STAT:2010/4200 2019 Statistical Methods and Computing

Intro to Regression

Lecture 4
Jan. 28, 2019

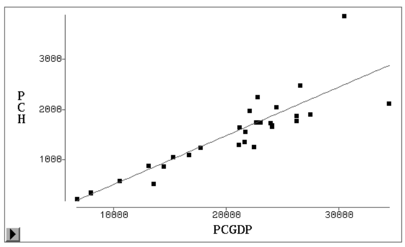
Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

Scatterplots

- represent the relationship between two different continuous variables measured on the same subjects
- each point represents the values for one subject for the two variables

Example: data reported by the Organization for Economic Development and Cooperation on its 29 member nations in 1998

- Per capita gross domestic product (a measure of wealth of the country) is on x-axis (horizontal)
- Per capita health care expenditures is on y-axis (vertical)



We can describe the overall pattern of a scatterplot by

- form or shape
- direction
- strength

Positive and negative association

- Two variables are *positively associated* when above-average values of one tend to occur in individuals with above-average values of the other, and below-average values of both also tend to occur together.
- Two variables are *negatively associated* when above-average values of one tend to occur in individuals with below-average values of the other, and vice-versa.

Linear relationships

- The form of a relationship shown by a scatterplot is linear if the points lie in a straight-line pattern.
- The linear relationship is strong if the points lie close to a line, with little scatter.

Example: per capita health care expenditures and gross domestic product

- "individuals" studied are countries
- form of relationship is roughly linear
- direction of relationship is positive
- strength: determined by how closely the points follow a clear pattern
  - quite strong

Correlation

- a numeric measure of the direction and strength of the linear relationship between two continuous variables measured on the same subjects
- terminology and notation
  - sample correlation coefficient r

Computing the sample correlation coefficient

- We have measured two different variables X and Y on the subjects in a study.
- There are n subjects.
- Let $\bar{x}$ and $\bar{y}$ be the sample means of the two variab les.
- Denote the sample standard deviation of the x variable as s_x and the sample standard deviation of the y variable as s_y.
- Then the sample correlation coefficient is computed as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Note that the first step in computing r is to *standardize* the measurements.
- Example: suppose X is heart rate in beats per minute and Y is body temperature in degrees Fahrenheit, and we have both heart rate and temperature measurements on $n = 10$ people.
  - The quantity

$$\frac{x_i - \bar{x}}{s_x}$$

    is the standardized heart rate for person i
    - how many standard deviations above or below the mean herat rate person i's heart rate is
  - Standardized values are no longer in their original units (e.g., the st andardized heart rates are not in beats per minute)

  - The sample correlation coefficient r is an average of the products of the standardized heart rates and temperatures for the 10 people.

Facts about correlation

- Correlation requires that both variables be quantitative, so that we can do arithmetic computations with them.
- r has no units, and, because it uses standardized values, it does not change when we change the units of measurements of x, y, or both.
  - For the same 10 people, r would not change whether we measured the heights and weights in inches and pounds or in centimeters and kilograms.
- $r > 0$ indicates a positive association between the two variable; $r < 0$ indicates a negative association
- r is always between -1 and +1
  - values of r near 0 mean a very weak linear relationship

  - values near +1 indicate a very strong positive relationship (all points lie almost exactly on a straight line)
  - values near -1 indicate a very strong negative relationship (all points lie almost exactly on a straight line)
- Correlation measures only the strength of *linear* relationships. r may be close to 0 even if the relationship between two variables is strong, if that relationship is curved.
- The sample correlation coefficient is very sensitive to outliers.
- A high correlation between two variables does not by itself imply a causal relationship.

Correlation and regression

- Correlation enables us to *assess the strength* of a linear relationship between two variables, but it does not enable us to *predict* the value of one variable for a subject for whom we know the value of the other variable.
- Prediction often is an important goal of statistical analysis.
- Example: we may wish to predict an infant's birthweight based on a laboratory measurement taken on the mother during pregnancy

Response variables and explanatory variables

- **response variable**
  - what we want to explain or predict
  - also called "dependent" or "outcome" variable
- **explanatory variable**
  - a variable that explains or influences differences in a response variable
  - also called "predictor" variables, "covariates," or "independent" variables
- When making a scatterplot of such data:
  - response variable goes on y-axis (vertical)
  - explanatory variable goes on x-axis (horizontal)

- Note: Correlation analysis does not distinguish between response and explanatory variables.

- Example: The admissions director of the University of Iowa wants to guess how successful incoming students are likely to be.
- The high school GPA is part of each incoming student's record. The admissions director wishes to predict the student's UI GPA.
- What is the response variable and what is the explanatory variable?

Simple Linear Regression

- If a scatterplot suggests a linear relationship between 2 variables, we want to summarize the relationship by drawing a straight line on the plot.
- A *regression line* summarizes the relationship between a response variable and an explanatory variable.
  - Both variables must be quantitative.
- definition: A *regression line* is a straight line that describes how a response variable Y changes as an explanatory variable X changes.
  - often used to predict the value of Y that corresponds to a given value of X.

Recall straight lines

$$y = a + bx$$

- a : intercept; the value of Y when X = 0
- b : slope; how much Y changes when X increases by 1 unit

Least squares: choosing the "best" estimated line for a set of sample data

a and b are estimated by choosing a line as follows:

- for each observed value y_i in the sample data, compute the distance from y_i to the line
- square each of the distances
- add up all the squared distances
- choose the line that makes the sum of these squared distances the smallest

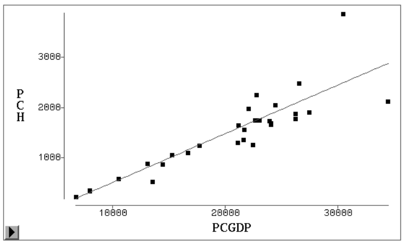
Using sample data to estimate the intercept and slope

- We will write an estimated regression line based on sample data as

$$\hat{y} = a + bx$$

- a is the estimated intercept, and b is the estimated slope
- The hat over the y means that $\hat{y}$ is the *predicted* value of the response variable, not an actual observed value

- Example: the estimated regression line for the health care expenditures and gross domestic product is



$$\hat{y} = -465.7 + 0.0968x$$

- This means that if country A has 1 unit higher PCGDP than country B, we would expect country A to have 0.0968 higher PCH than country B.
  - Since we are measuring PCH in dollars and PCGDP in dollars, this means for every additional dollar in PCGDP, we expect about a 9.7-cent increase in PCH.

- Note that it makes no sense in this problem to say that the intercept (-465.7) is the amount of per capital health care expenditure that we would expect in a country with PCGDP = 0.
- **An estimated regression line is meaningful only for the range of X values actually observed.**
  - In the PCH/PCGDP problem, this is about $8000 - 33000. The estimated intercept makes the linear relationship come out right over this range of X values.