# Regression

Patrick Breheny

December 1

# 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

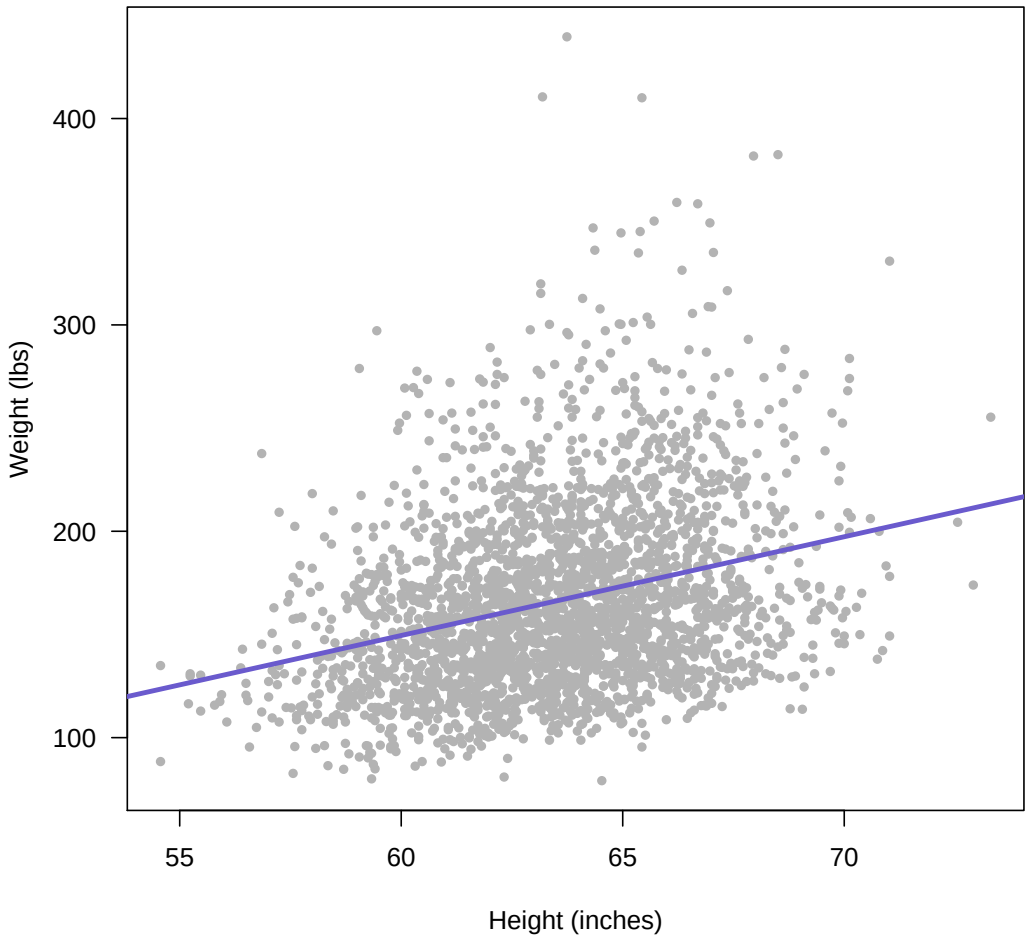
# Predicting weight from height (cont'd)

- How about a woman who is 5 feet, 6.25 inches?
- 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 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
- What about a woman who is 5 feet, zero inches?
- By the same rationale, we would expect her to be $166 - (1.27)(0.3)(44.5) = 149$ pounds

# 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 the major topic in Biostatistical Methods II (BIOS 5720)

# The equation of the regression line

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

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

  where $\alpha$ is the intercept and $\beta$ is the slope
- For the height/weight NHANES data, the intercept is -137 pounds and the slope is 4.8 pounds/inch

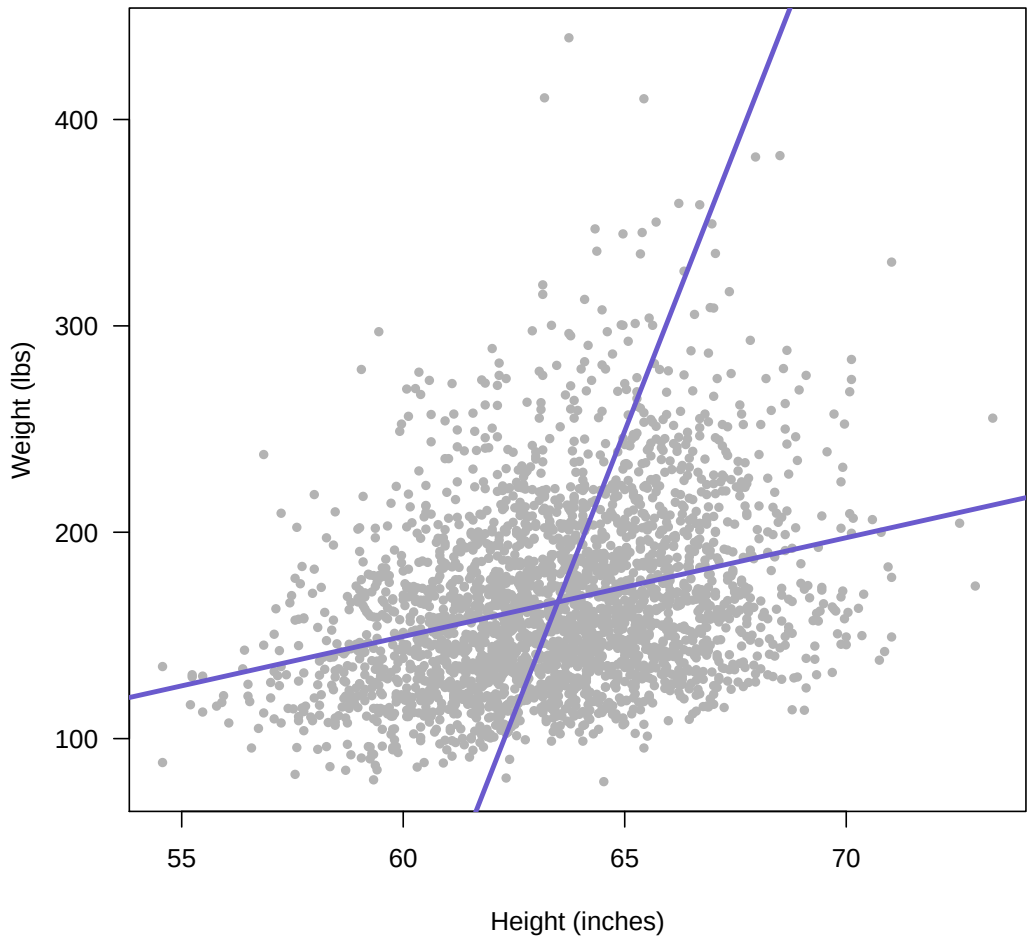
# $\beta$ vs. $\rho$

- Note the similarity and the difference between the slope of the regression line ($\beta$) and the correlation coefficient ($\rho$):
  - The correlation coefficient says that if you go up in height by one standard deviation, you can expect to go up in weight by $\rho = 0.3$ 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 correlation doesn't depend on measurement scale; the correlation between weight and height is 0.3 whether we measure height in inches or meters
- This is *not* true for regression; we will get different values of $\beta$ depending on how we measure height and weight
- We also said that the correlation between weight and height is the same as the correlation between height and weight
- This is also *not* true for regression
- The regression of weight on height will give a different answer than the regression of height on weight

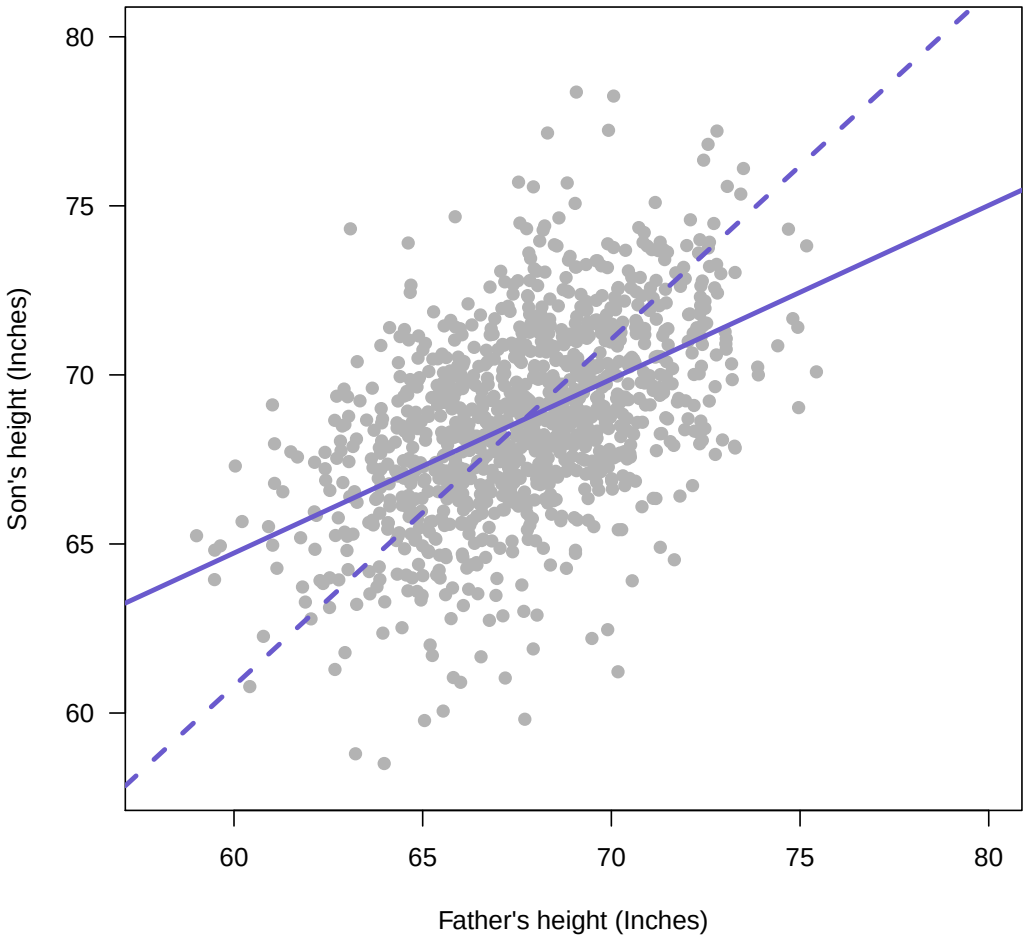
# The two regression lines



# Why only $\rho$ standard deviations?

- Only moving $\rho$ 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*

# Example: The regression fallacy

- 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?

- No; 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 (>50 times more likely)

# What makes regression a good method for prediction?

- We've said that the regression line is a good way of predicting one outcome on the basis of another piece of information
- But what makes it a good method for prediction? Is there some objective criterion by which it can be said to be the best way of making predictions?
- As it turns out, yes: the regression line is the line with the lowest prediction error, as measured by squaring the prediction errors

# Regression and root-mean-square error

- The amount by which the regression prediction is off is called the *residual*:

$$r_i = y_i - (\hat{\alpha} + x_i\hat{\beta})$$

- 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? I.e., which one minimizes $\sum_i r_i^2$?
- The regression line
- Because of this, the regression line is also called the "least squares" fit

# $\widehat{\beta}$ as least-squares solution

- We'll now derive a few theoretical results concerning $\widehat{\beta}$
- In what follows, we'll assume without loss of generality that $x$ and $y$ are centered (i.e., have mean zero); this simplifies the math without changing the results
- **Theorem:** The value of $\beta$ that minimizes $\sum_i r_i^2$ is given by:

$$\widehat{\beta} = \frac{\sum_i \tilde{x}_i \tilde{y}_i}{\sum_i \tilde{x}_i \tilde{x}_i} = \frac{\langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle}{\langle \tilde{\mathbf{x}}, \tilde{\mathbf{x}} \rangle},$$

where $\tilde{x}, \tilde{y}$ denote the centered version of $x$ and $y$

# Distribution of $\widehat{\beta}$

- Let us now assume that $y_i \sim \mathrm{N}(\alpha + x_i\beta, \sigma^2)$
- In this situation, $\hat{\beta}$ is simply a linear combination of normally distributed variables and therefore follows a normal distribution itself:
- **Theorem:** Suppose $y_i \sim \mathrm{N}(\alpha + x_i\beta, \sigma^2)$. Then

$$\widehat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum_i \tilde{x}_i^2}\right)$$

- **Corollary:** Suppose $y_i \sim \mathrm{N}(\alpha + x_i\beta, \sigma^2)$. Then

$$\frac{\widehat{\beta} - \beta}{\hat{\sigma}/\sqrt{\sum_i \tilde{x}_i^2}} \sim t_{n-2},$$

where $\hat{\sigma}^2 = \sum_i r_i^2/(n-2)$

# Confidence interval for $\widehat{\beta}$: Example

- Previously, we stated that the regression coefficient for the NHANES height-weight data was $\widehat{\beta} = 4.8$ pounds per inch; now let's construct a confidence interval for $\widehat{\beta}$
- With $n = 2649$ and $\hat{\sigma} = 42.6$, the standard error is

$$\text{SE} = \frac{42.6/\sqrt{2649}}{2.75} = 0.301$$

- With such a large sample size, the $t$ is very similar to the normal, and $\pm 1.96$ contains the middle 95% of the distribution, so our confidence interval is

$$4.8 \pm 1.96\text{SE} = [4.20, 5.38]$$

# Summary

- Given means, SDs, and the correlation coefficient between two variables, we can predict one outcome based on the other
- The correlation coefficient ($r$) is a unit-less version of the regression slope ($\beta$)
  - 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
- If we assume $Y$ follows a normal distribution, $\widehat{\beta}$ follows a normal distribution, which allows us to construct confidence intervals in a straightforward manner