Regression

Patrick Breheny

February 12

NHANES

- Every few years, the CDC conducts a huge survey of randomly chosen Americans called the National Health and Nutrition Examination Survey (NHANES)
- Hundreds of variables are measured on these individuals:
  - Demographic variables like age, education, and income
  - Physiological variables like height, weight, blood pressure, and cholesterol levels
  - Dietary habits
  - Disease status
  - Lots more: everything from cavities to sexual behavior

Predicting weight from height

- For the 2,649 adult women in the NHANES data set:
  - average height = 5 feet, 3.5 inches
  - average weight = 166 pounds
  - SD(height) = 2.75 inches
  - SD(weight) = 44.5 pounds
  - correlation between height and weight = 0.3
- Suppose you were asked to predict a person's weight from their height
- First, an easy case: suppose the woman was 5 feet, 3.5 inches
- Since the woman is average height, we have no reason to guess anything other than the average weight, 166 pounds

Predicting weight from height (cont'd)

- How about a woman who is 5'6?
- She's a bit taller than average, so she probably weighs a bit more than average
- But how much more?
- To put the question a different way, she is almost one standard deviation above the average height; how many standard deviations above the average weight should we expect her to be?

Using the correlation coefficient

- The answer turns out to depend on the correlation coefficient
- Since the correlation coefficient for this data is 0.3, we would expect the woman to be 0.3 standard deviations above the mean weight, or $166 + 0.3(44.5) = 179$ pounds

Procedure

Suppose we are interested in predicting y based on x; to put the thought process from the previous slide into an explicit procedure, we have:

(1) Standardize x:

$$z_x = \frac{x - \bar{x}}{SD_x}$$

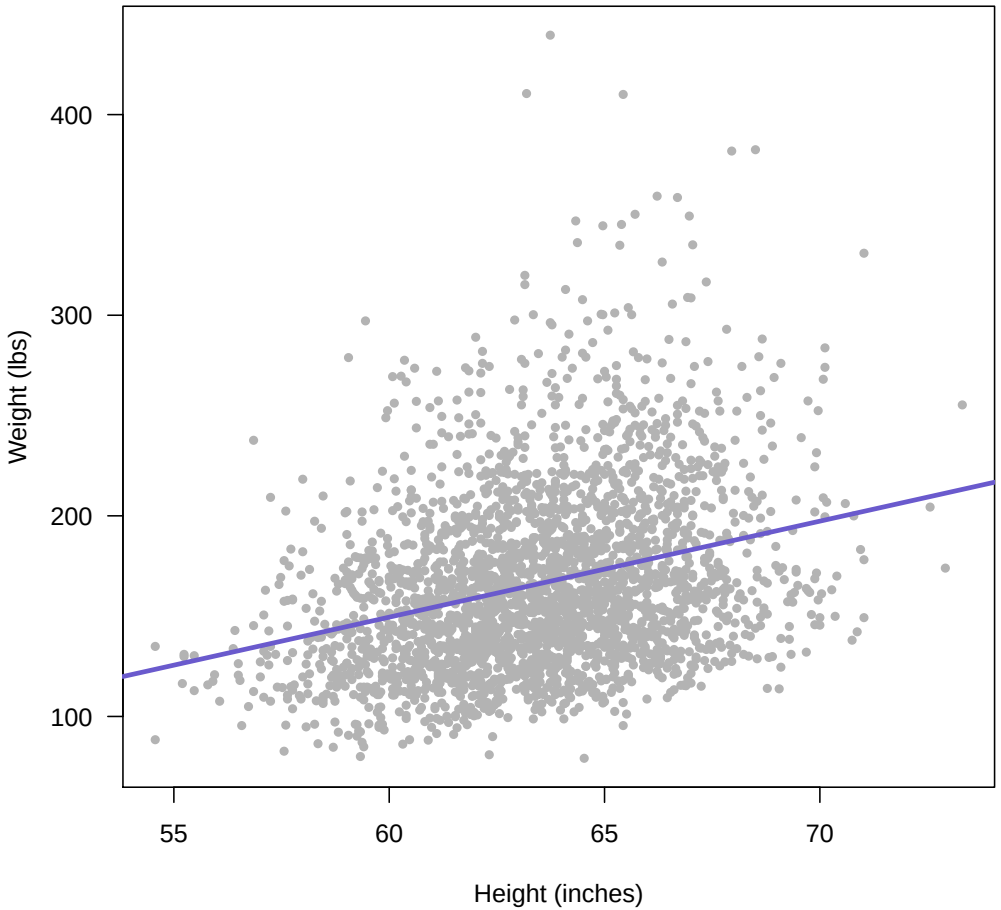
(2) Relate the standardized values of x and y:

$$z_y = r z_x$$

(3) Unstandardize y:

$$y = \bar{y} + z_y SD_y$$

Graphical interpretation



The regression line

- This line is called the *regression* line
- It tells you, for any height, the average weight for women of that height
- Here, we were trying to predict one variable based on one other variable; if we were trying to predict weight based on height, dietary habits, and cholesterol levels, or trying to study the relationship between cholesterol and weight while controlling for height, then this is called *multiple regression*
- Multiple regression is beyond the scope of this course, but is a major topic in Design and Analysis of Biomedical Studies (BIOS 5120)

The equation of the regression line

- Like all lines, the regression line may be represented by the equation

$$y = \alpha + \beta x,$$

  where α is the intercept and β is the slope
- For the height/weight NHANES data, the intercept is -137 pounds and the slope is 4.8 pounds/inch

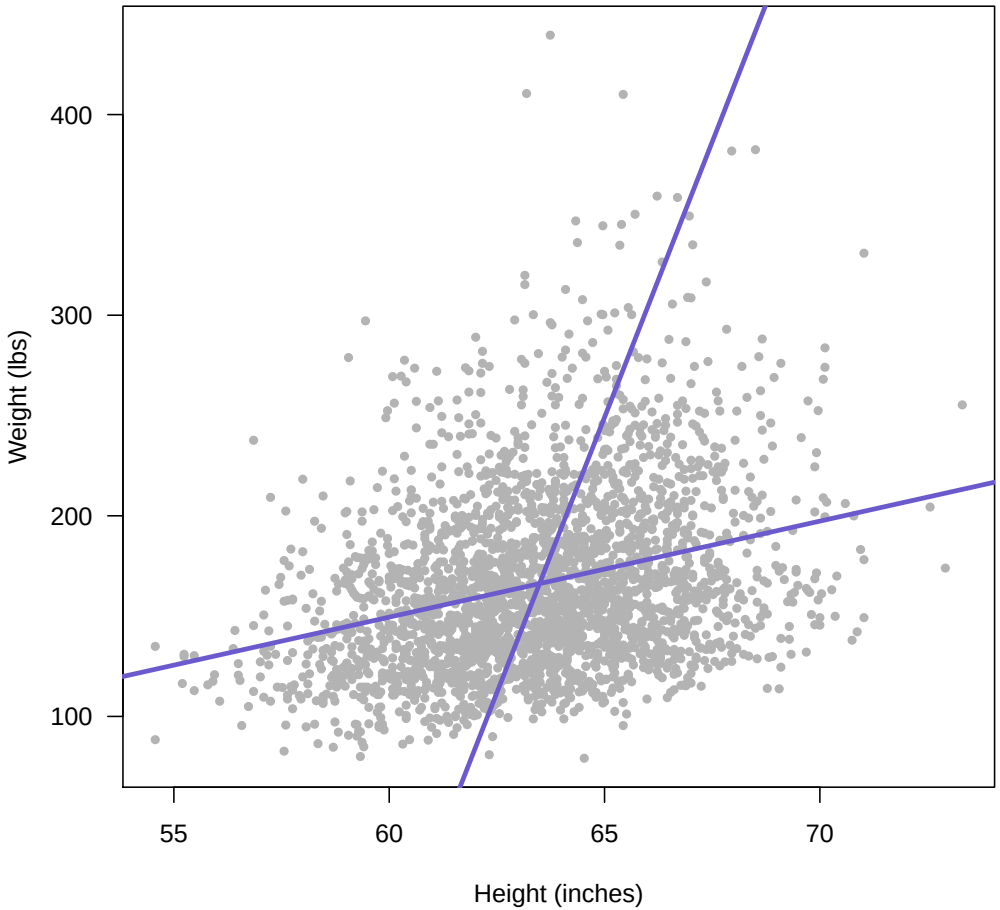
β vs. r

- Note the similarity and the difference between the slope of the regression line (β) and the correlation coefficient (r):
  - The correlation coefficient says that if you go up in height by one standard deviation, you can expect to go up in weight by $r = 0.3$ standard deviations
  - The slope of the regression line tells you that if you go up in height by one inch, you can expect to go up in weight by $\beta = 4.8$ pounds
- Essentially, they tell you the same thing, one in terms of standard units, the other in terms of actual units
- Therefore, if you know one, you can always figure out the other simply by changing units (which here involves multiplying by the ratio of the standard deviations)

There are two regression lines

- We said that the correlation between weight and height is the same as the correlation between height and weight
- This is not true for regression
- The regression of weight on height will give a different answer than the regression of height on weight

The two regression lines



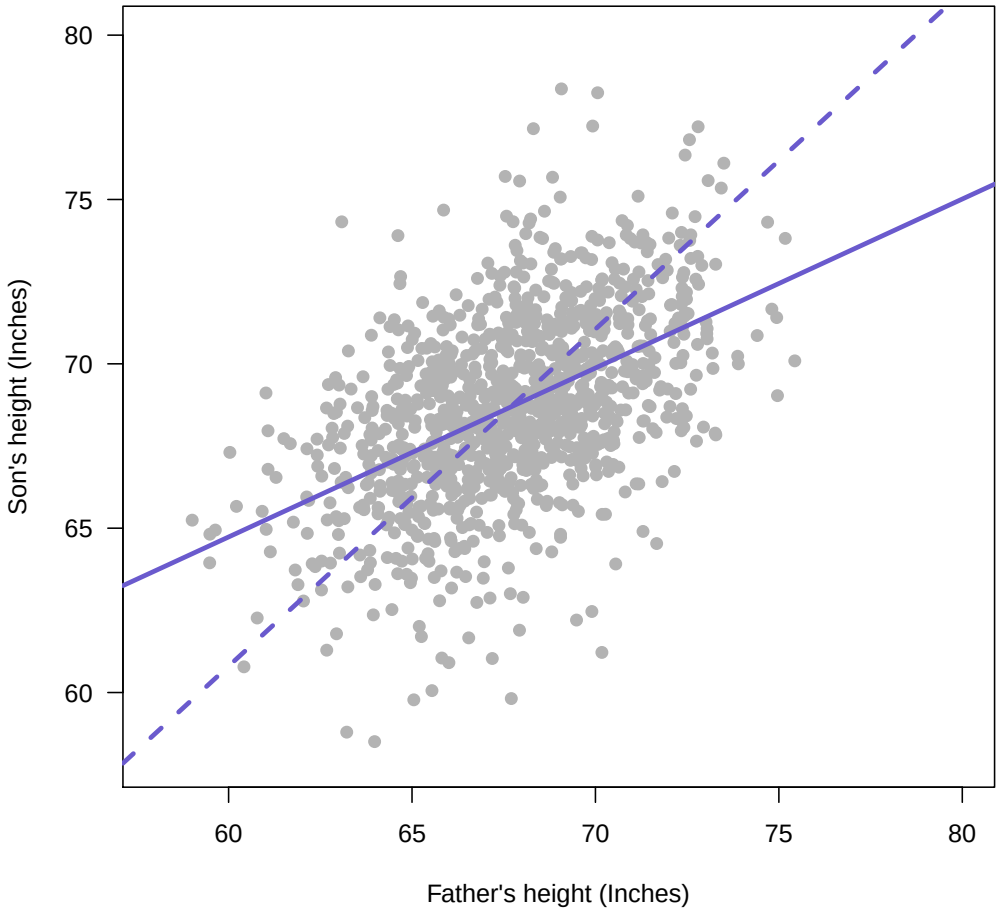
Regression and root-mean-square error

- The amount by which the regression prediction is off is called the *residual*
- One way of looking at the quality of our predictions is by measuring the size of the residuals
- Out of all possible lines that you could draw, which one has the lowest possible root-mean-square of the residuals?
- The regression line
- Because of this, the regression line is also called the "least squares" fit

Why only r standard deviations?

- Only moving r standard deviations away from the average may be counterintuitive; if height goes up by one SD, shouldn't weight too?
- Here's an example that I hope will help clarify this concept:
  - A student is taking her first course in statistics, and we want to predict whether she will do well in the course or not
  - Suppose we know that last semester, she got an A in math
  - Now suppose that we know that last semester, she got an A in pottery
- These two pieces of information are not equally informative for predicting how well she will do in her statistics class
- We need to balance our baseline guess (that she will receive an average grade) with this new piece of information, and the correlation coefficient tells us how much weight the new information should carry

Fathers and sons again



How regression got its name

- Because the correlation coefficient is always less than 1, the regression line will always lie beneath the "x goes up by 1 SD, y goes up by 1 SD" rule
- Galton called this phenomenon "regression to mediocrity," and this is where regression gets its name
- People frequently read too much into the regression effect – this is called the *regression fallacy*

The regression fallacy, example #1

- A group of subjects are recruited into a study
- Their initial blood pressure is taken, then they take an herbal supplement for a month, and their blood pressure is taken again
- The mean blood pressure was the same, both before and after
- However, subjects with high blood pressure tended to have lower blood pressure one month later, and subjects with low blood pressure tended to have higher blood pressure later
- Does this supplement act to stabilize blood pressure?

Why does regression to the mean happen?

- Not really; the same effect would occur if they took placebo
- Why?
- Consider a person with a blood pressure 2 SDs above average
- It's possible that the person has a true blood pressure 1 SD above average, but happened to have a high first measurement; it's also possible that the person has a true blood pressure 3 SDs above average, but happened to have a low first measurement
- However, the first explanation is much more likely

The regression fallacy, example #2

- In professional sports, some first-year players have outstanding years and win "Rookie of the Year" awards
- They often fail to live up to expectations in their second years
- Writers call this the "sophomore slump", and come up with elaborate explanations for it

The regression fallacy, example #3

- An instructor standardizes her midterm and final so that the class average is 50 and the SD is 10 on both tests
- She has taught this class many times and the correlation between the tests is always around 0.5
- This year, she decides to do something different – she takes the 10 students with the lowest scores on the midterm and gives them special tutoring
- On the final, all ten students score above 50; can this be explained by the regression effect?
- No!
- The regression effect can only take these students closer to the average; the fact that they all score *above* average indicates that the tutoring really did work

Summary

- Given means, SDs, and the correlation coefficient between two variables, we can predict one outcome based on the other (know how to do this!)
- The correlation coefficient (r) is a unit-less version of the regression slope (β)
  - They tell you how much weight to give to variable A when predicting the outcome of variable B
  - Given SDs, you can convert between them
- Unless two variables are perfectly correlated, outcomes will tend to lie closer to the average than you would expect from the "x goes up by 1 SD, y goes up by 1 SD" rule