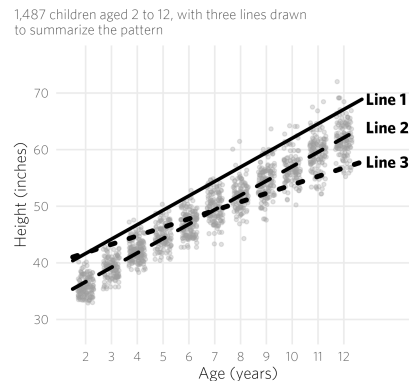


PSYCH 10 · Fall 2026 · Worksheet for Lecture 11: Regression and correlation · Answer key

The answers are in blue. Where a question has more than one defensible answer, the key says so and describes what a good answer contains.

Part 1 · Reading a regression line



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.

a. *Line 2. Line 1 sits above nearly every point: at every age most of the children are below it, so almost every residual would be negative. Line 3 is too flat: it runs above the youngest children and below the oldest, so its residuals would be positive at one end of the age range and negative at the other. Line 2 has children above and below it in about equal numbers at every age, which is the requirement for a line meant for prediction. A good sentence about a rejected line says where the points lie relative to it (above, below, or both) and at which ages; "it looks wrong" is not enough.*

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.

b. *B and E. E is B for one pair of ages: 2.5 inches per year, times one year. A reads the slope as the growth of one child, but each child in the survey was measured once; the slope compares children of different ages at one time. C reads the slope as a proportion of variation; 2.5 is in inches per year, and the proportion that age accounts for is $PRE = 0.92$. D reads the slope as a correlation; r for these data is 0.96, and no correlation can exceed 1. A good reason says what the 2.5 is: the difference in predicted height, in inches, between children one year apart in age.*

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.

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.

c. *A, B, and D. The 6-level model's PRE is 0.893. A and B are the nesting argument from Lecture 10: splitting a group into smaller groups can only lower the error, so PRE rises from 2 levels (0.67) to 6 (0.893) to 11 (0.919), and the 6-level value has to sit between its neighbors. C is wrong: the 6-level model has more parameters than the line but a lower PRE, because six flat steps describe a rising pattern worse than one sloped line does; parameters help only when they match the shape of the pattern. D is right: the eleven yearly means lie close to a line, so the line's two parameters recover 0.917 of the 0.919. E is wrong: the line matches the eleven means to two decimals with nine fewer parameters and gives a prediction at every age, including ages between the years. A good reason names the nesting (which groups are splits of which) or the shape (flat steps against a slope).*

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.

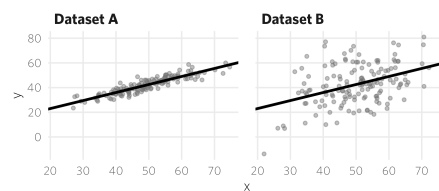
Rounded to simple numbers, the regression line is **height = 32 + 2.5 x age**, with height in inches and age in years.

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.

d. *A strong answer says the line should not be used at age 24, because it was fitted to children aged 2 to 12 and the data say nothing about what happens past 12. The impossible prediction is itself evidence that the pattern changes: growth slows in the teens and then stops, which is why the survey's 13-to-17-year-olds fall farther below the line each year. Predicting adults' heights would take data on adults. Weaker answers: "the line is wrong" (it works between ages 2 and 12), "a 24-year-old is about 68 inches" (one unsupported number in place of another), and "the arithmetic is right" (it is, and that is not the question).*

Part 2 · The correlation coefficient

Two simulated datasets of 150 points, each with the fitted line $y = 10 + 0.65x$



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.

a. *A. The two lines have the same slope, so the tilt cannot tell the datasets apart; in A the points lie close to the line and in B far from the line. r measures the spread of the points around the line, not the steepness of the line. (r is 0.90 in A and 0.40 in B.)*

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.

b. *"Sleep accounts for a quarter of the variation in exam scores." The proportion of variation accounted for is r squared, $0.5 \times 0.5 = 0.25$, not r itself; accounting for half of the variation would take a correlation of about 0.71.*

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.

c. *B and E. A takes the number at face value, but software computes a correlation for any two columns of numbers, meaningful or not. C does not help: a slope in dollars per major number is no more meaningful than the correlation, because a step of one in major number is not a step in anything. D locates the problem in the number of students, but the problem is what the numbers mean; a million students would give the same uninterpretable correlation, more precisely. E is right, and it shows why B is right: the correlation depends entirely on the order in which the labels were assigned, so a numbering chosen by salary manufactures a large correlation out of nothing. A good reason says what the major numbers are: labels in catalog order, not amounts of anything.*

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.