

Psych 10: Intro to Statistics

Lecture 11 · Regression and correlation · Wed Oct 28, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

- CourseKata **Chapter 9** is due Friday.
- Section this week: **Milestone 5**, the preregistration, due Sunday, November 1 on Gradescope.
- Friday: **Case #5**, in your tutorial session. Check your discussion section's Slack channel for your time slot. Bring a pen.

Looking ahead to the exam

- The **practice exam** is Monday, November 9, in class. The exam is Wednesday, November 18.
- I expect to be able to release **pre-practice exam study questions** early next week, likely Monday.
- We'll be holding a **review session** on Wed, November 11th. Check your section's Slack channel for a link to a Google Form where you can suggest topics to cover in that session; I'll be collecting suggestions through November 4th.

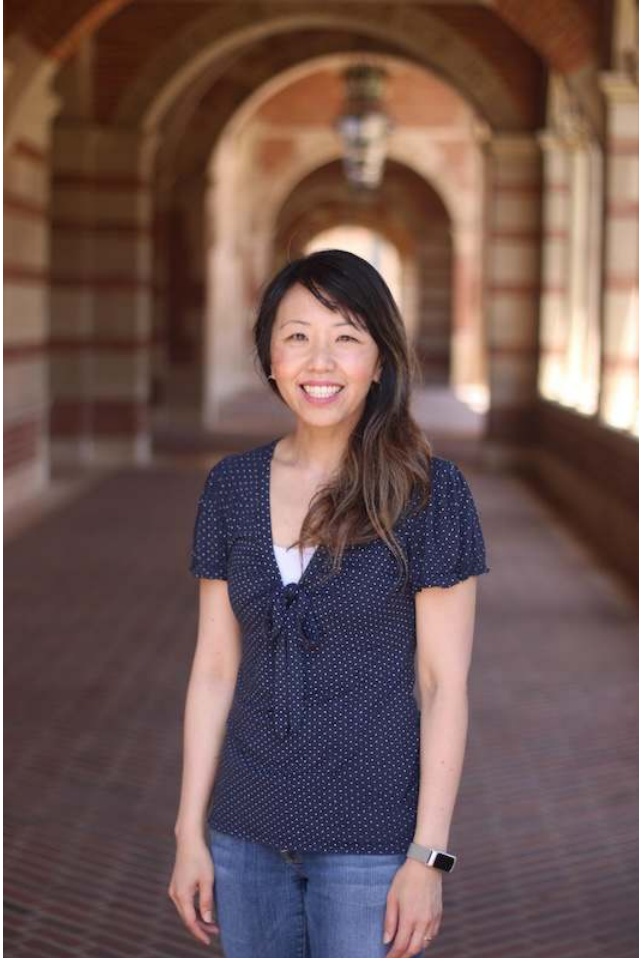
The mid-quarter survey

Mid-Quarter Survey.

Please set aside 10 minutes to complete the [mid-quarter survey](#). It asks how you are feeling, what you are learning, and anything else you would like us to know about your experience in the course so far, and it helps the teaching team shape the second half of the quarter. Good-faith completion of the mid-quarter survey makes you eligible to receive a full percentage point of extra credit that counts toward your final grade.

- The link is on the syllabus page, below the student background survey.
- It asks how you're feeling, what you're learning, and anything else you want us to know, and it helps us shape the second half of the quarter.
- A good-faith response makes you eligible for a full percentage point of extra credit toward your final grade.

Next Wednesday: a guest lecture by Ji Son



- Next Wednesday, November 4, we'll have a guest lecture from none other than the author of CourseKata, Ji Son.
- Ji Son is a Professor of Psychology at Cal State LA and co-author, with James Stigler, of the CourseKata textbook this course is built on.

Today

Regression and correlation



Guest lecture: tidying and
wrangling data, with Jane
Stephenson



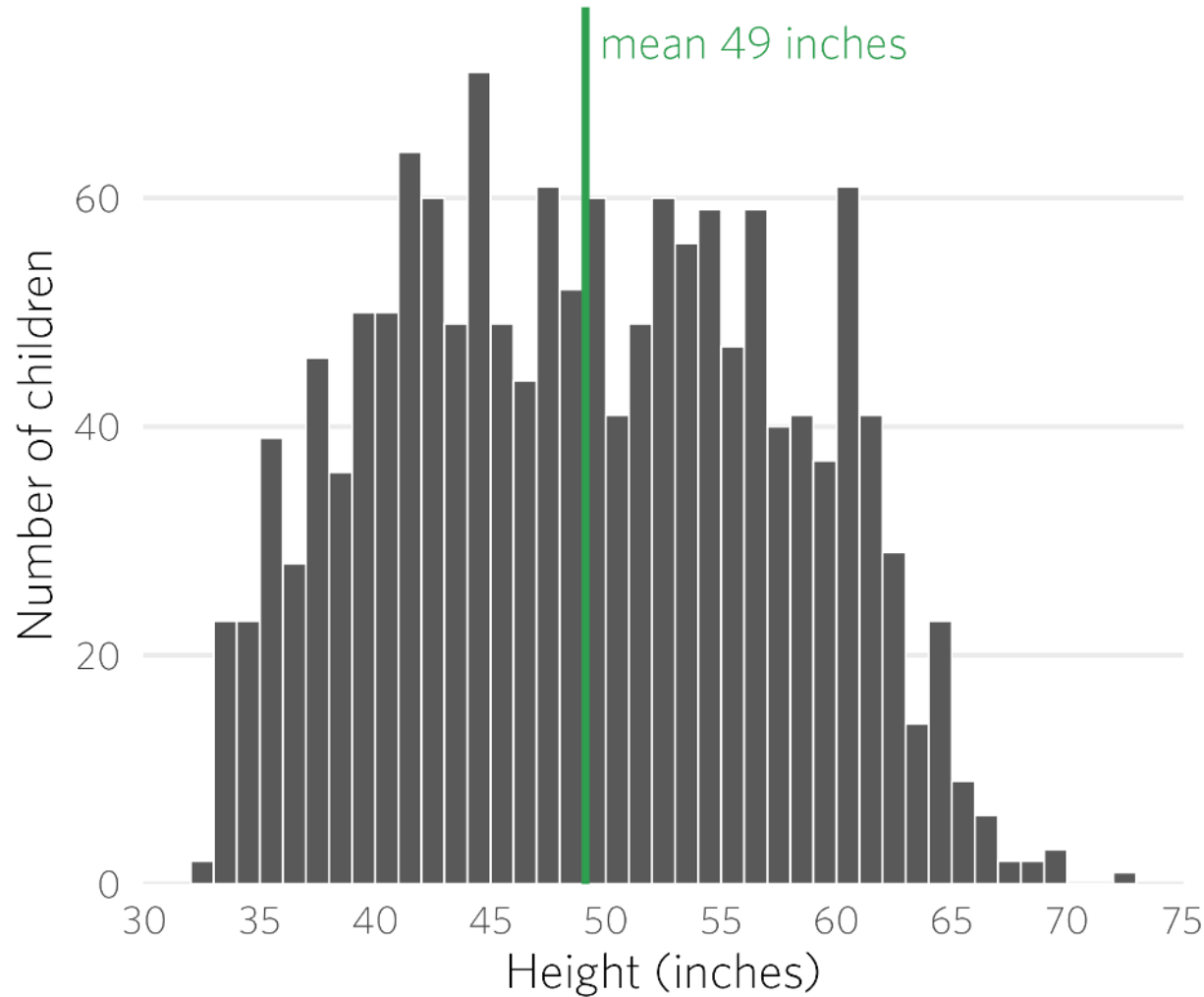
The slope and the intercept



How strong is the
relationship?

The heights of 1,487 children

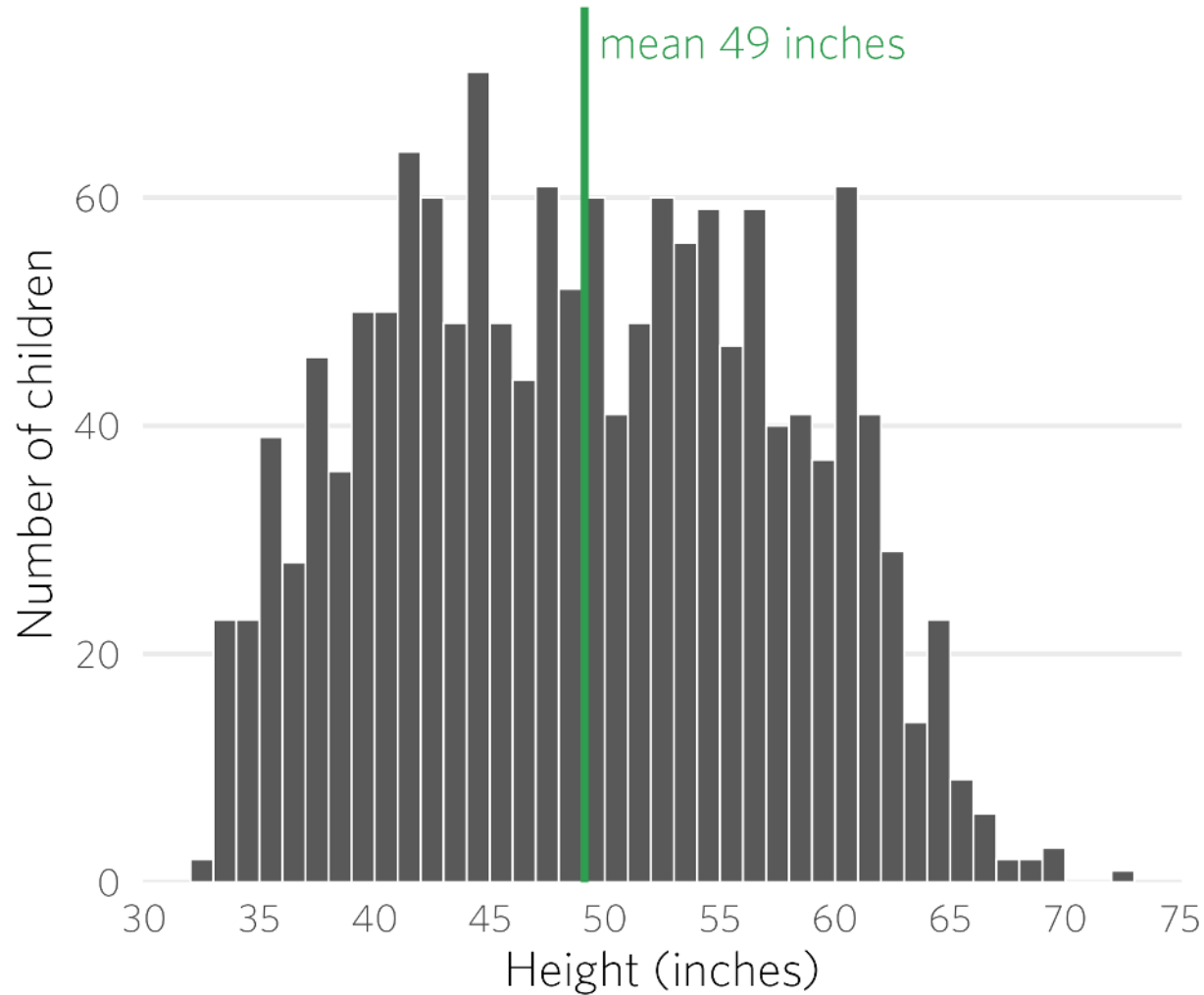
1,487 children aged 2 to 12 (NHANES); SD 8.5 inches



- The NHANES children from Lecture 5, restricted to the 1,487 aged 2 to 12.
- With nothing else known about a child, the mean, 49 inches, is the prediction for every child, as in Lecture 5.
- The SD is 8.5 inches: a typical child's height is about 8.5 inches from the mean.

Predict a child's height

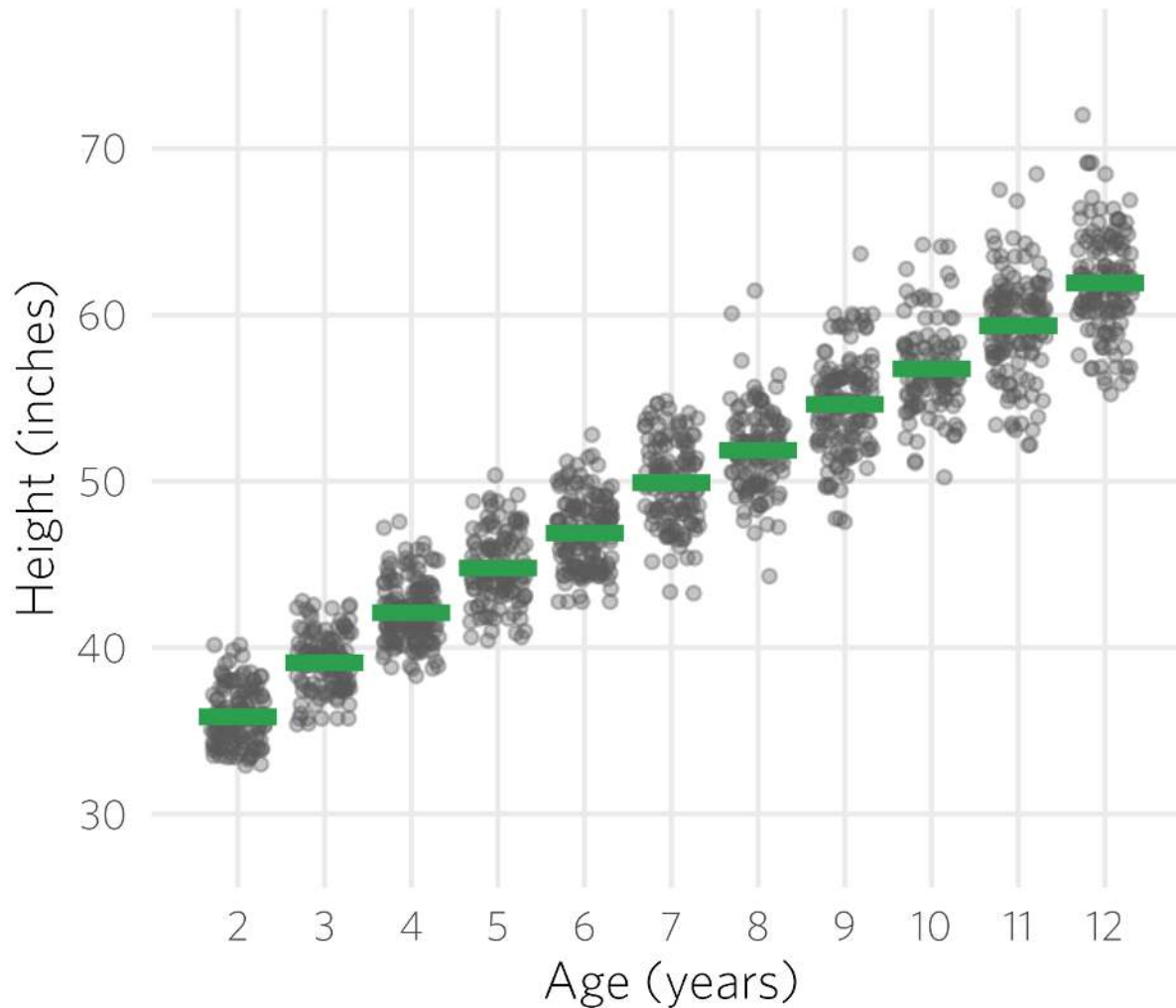
1,487 children aged 2 to 12 (NHANES); SD 8.5 inches



- A child is picked at random from these 1,487.
What height do you predict, in inches?
- Now you learn that the child is 10 years old.
What height do you predict for this child?

One mean per year of age

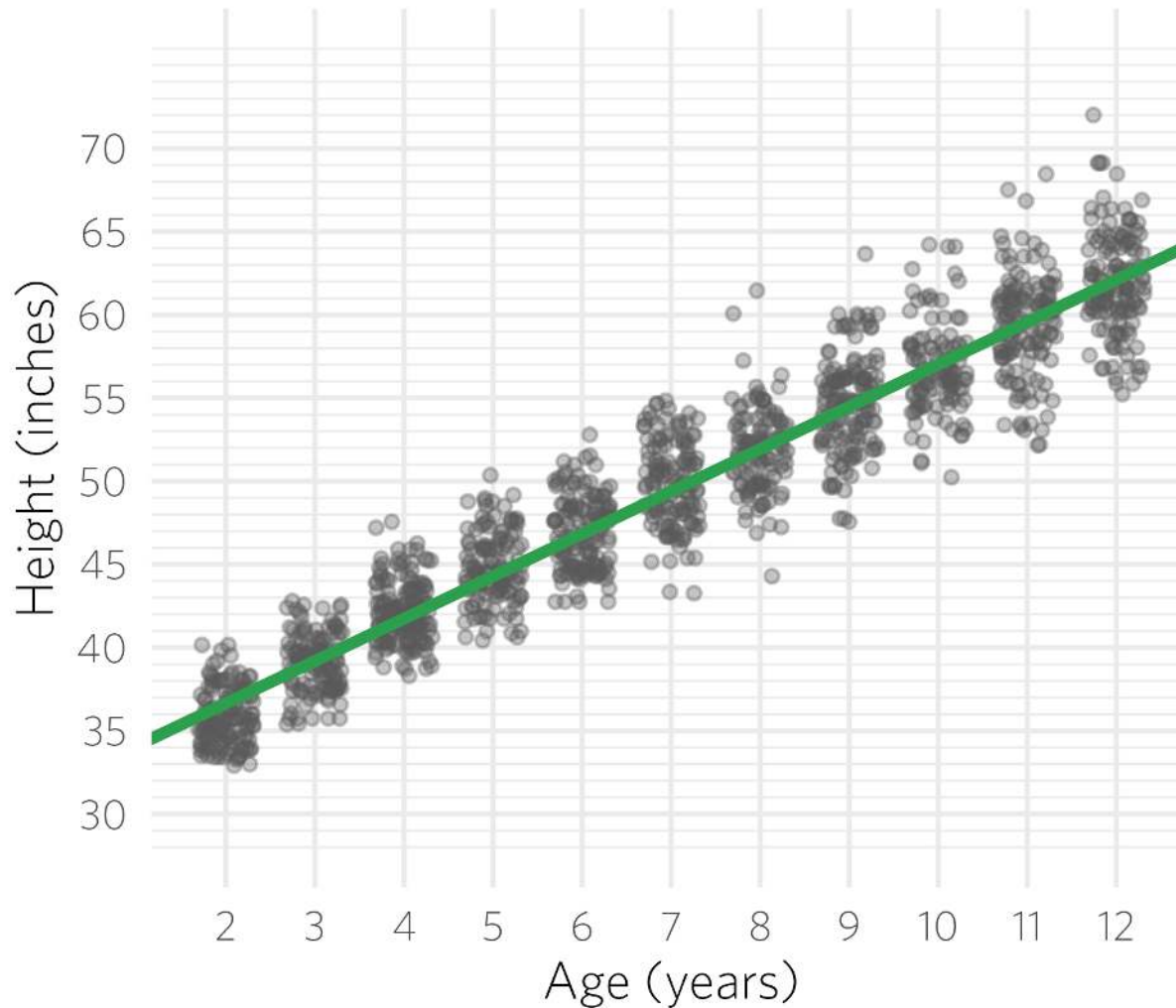
One mean per year of age: eleven means (PRE 0.92)



- Eleven means, ten parameters more than the empty model: PRE 0.92.
- The means rise by about 2.5 inches per year of age, close to a straight line.

A line in place of eleven means

The regression line of height on age (PRE 0.92)



- The line predicts height from age with two numbers: an intercept and a slope.
- Its PRE is 0.92, the same as eleven means, with nine fewer parameters.

$$Y_i = b_0 + b_1 X_i + e_i$$

- This is the same equation as last week's two-group model.
- There, X_i was 0 or 1. Here, X_i is the child's age in years.

The line as R reports it

```
lm(inches ~ Age, data = children)
```

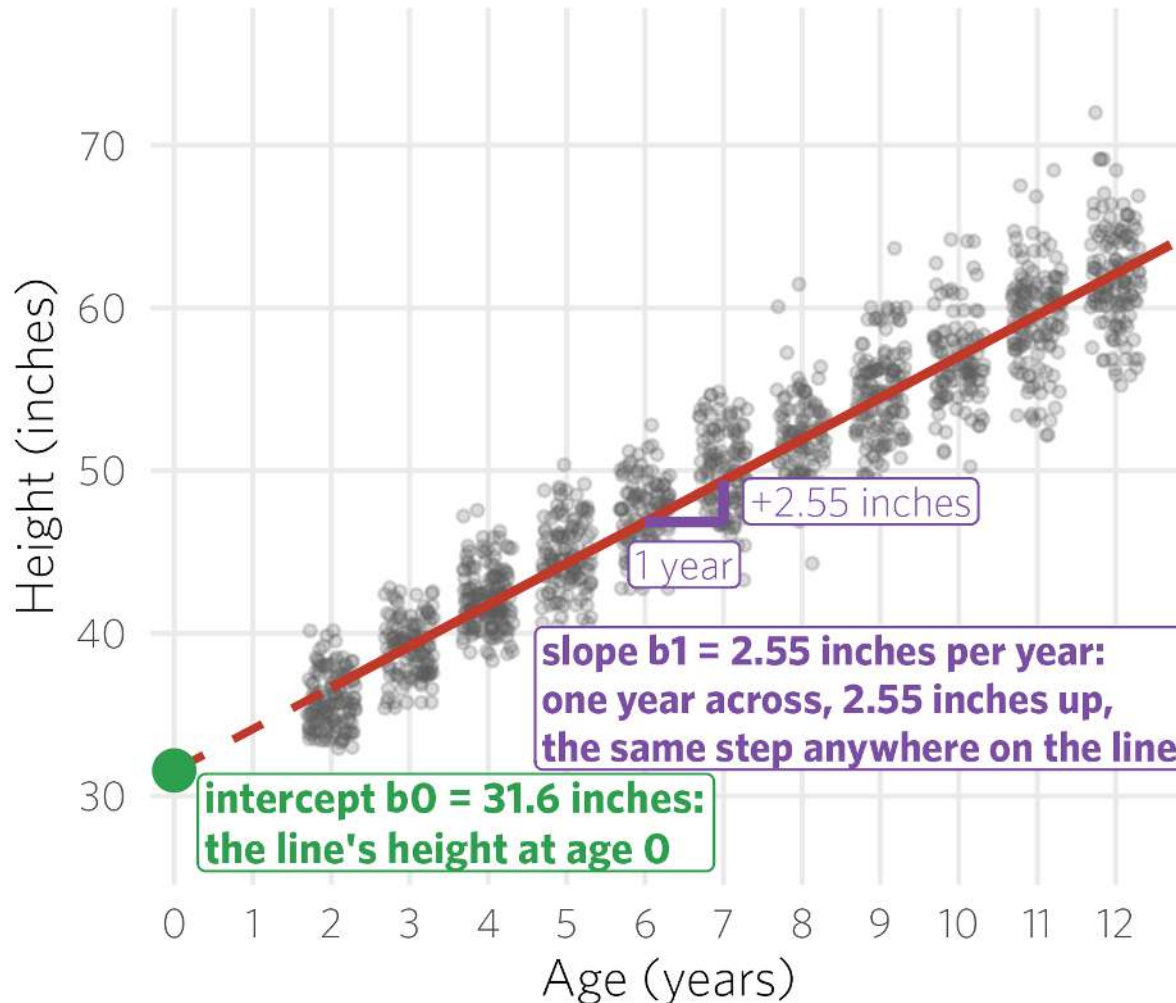
(Intercept)	Age
31.56	2.55

$$\text{height} = 31.6 + 2.55 \times \text{age}$$

- The slope: for each additional year of age, the predicted height is **2.55 inches** greater.
- Its units are inches per year: the outcome's units per unit of the explanatory variable.
- Last week, b_1 was the difference between two group means. Now it is the change in the predicted height per year of age.

The intercept and the slope, on the picture

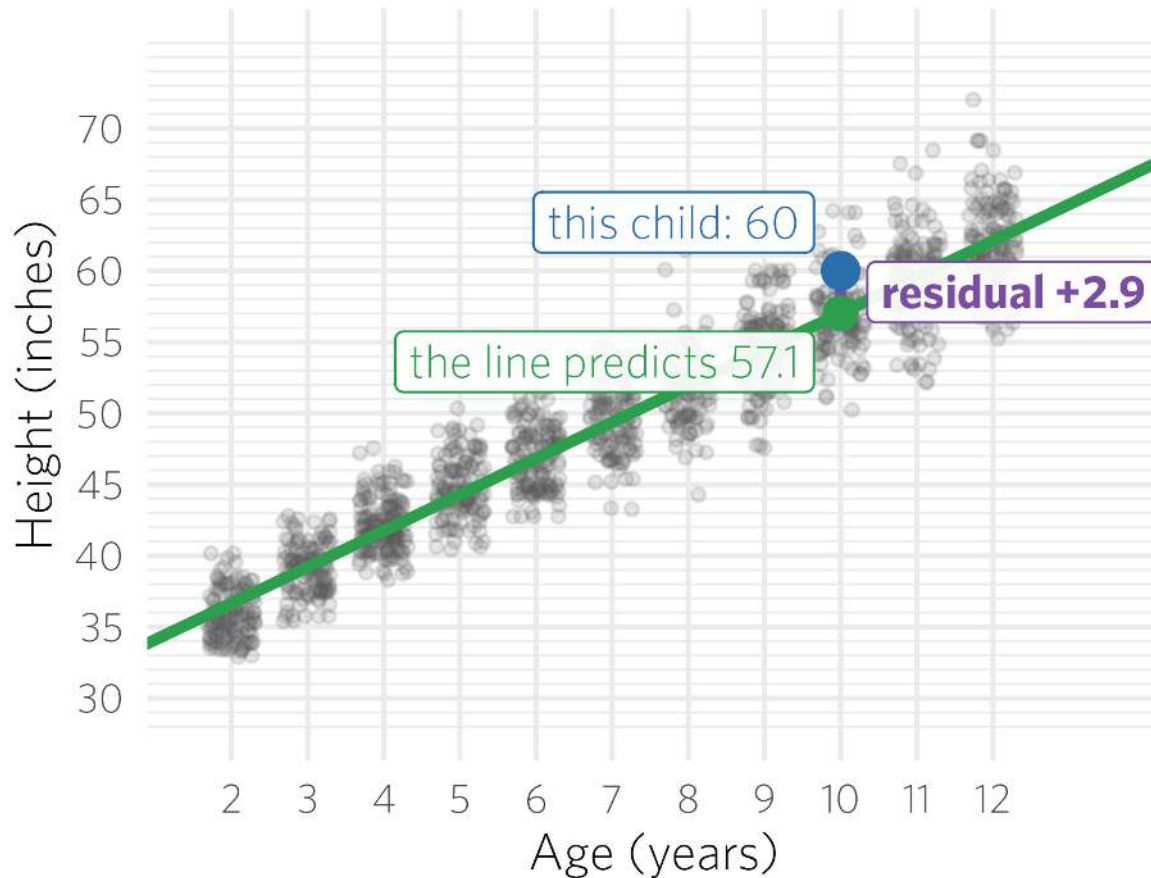
$$\text{height} = 31.6 + 2.55 \times \text{age}$$



- The fitted line, with the age axis extended to 0.
- The intercept, $b_0 = 31.6$: the line's height at age 0, where the line meets the vertical axis.
- The slope, $b_1 = 2.55$: one year across, 2.55 inches up, the same step anywhere on the line.

Predictions and residuals

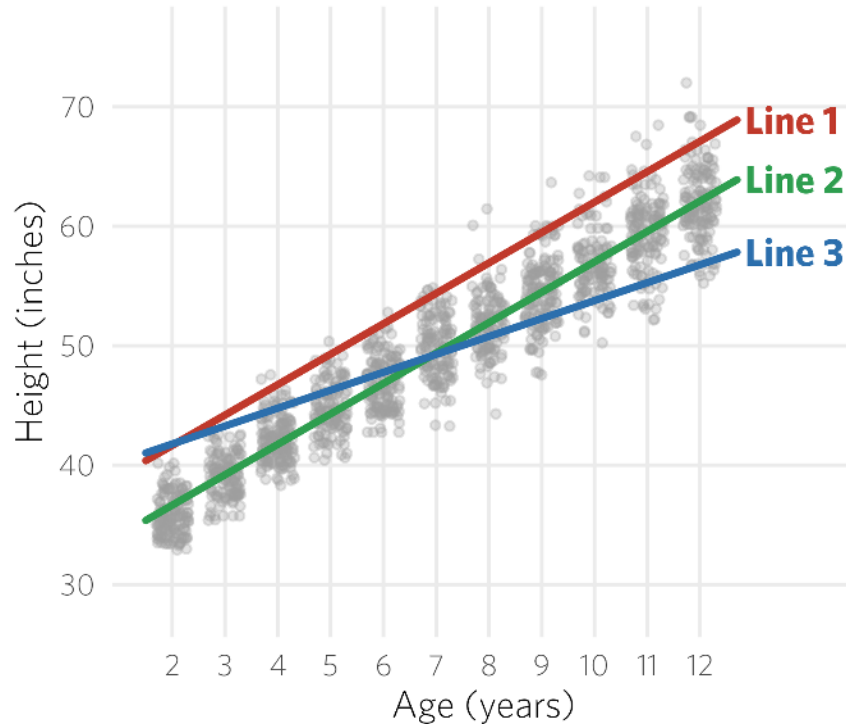
A 10-year-old who is 60 inches tall



- Take one child from the survey, a 10-year-old who is 60 inches tall.
For a 10-year-old the line predicts $31.6 + 2.55 \times 10 = 57.1$ inches.
- Residual = actual minus predicted:
 $60 - 57.1 = +2.9$ inches, the vertical gap from the point to the line.
- A positive residual: the child is taller than the line predicts for their age.

Worksheet - Part 1, a-b

Three lines drawn to summarize the pattern



The figure shows the heights of 1,487 children aged 2 to 12 from the NHANES health survey, with three lines that three students drew to summarize the pattern.

a. Which line is the best summary of the pattern for predicting a child's height from age? Select one: *Line 1*, *Line 2*, or *Line 3*. Then, for each of the two lines you did not select, write one sentence explaining why it is a worse summary of the data than the line you selected. Refer to where the children's points lie: above the line, below it, or both, and at which ages.

b. The regression of height on age yields a slope of 2.5 inches per year of age. Select every statement that is a correct reading of that slope. For each statement, write a few words on your sheet saying why you selected it or why you did not.

- A. Each child in the survey grew 2.5 inches a year.
- B. Children who are one year older than other children are predicted to be 2.5 inches taller, on average.
- C. Age accounts for 2.5% of the variation in the children's heights.
- D. The correlation between age and height is 2.5.
- E. The predicted height of a 10-year-old is 2.5 inches more than the predicted height of a 9-year-old.

Worksheet - Part 1, c

c. Age can enter a model as a categorical or as a continuous variable. With 2 levels (ages 2 to 7 and 8 to 12): 2 parameters, PRE = 0.67. With 11 levels, one mean per year of age: 11 parameters, PRE = 0.919. As a continuous variable, with the regression line: 2 parameters, PRE = 0.917. A fourth model has 6 levels of two years each (2 to 3, 4 to 5, and so on): 6 parameters, PRE not given. Select every statement that is right. For each statement, write a few words on your sheet saying why you selected it or why you did not.

A. The 6-level model's PRE is at least 0.67, because each of the 2 levels is split into 6-level groups, and splitting a group can only lower the error.

B. The 6-level model's PRE is at most 0.919, because each 6-level group is split into single years in the 11-level model.

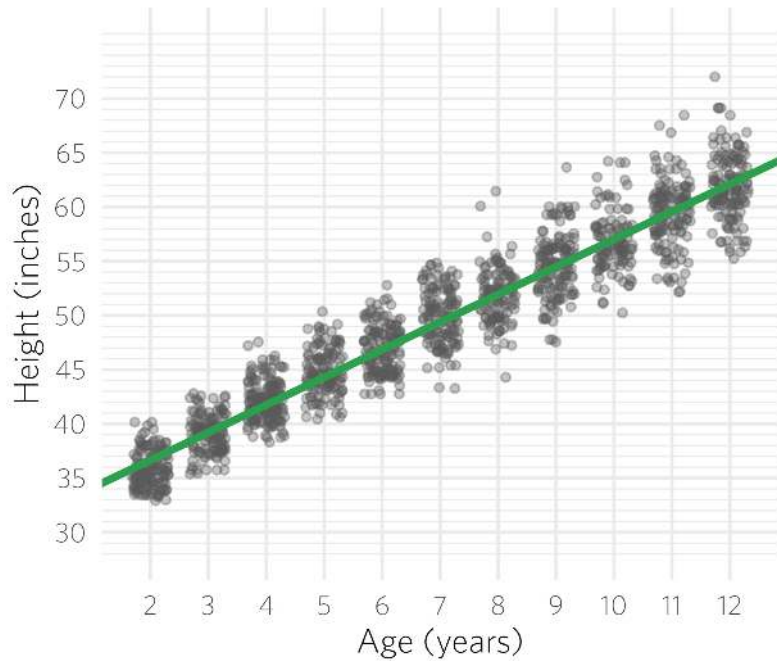
C. The 6-level model's PRE is at least 0.917, the line's, because 6 parameters are more than 2.

D. The line's 2 parameters do nearly the work of 11 means, because the yearly means lie close to a straight line.

E. The 11-level model is the best choice for predicting a new child's height, because its PRE is the highest.

Worksheet - Part 1, d

The regression line of height on age (PRE 0.92)



Rounded to simple numbers, the regression line is **height = 32 + 2.5 x age**, with height in inches and age in years.

d. The oldest children in these data are 12. For a 24-year-old, the line gives $32 + 2.5 \times 24 = 92$ inches, about 7 feet 8 inches. Should the line be used to predict the height of a 24-year-old? Write *yes* or *no*, then one or two sentences giving your reason; refer to the data the line was fitted to.

Formulas for Part 1

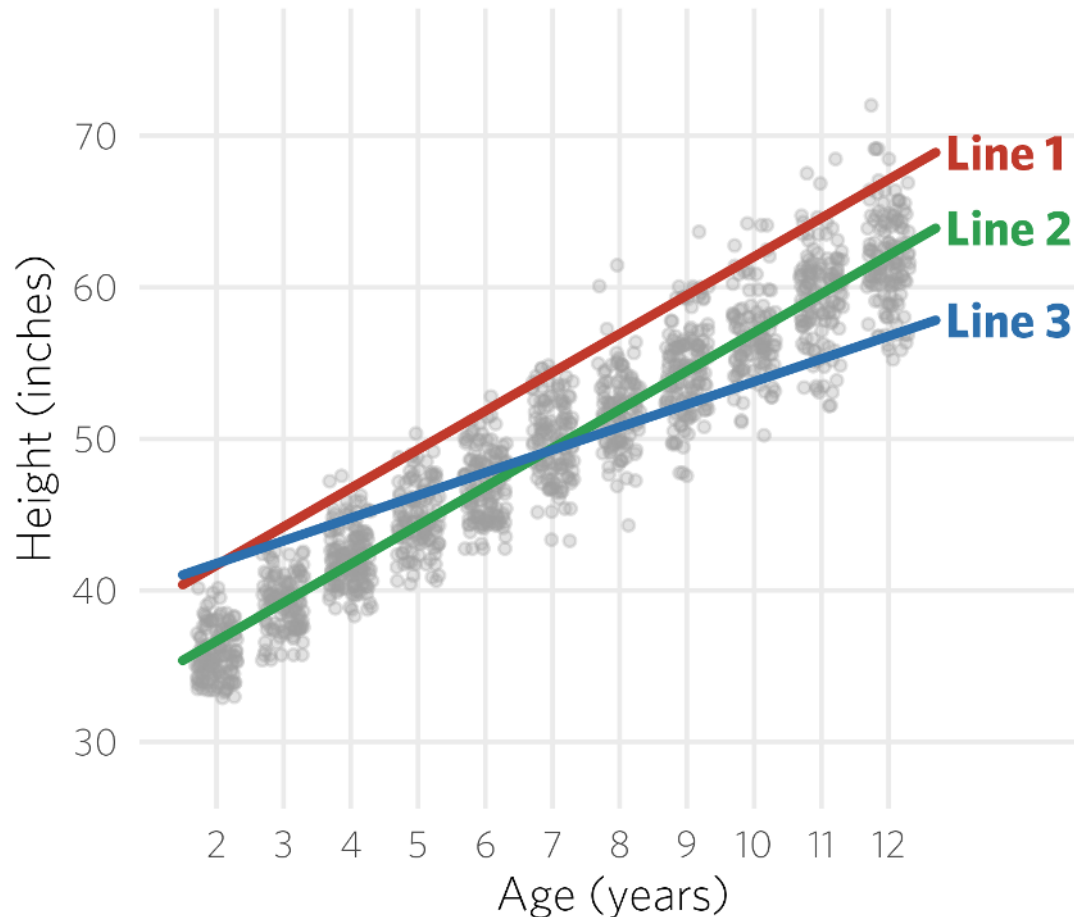
$$\text{predicted height} = 32 + 2.5 \times \text{age}$$

$$\text{PRE} = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

The line gives inches of height for an age in years; PRE is the proportion of the total variation that a model removes.

Part 1a - Which line?

Three lines drawn to summarize the pattern



- **a.** *The best line for prediction.* Line 2: children above and below it in about equal numbers at every age.
- Line 1 sits above nearly every point: 98% of the children are below the line.
- Line 3 is too flat: above every 2-year-old, below 95% of the 12-year-olds.
- A line for prediction summarizes the whole cloud; the check is where the points fall across the whole range of ages.

Part 1b · The five readings of the slope

- **b.** *Correct readings of 2.5 inches per year.* B and E.
- **A, no.** Each child was measured once; the slope compares children of different ages, not one child over time.
- **B, yes.** Children one year older are predicted to be 2.5 inches taller, on average: the slope in words, with its units.
- **C, no.** 2.5 is in inches per year, not a proportion. The proportion of variation is $PRE = 0.92$.
- **D, no.** $r = 0.96$ here, and no correlation exceeds 1.
- **E, yes.** B for one pair of ages: 2.5 inches per year, times one year.

Part 1c · Age as 2, 6, or 11 groups, or as a line

Age as	Parameters	PRE
2 levels (2 to 7, 8 to 12)	2	0.67
6 levels (two years each)	6	0.89
11 levels (one per year)	11	0.919
a continuous variable (the line)	2	0.917

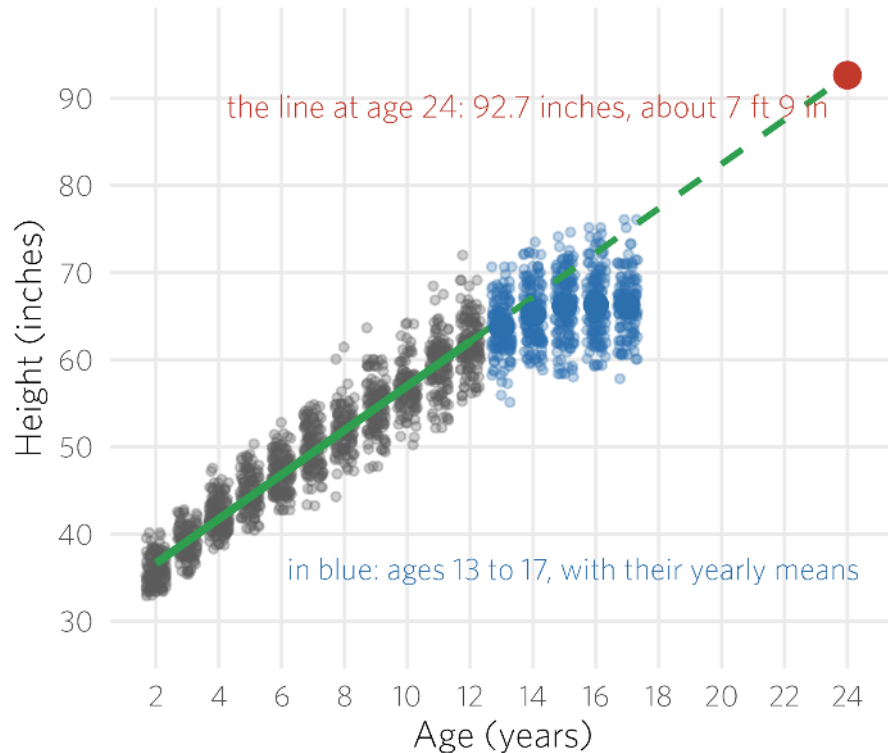
- **c.** *Which statements are right?* A, B, and D. The 6-level PRE, withheld on the sheet, is 0.89.
- A finer split of the same groups can only lower the error, so PRE rises from 2 levels to 6 to 11.
- More parameters do not guarantee a better fit: six flat steps lose to one sloped line.

Part 1c · The five statements

- **A, yes.** Each 2-level group is split into 6-level groups, and splitting a group can only lower the error: the 6-level PRE is at least 0.67.
- **B, yes.** Each 6-level group is split into single years in the 11-level model: the 6-level PRE is at most 0.919.
- **C, no.** Six parameters are more than two, but six flat steps fit a rising pattern worse than one sloped line: 0.89 against 0.917.
- **D, yes.** The yearly means lie close to a straight line, so the line's two parameters recover 0.917 of the 0.919.
- **E, no.** The line matches the eleven means to two decimals with nine fewer parameters, and predicts at every age.

Part 1d - The line at age 24

The line fitted to ages 2 to 12, extended to age 24



- **d.** *Should the line be used at age 24?* No. The line was fitted to ages 2 to 12; the data say nothing about age 24.
- The impossible prediction is evidence that the pattern changes: growth slows in the teens and stops. The 13-to-17-year-olds fall farther below the line each year.
- Predicting adults' heights would take data on adults.

Today

Regression and correlation



Guest lecture: tidying and
wrangling data, with Jane
Stephenson



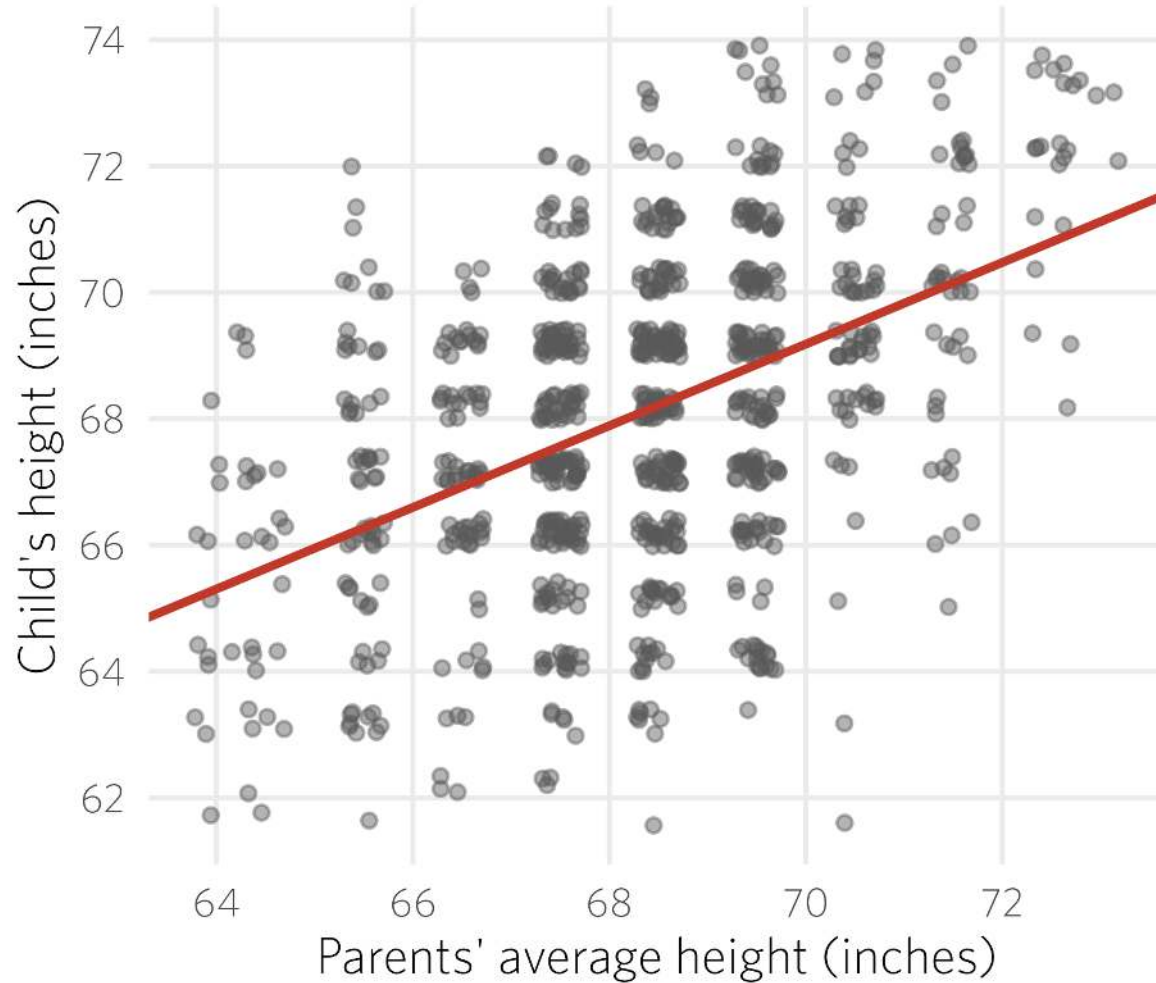
The slope and the intercept



How strong is the
relationship?

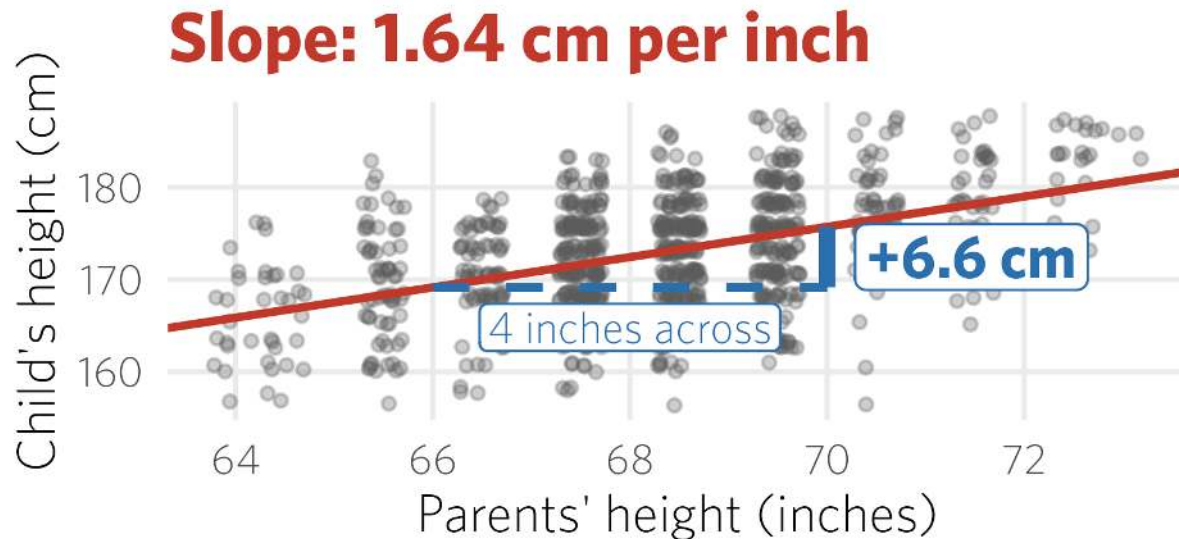
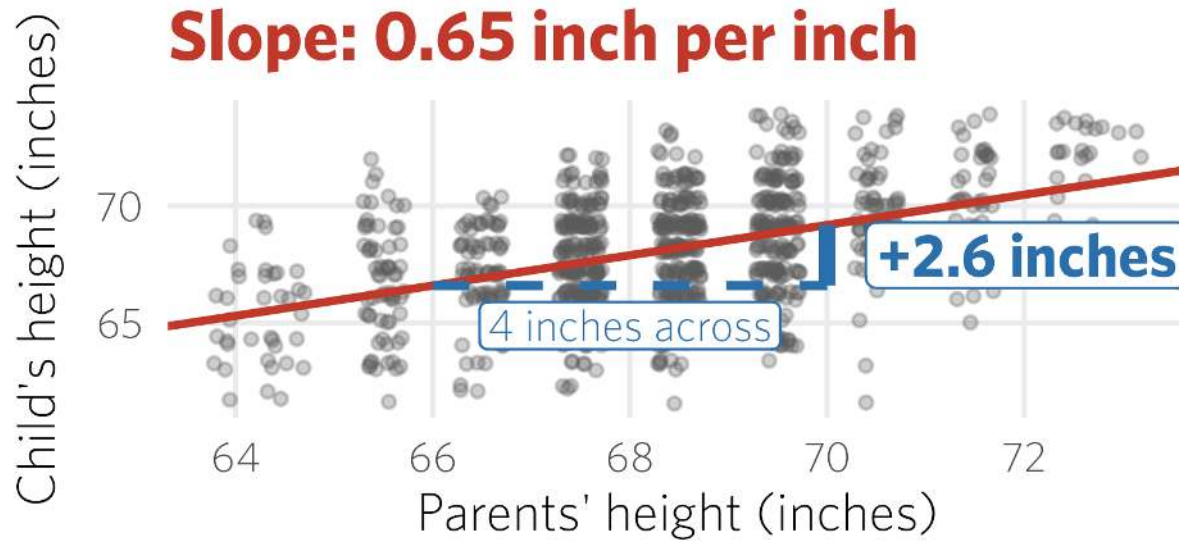
Galton's parents and children

928 adult children and their parents (Galton, 1886)



- The next line predicts an adult child's height from the parents' average height:
Francis Galton's data from 1886, 928 adult children.
- The fitted line is $\text{child} = 23.9 + 0.65 \times \text{parent}$:
each additional inch of parents' height predicts 0.65 inch more in the child's height.

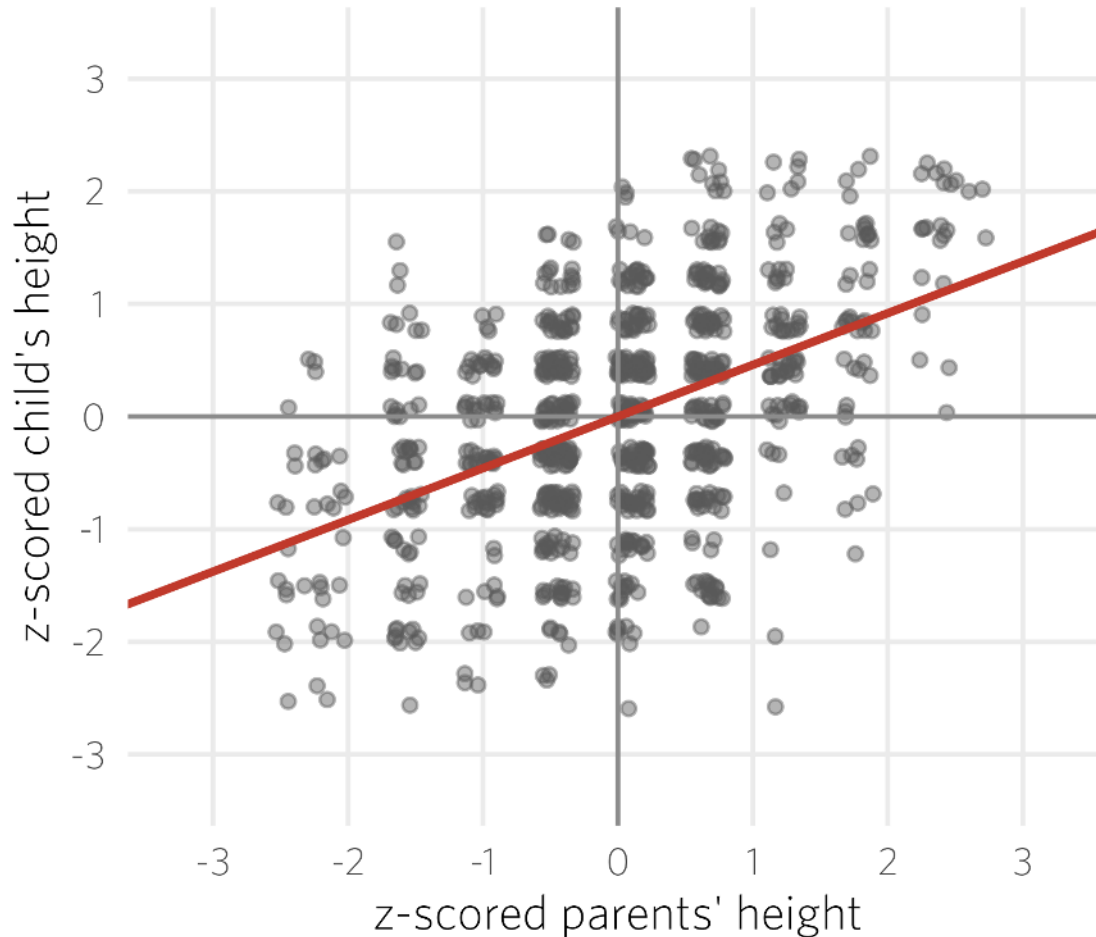
The slope depends on the units



- The same 928 pairs twice: the children's heights in inches, then in centimeters.
Which panel shows the stronger relationship?
- Neither. One cloud, two slopes: 0.65 inch per inch and 1.64 cm per inch.
- A slope has units. Strength needs a number with no units.

Both variables as z-scores

The same 928 pairs as z-scores: slope 0.46



- To compare strength across variables, remove the units: z-score each variable, as in Lecture 7.

$$z = \frac{x - \bar{x}}{s}$$

- Each height becomes its distance from the mean, in SDs. The cloud keeps its shape.
- The line passes through (0, 0); its slope is 0.46.

Calculating r

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

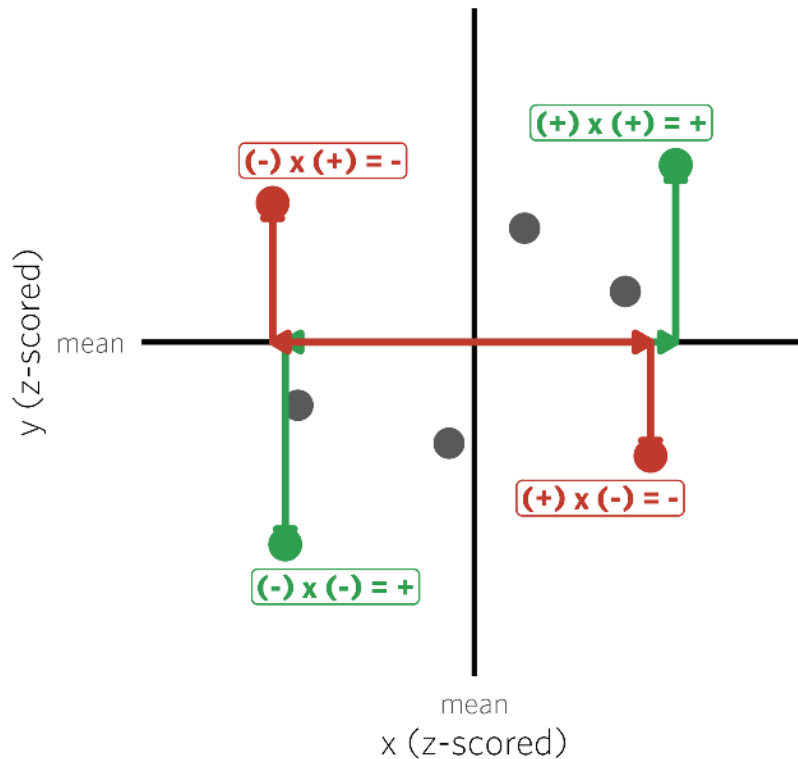
$$\text{covariance} = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{n - 1}$$

$$r = \frac{\text{covariance}}{s_x s_y}$$

- The variance, from Lecture 6: each deviation from the mean times itself, summed, divided by n - 1.
- The covariance: each case's x-deviation times its y-deviation, in place of the x-deviation twice.
- r divides the covariance by the two SDs: a standard scale from -1 to 1, the same number as the slope of the z-scored line.

Which points raise r, and which lower r

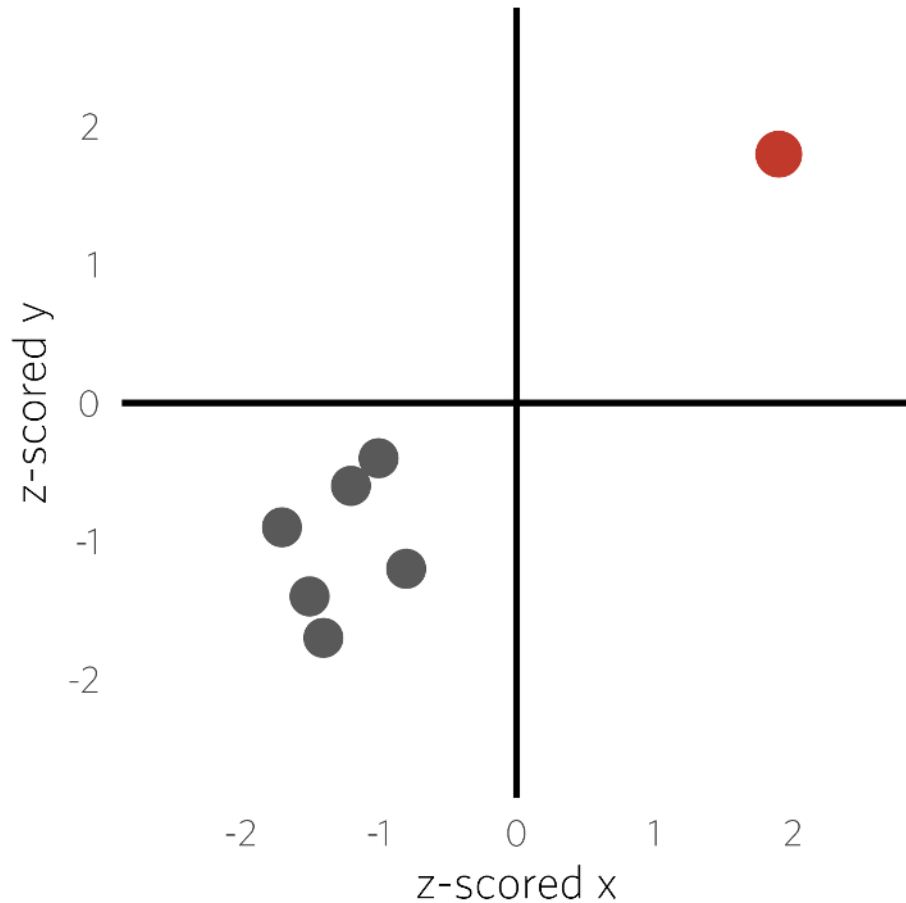
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1) s_x s_y}$$



- Top right: both deviations positive. The product is positive: the point raises r.
- Bottom left: both negative. The product is positive again: this point also raises r.
- Top left: one negative, one positive. The product is negative: the point lowers r.
- Bottom right: negative again, so the point lowers r.

$r = +1$, $r = -1$, $r = 0$, and one outlier

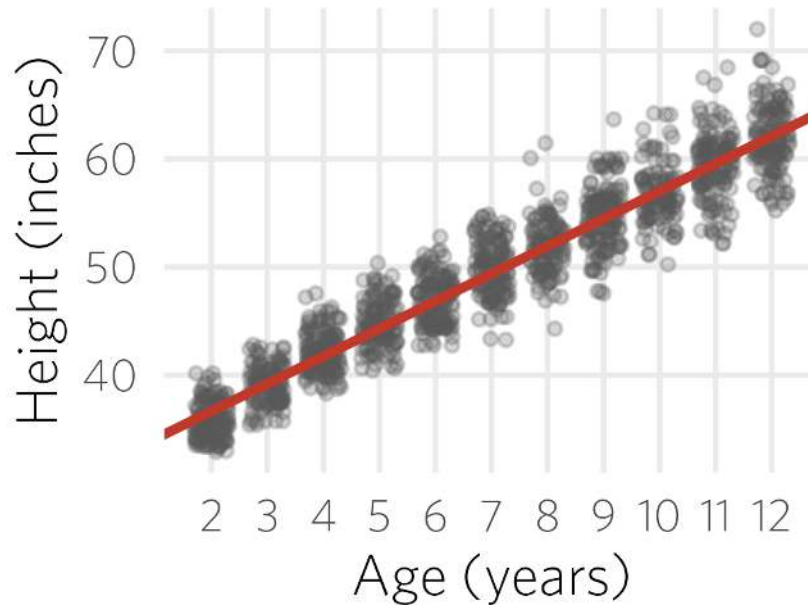
One outlier: $r = 0.92$ for a cloud with no tilt



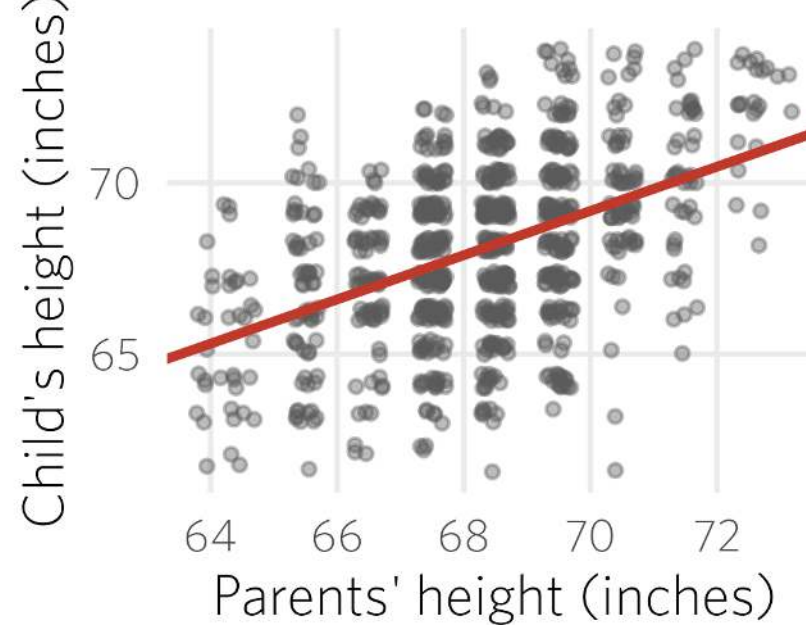
- $r = +1$ when every point lies on a rising line.
- $r = -1$ when every point lies on a falling line.
- $r = 0$ for a ball-shaped cloud with no tilt.
- One outlier can make r large for a cloud with no trend.
 r can be computed for any set of points,
so a scatter plot comes first.

Which pair is more strongly correlated?

A. 1,487 children: height by age



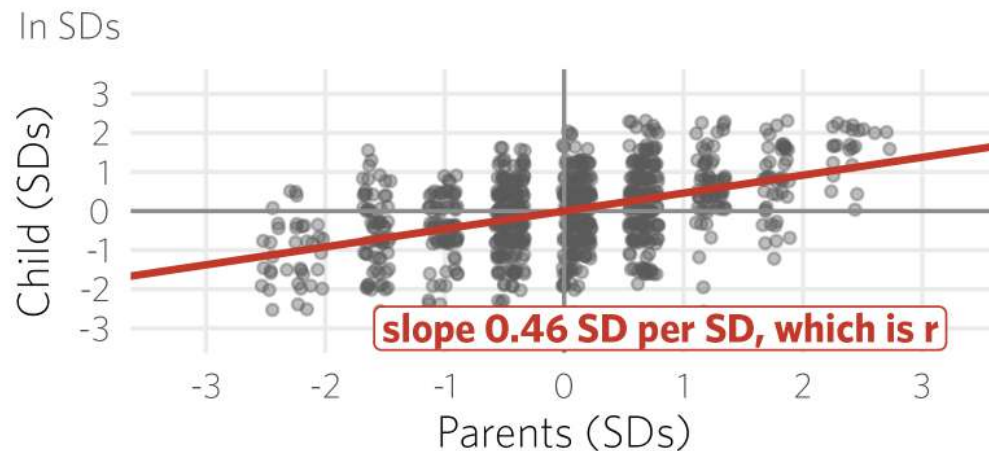
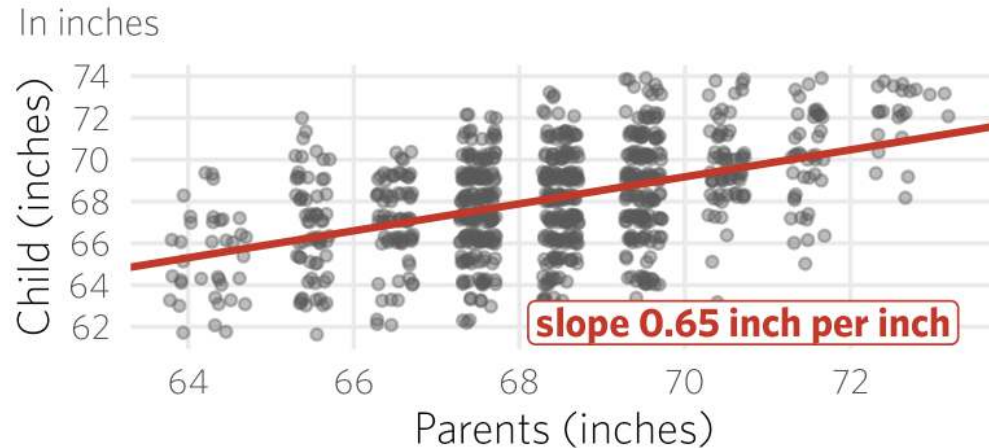
B. 928 adults: height by parents' height



Which pair of variables is more strongly correlated: A or B?

A: $r = 0.96$, against 0.46 in B. The two slopes, 2.5 inches per year and 0.65 inch per inch, are in different units and cannot be compared; r , with no units, allows the comparison.

r and the slope



- The same cloud twice: in inches above, in SDs below.
- The slope in SDs is r .
The slope in inches is r multiplied by the ratio of the two SDs.

$$b_1 = r \times \frac{s_y}{s_x}$$

- r says how close the points lie to the line;
the slope says how much y changes per unit of x , for predictions.
- With equal SDs, as for z -scores, the slope and r are the same number.

r^2 is the proportional reduction in error

$$r = 0.46$$

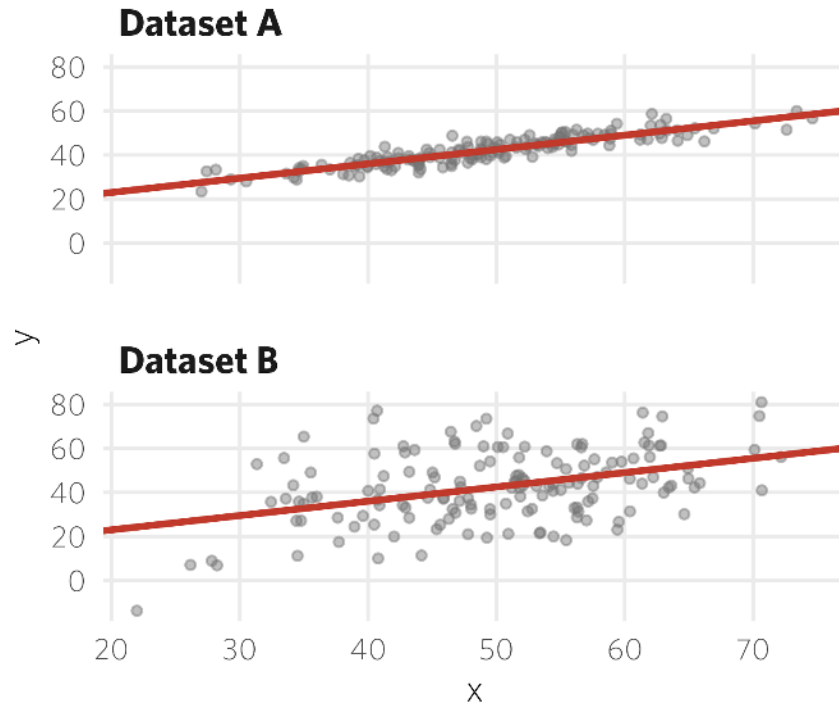
$$r^2 = 0.46 \times 0.46 = 0.21$$

$$\text{PRE} = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = 0.21$$

- Galton's correlation is $r = 0.46$.
- r^2 is the proportion of the variation in children's heights that parents' height accounts for: 21%, not 46%.
- Fit the same line with `lm()` and build its ANOVA table as on Monday: $SS_{\text{Model}} / SS_{\text{Total}} = 0.21$, the same number, the proportional reduction in error (PRE).
- PRE, r^2 , and R^2 are three names for one quantity.

Worksheet - Part 2, a-b

The same fitted line, $y = 10 + 0.65x$, in both



a. The figure shows two simulated datasets, A and B, of 150 points each. Both have the same fitted line, $y = 10 + 0.65x$. Which dataset has the larger correlation, r ? Select A or B, then write one sentence giving the reason.

b. A news story reports a correlation of $r = 0.5$ between students' hours of sleep and their exam scores, and says: "Sleep accounts for half of the variation in exam scores." This sentence is not quite right. Write a corrected version of the sentence.

Worksheet - Part 2, c

c. A university records each student's major as a number from 1 to 60, in the order the majors were added to the catalog. An analyst finds a correlation of 0.20 between major number and starting salary and concludes that students in higher-numbered majors earn more. Select every statement that is right. For each statement, write a few words on your sheet saying why you selected it or why you did not.

A. It is reasonable: a correlation of 0.20 is weak but positive.

B. It cannot be drawn: the major numbers are labels, and renumbering the majors would change the correlation.

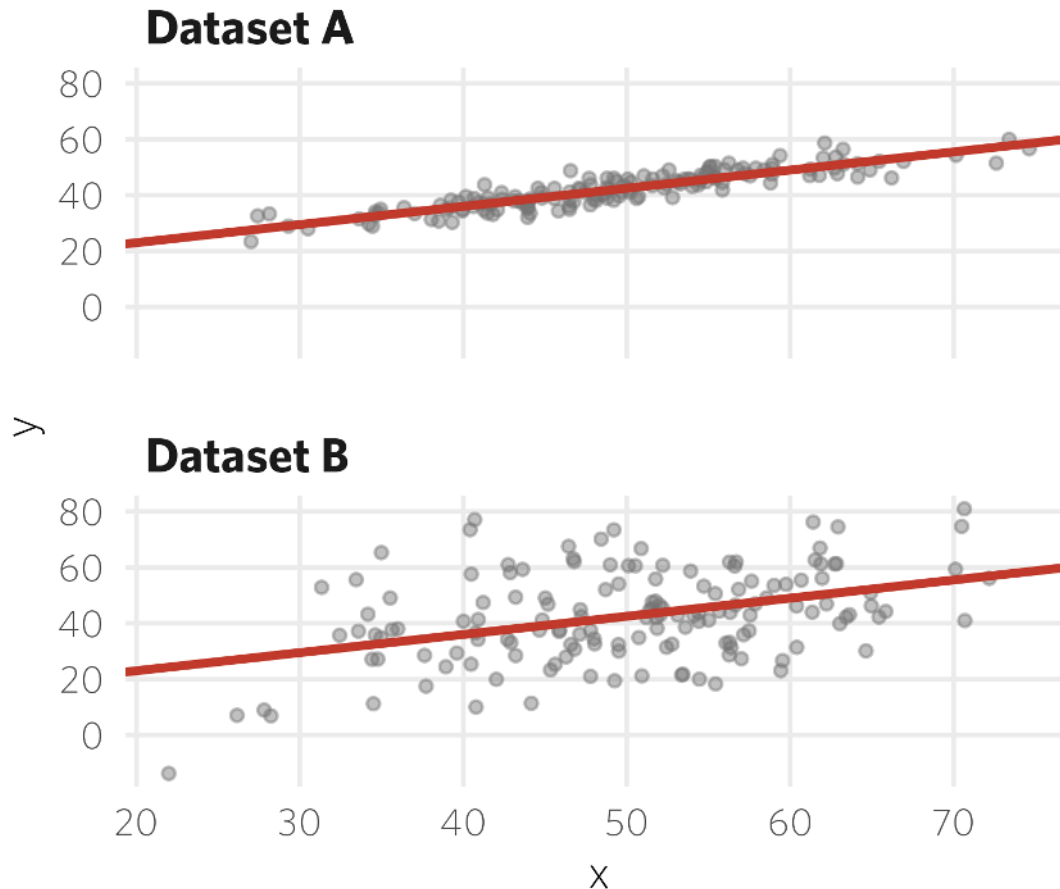
C. It is reasonable once the analyst also reports the slope, in dollars per major number.

D. It is wrong, because a correlation this small can be trusted only with at least 1,000 students.

E. Numbering the majors by average salary, 1 to 60, would produce a large positive correlation that says nothing new.

Part 2a - The same slope, two correlations

The same fitted line, $y = 10 + 0.65x$, in both



- **a.** Which dataset has the larger r ? A: $r = 0.90$, against 0.40 in B.
- The slopes are equal, so steepness cannot separate the two datasets.
- r measures the spread of the points around the line, not the steepness of the line.

Part 2b • r^2 , not r

- **b.** *The sleep sentence, corrected.* "Sleep accounts for a quarter of the variation in exam scores." The proportion is $r^2 = 0.5 \times 0.5 = 0.25$.
- Accounting for half of the variation would take $r = 0.71$.

Part 2c · The five statements about the majors

- **c.** *The majors numbered 1 to 60.* B and E.
- **A, no.** Software computes a correlation for any two columns of numbers, meaningful or not; 0.20 is a number, not a finding.
- **B, yes.** The major numbers are labels; renumbering the majors would change the correlation.
- **C, no.** A slope in dollars per major number has the same problem: a step of one in major number is not an increase in any quantity.
- **D, no.** More students would give the same uninterpretable correlation, more precisely.
- **E, yes.** Numbering the majors by salary would manufacture a large correlation out of nothing, which is the point of B.

Before Friday

- Finish CourseKata **Chapter 9**.
- Friday: **Case #5**, in your tutorial session. Check your Slack channel for your time slot.
- Milestone 5 is due Sunday.

Hand in your worksheet on the way out

On the last line, write down one thing you learned today.

Thanks!