bring a calculator and a laptop

Sunday, March 7, 6 PM

Mastery and Problem Set for Chapter 14 are due

For Tuesday, March 9: read the rest of section 14-6, for Topic 13