Correlation

Patrick Breheny

February 11

Introduction

- Box plots are a way to examine the relationship between a continuous variable and a categorical variable
- In lab, we saw bar charts as a way of comparing two (or more) categorical variables
- Now, we will discuss how to summarize and illustrate the relationship between two continuous variables

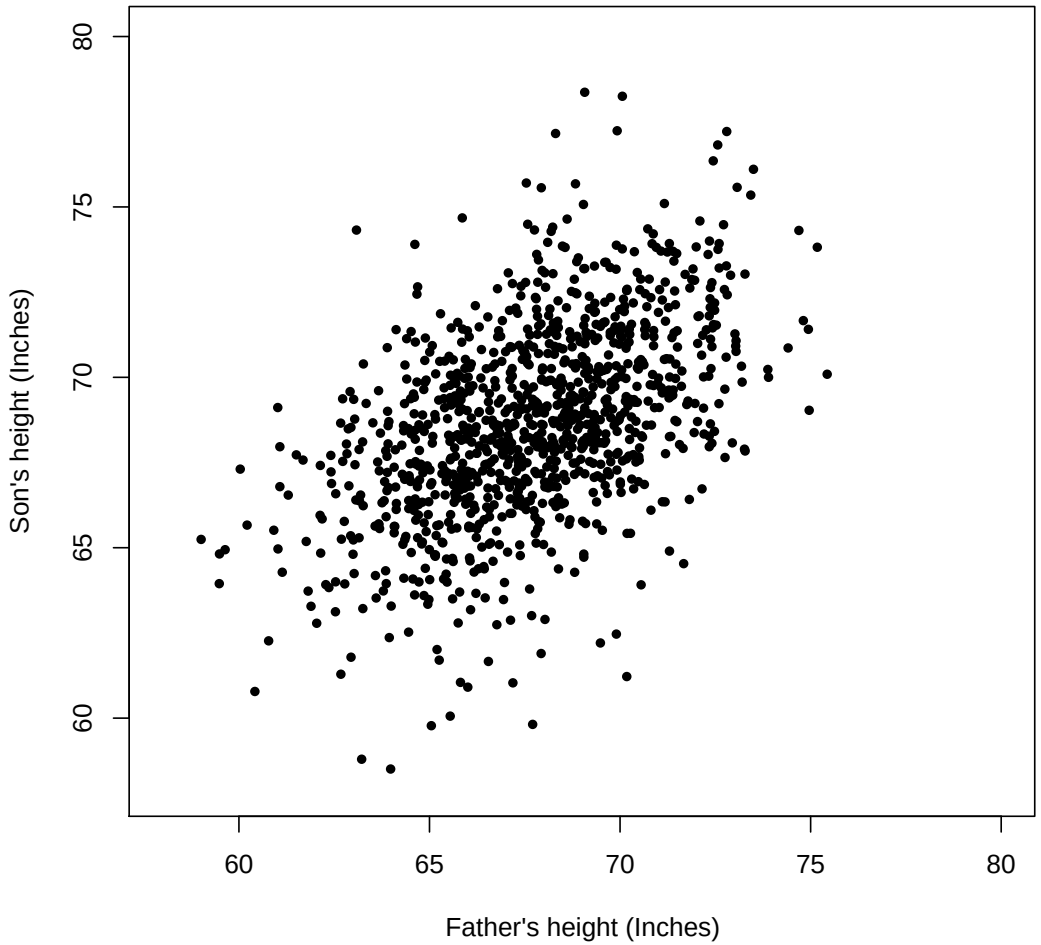
Pearson's height data

- Statisticians in Victorian England were fascinated by the idea of quantifying hereditary influences
- Two of the pioneers of modern statistics, the Victorian Englishmen Francis Galton and Karl Pearson were quite passionate about this topic
- In pursuit of this goal, they measured the heights of 1,078 fathers and their (fully grown) sons

The scatter plot

- As we've mentioned, it is important to plot continuous data – this is especially true when you have two continuous variables and you're interested in the relationship between them
- The most common way to plot the relationship between two continuous variables is the *two-way scatter plot*
- Scatter plots are created by setting up two continuous axes, then creating a dot for every pair of observations

Scatter plot of Pearson's height data



Observations about the scatter plot

- Taller fathers tend to have taller sons
- The scatter plot shows how strong this association is – there is a tendency, but there are plenty of exceptions

Standardizing a variable

- Before we summarize this relationship numerically, we must discuss the idea of *standardizing* a variable
- In Pearson's height data, one of the sons measured 63.2 inches tall
- Because the average height of the sons in the sample was 68.7 inches, another way of describing his height is to say that he was 5.5 inches below average
- Furthermore, because the standard deviation of the sons was 2.8 inches, yet another way of describing his height is to say that he was 1.9 standard deviations below the average

The standardization formula

- Putting this into a formula, we standardize an observation x_i by subtracting the average and dividing by the standard deviation:

$$z_i = \frac{x_i - \bar{x}}{SD_x}$$

  where $\bar{x}$ and SD_x are the mean and standard deviation of the variable x
- One virtue of standardizing a variable is interpretability:
  - If someone tells you that the concentration of urea in your blood is 50 mg/dL, that likely means nothing to you
  - On the other hand, if you are told that the concentration of urea in your blood is 4 standard deviations above average, you can immediately recognize this as a very high value

More benefits of standardization

- If you standardize all of the observations in your sample, the resulting variable will be "standardized" in the sense of having mean 0 and standard deviation 1
- Standardization therefore brings all variables onto a common scale – regardless of whether the heights were originally measured in inches, centimeters, or miles, the standardized heights will be identical
- As we will see momentarily, this allows us to study the relationship between two continuous variables without worrying about the scale of measurement
- The concept behind standardization – taking an observation, then subtracting the expected value and dividing by the variability – is fundamental to statistics and we will see variations on this idea many times in this course

The correlation coefficient

- The summary statistic for describing the strength of association between two variables is the *correlation coefficient*, denoted by r (and sometimes called Pearson's correlation coefficient)
- The correlation coefficient is always between 1 (perfect positive correlation) and -1 (perfect negative correlation), and can take on any value in between
- A *positive correlation* means that as one variable increases, the other one tends to increase as well
- A *negative correlation* means that as one variable increases, the other one tends to decrease

Calculating the correlation coefficient

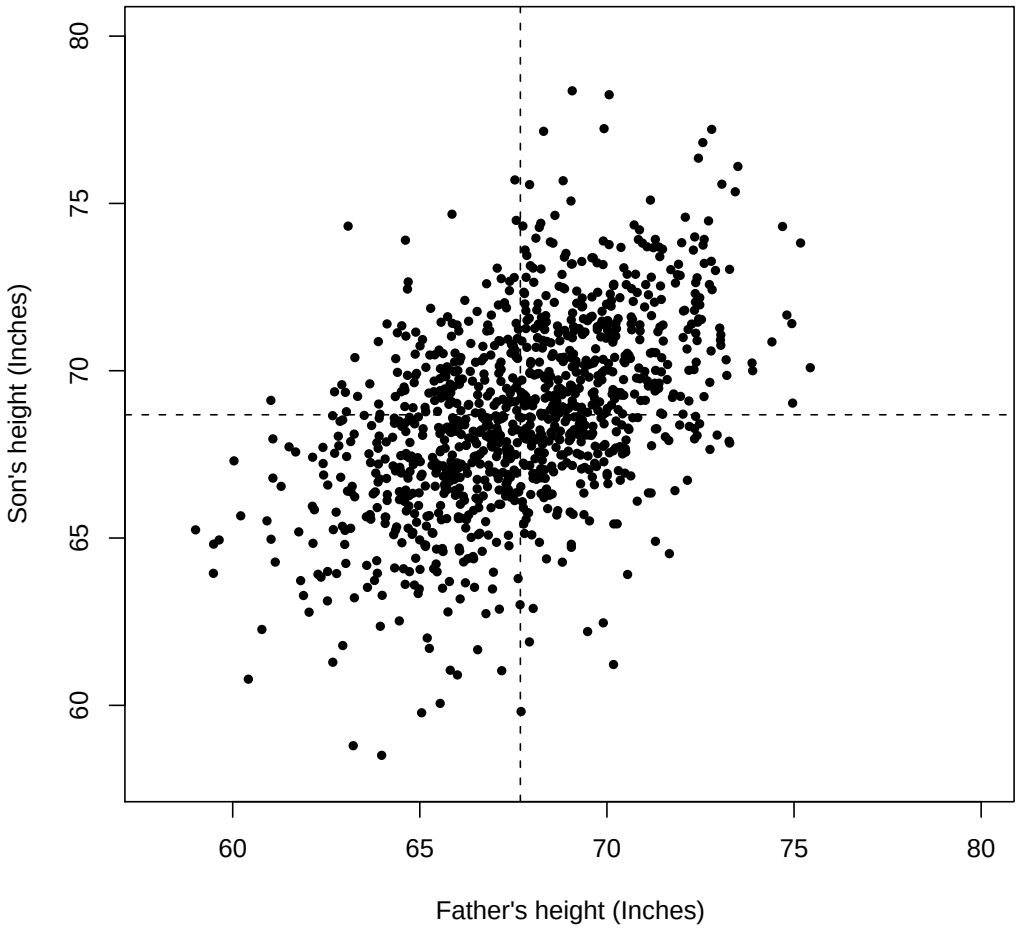
- The correlation coefficient is simply the average of the products of the standardized variables
- In mathematical notation,

$$r = \frac{\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i}{n-1},$$

where $\tilde{x}_i$ and $\tilde{y}_i$ are the standardized values of x and y (*i.e.*, the z_i's computed separately for variable x and variable y)

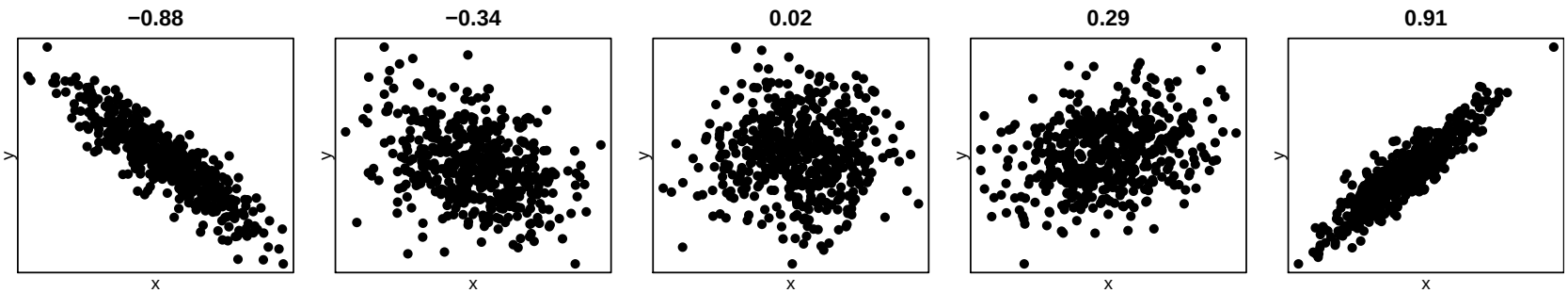
- Note: The "n versus $n-1$" issue has nothing to do with correlation; however, if $n-1$ is used when standardizing, it must be used again here

Meaning behind the correlation coefficient formula



For this data,
$r = 0.50$

The correlation coefficient and the scatter plot



More about the correlation coefficient

- Because the correlation coefficient is based on standardized variables, it does not depend on the units of measurement
- Thus, the correlation between father's and son's heights would be 0.5 even if the father's height was measured in inches and the son's in centimeters
- Furthermore, the correlation between x and y is the same as the correlation between y and x

Interpreting the correlation coefficient

- The correlation between heights of identical twins is around 0.93
- The correlation between income and education in the United States is about 0.44
- The correlation between a woman's education and the number of children she has is about -0.2
- When concrete physical laws determine the relationship between two variables, their correlation can exceed 0.9
- In the social sciences, this is rare – correlations of 0.3 to 0.7 are considered quite strong in these fields

Numerical summaries can be misleading!

From Cook & Swayne's *Interactive and Dynamic Graphics for Data Analysis*:

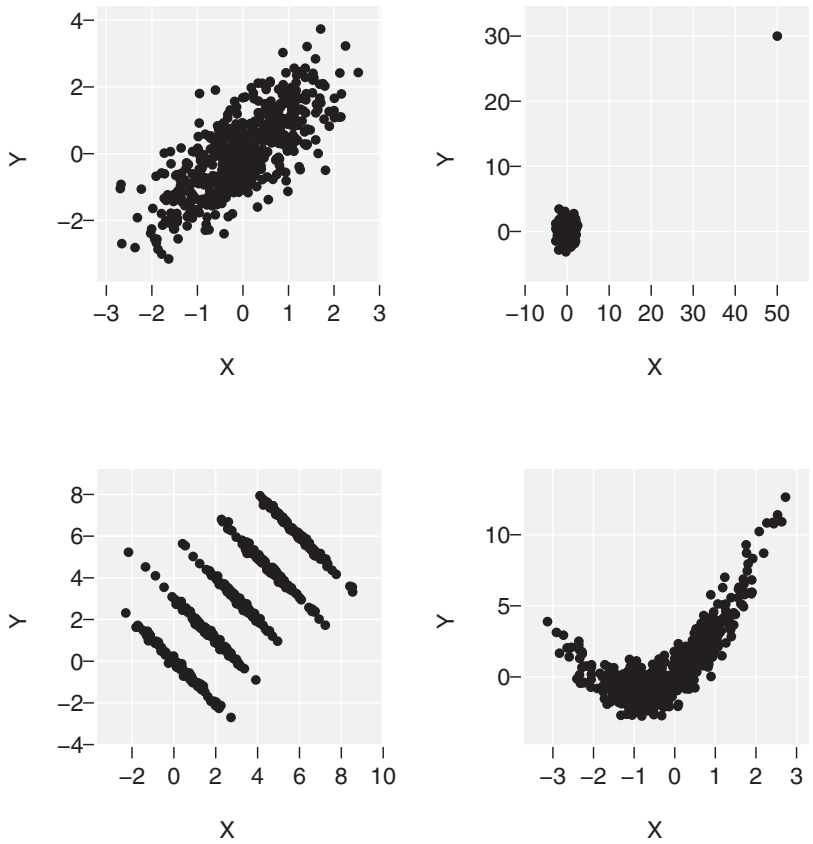


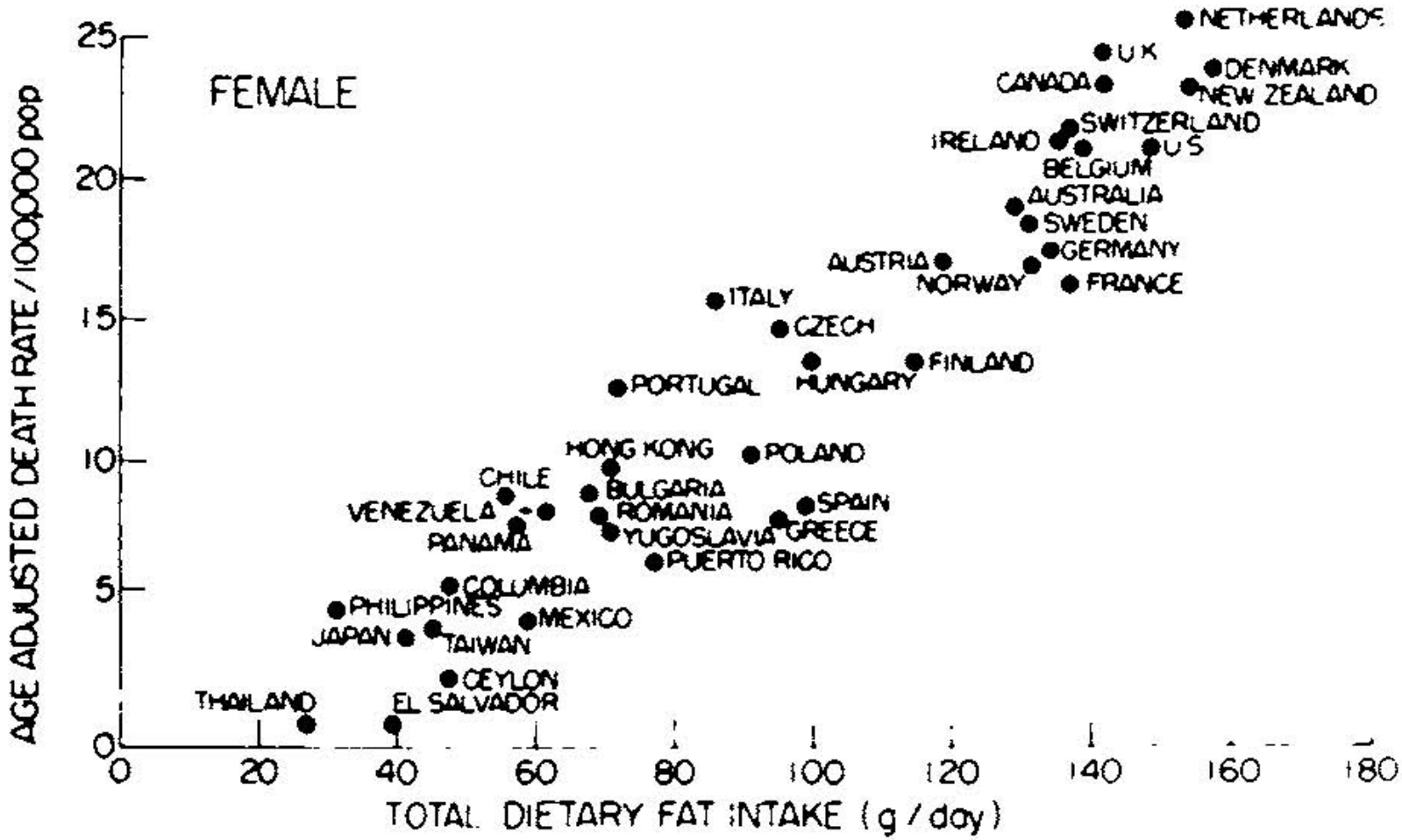
Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological correlations

- Epidemiologists often look at the correlation between two variables at the ecological level – say, the correlation between cigarette consumption and lung cancer deaths per capita
- However, people smoke and get cancer, not countries
- These correlations have the potential to be misleading
- The reason is that by replacing individual measurements by the averages, you eliminate a lot of the variability that is present at the individual level and obtain a higher correlation than there really is

Fat in the diet and cancer

From an article by Carroll in *Cancer Research* (1975):



Summary

- The standard way to display the relationship between two continuous variables is the scatter plot
- A standard summary statistic for this relationship is the correlation coefficient
- A standardized variable tells us how many standard deviations above/below the mean an observation is:

$$z_i = \frac{x_i - \bar{x}}{SD_x}$$

- Correlations at the ecological level are much higher than correlations at the individual level