Correlation

Patrick Breheny

November 17

Introduction

- Generally speaking, scientific questions often revolve around asking how changes in some quantity X affect some other quantity Y
- Thus far in this class, we've discussed studies in which both X and Y were categorical (contingency tables) and when X was categorical and Y was continuous (two-sample t-tests, Wilcoxon rank sum test, etc.)
- But what about when X is continuous?
- This requires a new way of thinking in terms of how we measure, describe, and model the relationship between X and Y

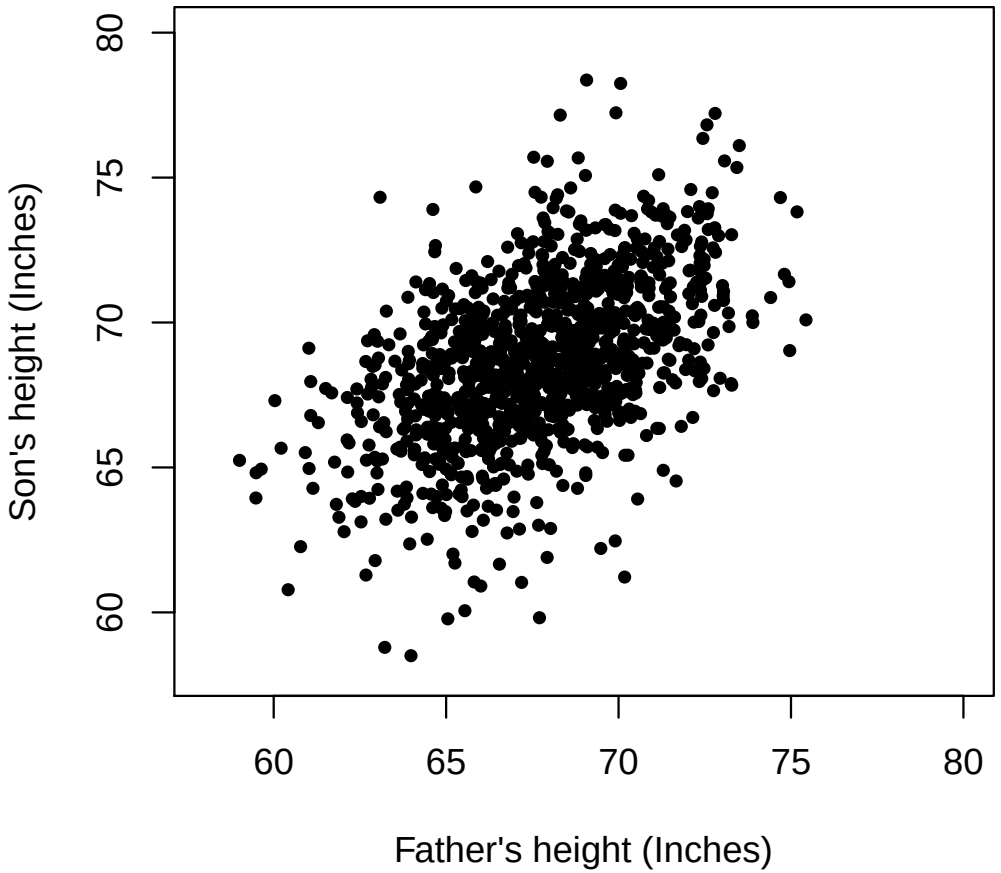
Pearson's height data

- 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

The correlation coefficient

- A simple, widely used summary statistic for describing the strength of association between two variables is the *correlation coefficient*, denoted by either r or $\hat{\rho}$ (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

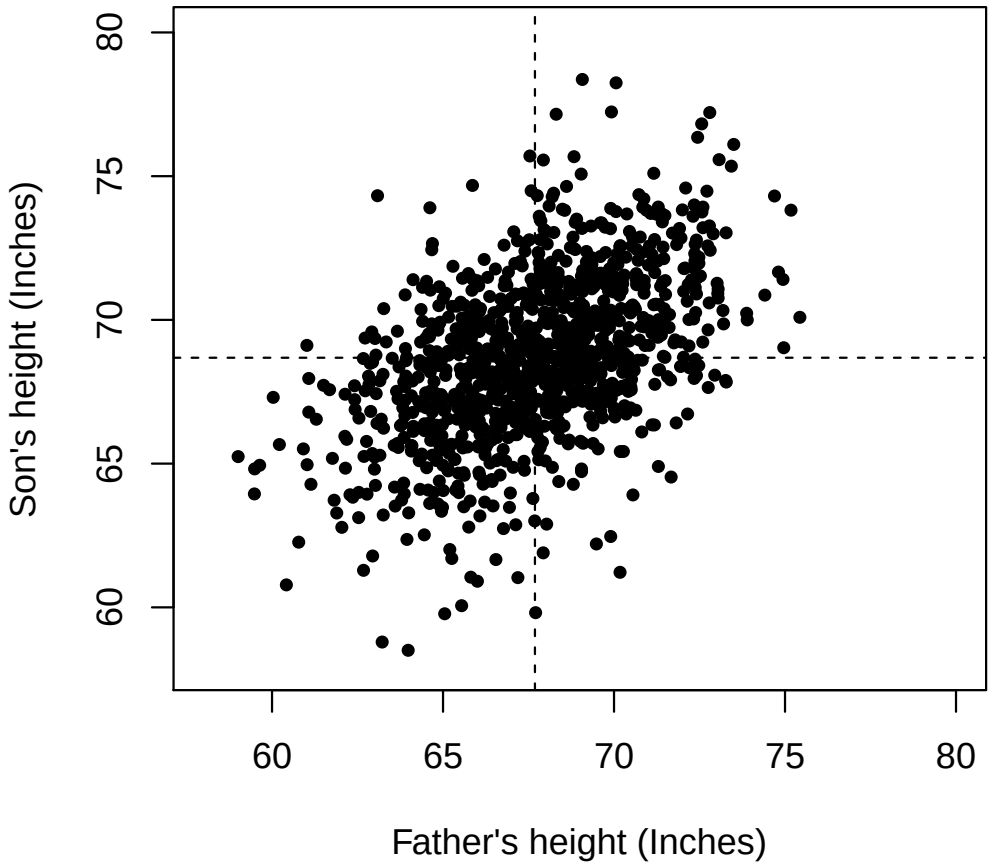
Calculating the correlation coefficient

- The correlation coefficient is simply the average of the products of the standardized variables
- In mathematical notation, letting $\{u_i\}$ denote the standardized values of $\{x_i\}$ and $\{v_i\}$ denote the standardized values of $\{y_i\}$,

$$\hat{\rho} = \frac{\sum_{i=1}^{n} u_i v_i}{n-1}$$

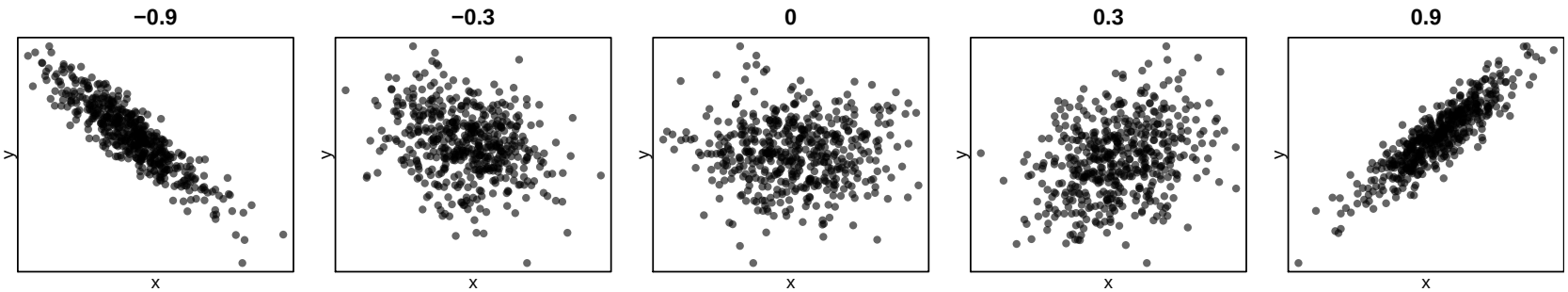
- Note: The $n-1$ in the denominator has nothing to do with correlation; if you use n for the standard deviations when standardizing, use an n in the denominator in the above equation; you just have to be consistent
- Note: By the Cauchy-Schwarz inequality, $\hat{\rho} \in [-1, 1]$

Intuition 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*:

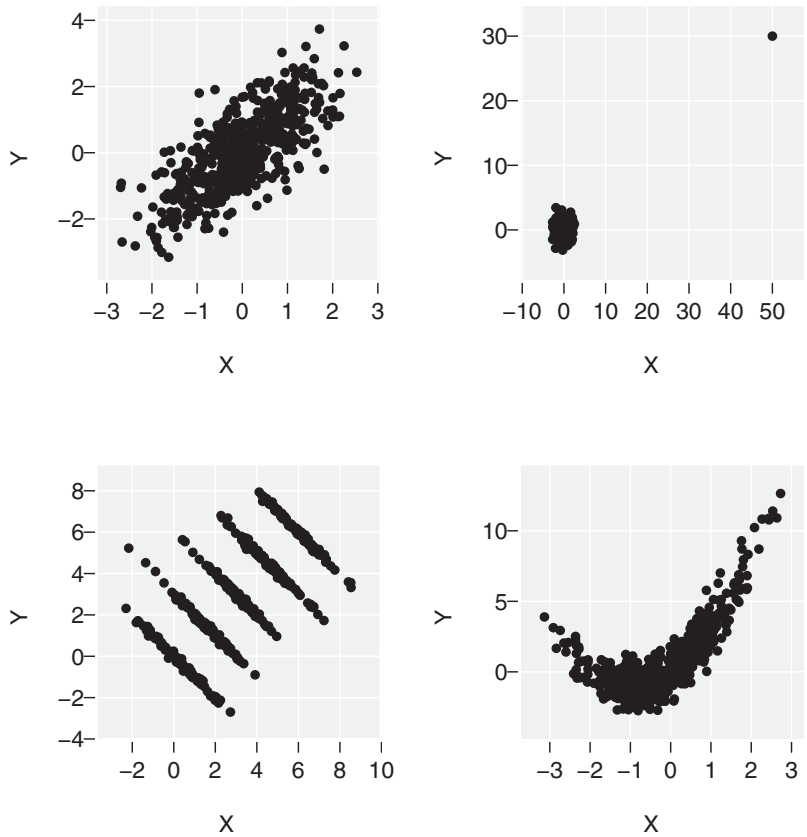


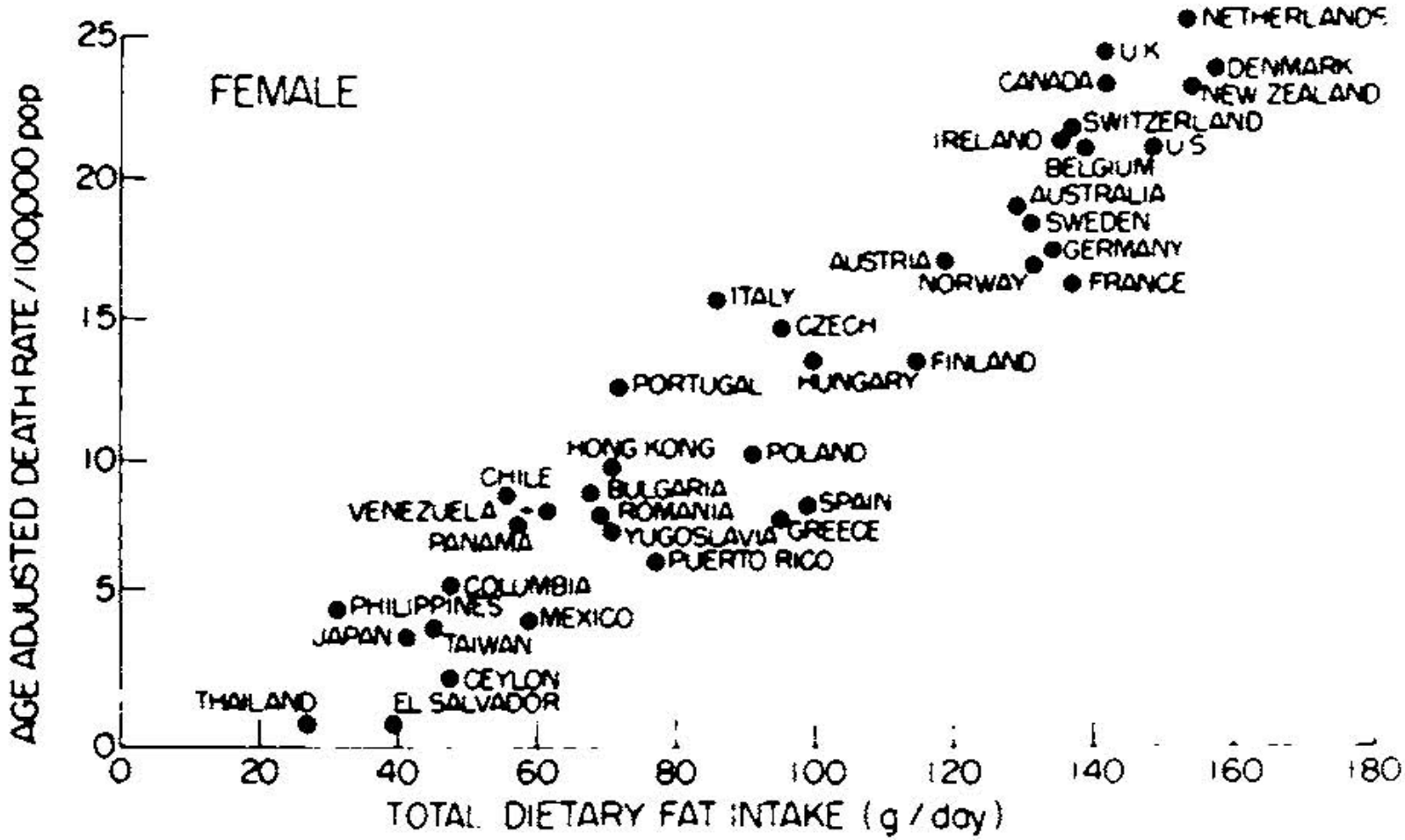
Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

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):



Approximate distribution of $\hat{\rho}$

- Let's turn our attention to inference concerning the population correlation coefficient, ρ:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

  where $\sigma_{xy} = \mathrm{E}[(X - \mu_x)(Y - \mu_y)]$, the *covariance* of X and Y
- It can be shown that $\hat{\rho}$ follows an approximate normal distribution with:

$$\mathrm{E}(\hat{\rho}) \approx \rho$$

$$\mathrm{Var}(\hat{\rho}) \approx \frac{(1 - \rho^2)^2}{n}$$

Possible approaches to inference

- Thus, somewhat reminiscent of the binomial distribution, the variance of $\hat{\rho}$ is smallest at the extremes (-1 and 1), and largest in the middle ($\rho = 0$)
- Thus, one approach to hypothesis testing and confidence interval construction would be to use the result on the previous slide, inverting the test as we did with the score interval in the binomial distribution to account for the fact that the variance of the estimator depends on the quantity we're estimating
- An alternative (and much more widely used) approach is to apply a clever transformation derived by Fisher that stabilizes the variance of our estimator, thereby constructing a pivotal quantity

Fisher's Z transformation

- The transformation Fisher proposed is the following:

$$f(\rho) = \operatorname{atanh}(\rho) = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$$

- **Theorem:** Suppose X and Y follow a bivariate normal distribution. Then

$$\frac{1}{2} \log \frac{1+\hat{\rho}}{1-\hat{\rho}} \dot{\sim} \mathrm{N}\left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right)$$

See our text for more details on the bivariate normal distribution

Variance-stabilizing transformations

- The remarkable thing about Fisher's transformation is that the resulting (approximate) distribution does not depend on ρ; its (approximate) variance is $1/(n-3)$ regardless of ρ
- Such a transformation – clearly desirable for the sake of constructing confidence intervals, but usually difficult to find – is known as a *variance-stabilizing* transformation
- Here, it allows us to construct a confidence interval for ρ by appropriately transforming the usual confidence interval for a standard normal random variable (i.e., ± 1.96 for a 95% confidence interval)

Confidence intervals for ρ

- So, for the height data, the $100(1-\alpha)$ confidence interval on the z-scale is

$$\operatorname{atanh}(\hat{\rho}) + z_{1-\alpha/2}\sqrt{\frac{1}{n-3}} = [0.491, 0.611]$$

- Transforming back to the original scale (by taking $\tanh$ of both ends of the interval) yields the interval $0.46, 0.55$
- Like the other transformed confidence intervals we have seen, the transformation introduces asymmetry in the resulting interval, although this is not apparent in the height example
- If we observed $\hat{\rho} = 0.9$ for a sample with $n = 10$, the Fisher Z-interval would be [0.62, 0.98]

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
- Correlations at the ecological level are much higher than correlations at the individual level
- The Fisher Z-transformation is a variance-stabilizing transformation that can be used to construct pivotal confidence intervals for the population correlation coefficient ρ