Lecture 03. Statistics and probability theory.

Complement

- skewness:
- making boxplots, exercise 4/2
- a uniform distribution in an interval, exercise 5/2

Relationships between variables

Association between variables
Two variables measured on the same cases are associated if knowing the value of one of the variables tells you something about the values of the other variable that you would not know without this information.

Examining relationships
When you examine the relationship between two or more variables, first ask the following preliminary questions:

- What individuals or cases do the data describe?
- What variables are present? How are they measured?
- Which variables are quantitative and which are categorical?

Response variable, explanatory variable
A response variable measures an outcome of a study. An explanatory variable explains or causes changes in the response variables.

Scatterplot
A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.
Example
Example 1 Coffee in Starbuck, prizes of three sizes

Tall	Grande	Venti
3.15	3.65	4.15

Example 2 Length of life of breeds (on average in years)

poodles	Great Danes	Irish wolfhounds
9.3	4.6	4.15

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Exercise

Examining scatterplot
In any graph of data, look for the overall pattern and for striking deviations from that pattern. You can describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship. An important kind of deviation is an outlier, an individual value that falls outside the overall pattern of the relationship.

Positive association, negative association
Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and below- average values also tend to occur together. Two variables are negatively associated when above-average values of one accompany below-average values of the other, and vice versa.

2.1 Summary

To study relationships between variables, we must measure the variables on the same group of individuals or cases. If we think that a variable x may explain or even cause changes in another variable y, we call x an explanatory variable and y a response variable. A scatterplot displays the relationship between two quantitative variables. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph. Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis. Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot. In examining a scatterplot, look for an overall pattern showing the form, direction, and strength of the relationship, and then for outliers or other deviations from this pattern.

- Form: Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and clusters are other forms to watch for.

- Direction: If the relationship has a clear direction, we speak of either positive association (high values of the two variables tend to occur together) or nega- tive association (high values of one variable tend to occur with low values of the other variable).

- Strength: The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

To display the relationship between a categorical explanatory variable and a quantitative response variable, make a graph that compares the distributions of the response for each category of the explanatory variable.

2.2 Correlation

A scatterplot displays the form, direction, and strength of the relationship between two quantitative variables. Linear (straight-line) relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relationship is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a relationship is. The two scatterplots in Figure 2.9

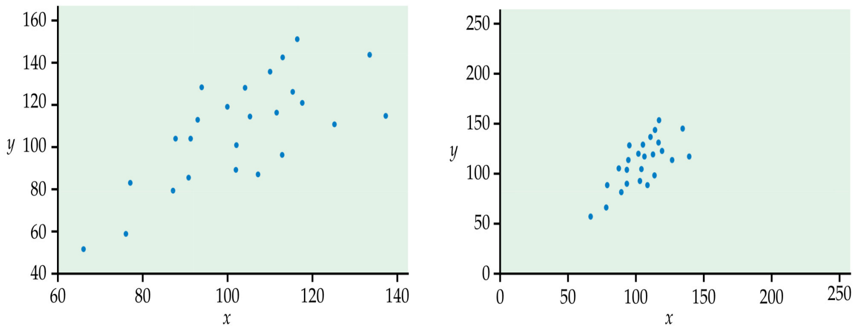

FIGURE 2.9 Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

depict exactly the same data, but the plot on the right is drawn smaller in a large field. The plot on the left seems to show a stronger relationship. Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot. We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. Correlation is the measure we use.

The correlation measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r. Suppose that we have data on variables x and y for n individuals. The means and standard deviations of the two variables are x and s_x for the x-values, and y and s_y for the y-values. The correlation r between x and y

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

Comments

Let you note that after the linear transformation $x_{new} = a + bx$ the mean value and standard deviation of the resulting sequence are as follows

$$\begin{aligned} \bar{x}_{new} &= a + b\bar{x} \\ s_{x_{new}} &= |b| \cdot s_x \end{aligned}$$

As a consequence, after the transformation $x_{new} = \frac{x-\bar{x}}{s_x}$ the resulting sequence has mean value $\bar{x}_{new} = 0$ and the standard deviation $s_{x_{new}} = 1$, hence this transformation is called the standardization of data set.

The scatterplots in Figure 2.10 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the essential meaning of r clear, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the standardized values of the variables and therefore cannot change the correlation.

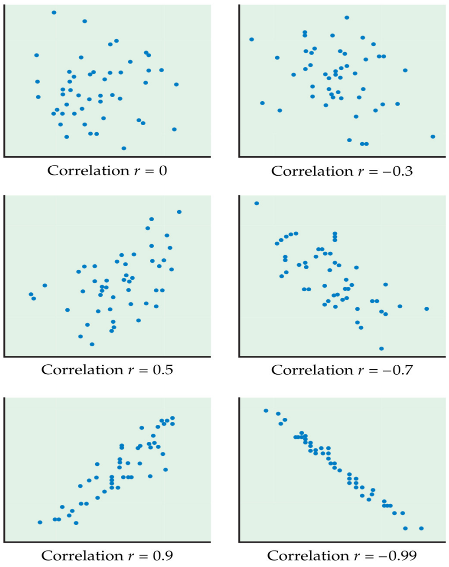

FIGURE 2.10 How the correlation r measures the direction and strength of a linear association.

2.2 Summary

Although you can calculate a correlation for any scatterplot, r measures only linear relationships. Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association. Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points lie exactly on a straight line. Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

Regression line

Example

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Fitting the straight line to the data set

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form $y = a + bx$. In this equation, b is

the slope, the amount by which y changes when x increases.
To find coefficients a and b, we consider a function

$$L(a,b) = \sum_{i=1}^{n} (y_i - bx_i - a)^2.$$

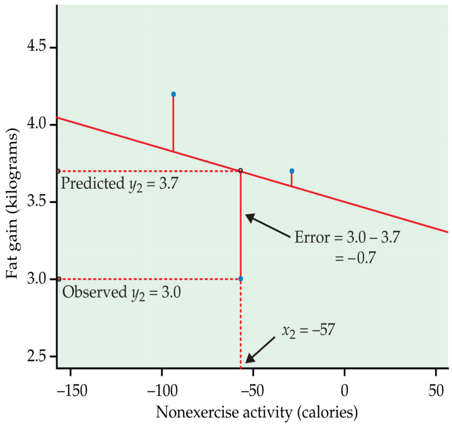

FIGURE 2.13 The least-squares idea: make the errors in predicting *y* as small as possible by minimizing the sum of their squares.

We determine the minimum of that function using tools from mathematical analysis

$$\begin{cases} 0 = \frac{\partial L}{\partial a} = -2\sum\limits_{i=1}^{n}(y_i - bx_i - a) \\ 0 = \frac{\partial L}{\partial b} = -2\sum\limits_{i=1}^{n}(y_i - bx_i - a)x_i \end{cases}$$

$$\begin{cases} \left(\sum\limits_{i=1}^{n} x_i\right) b + na = \sum\limits_{i=1}^{n} y_i \\ \left(\sum\limits_{i=1}^{n} x_i^2\right) b + \left(\sum\limits_{i=1}^{n} x_i\right) a = \sum\limits_{i=1}^{n} x_i y_i \end{cases}$$

$$W = \det \begin{bmatrix} \sum_{i=1}^{n} x_i & n \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \end{bmatrix} = n^2\bar{x}^2 - n\sum_{i=1}^{n} x_i^2$$

$$W_1 = \det \begin{bmatrix} \sum_{i=1}^{n} y_i & n \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \end{bmatrix} = n^2\bar{x}\bar{y} - n(\sum_{i=1}^{n} x_i y_i)$$

$$W_0 = \det \begin{bmatrix} \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \end{bmatrix} = n\bar{x}\sum_{i=1}^{n} x_i y_i - n\bar{y}\sum_{i=1}^{n} x_i^2$$

$$b = \frac{W_1}{W} = r\frac{s_y}{s_x}, \quad a = \frac{W_0}{W} = \bar{y} - b\bar{x}$$

Let you remember that the regression line depends on the outliers.

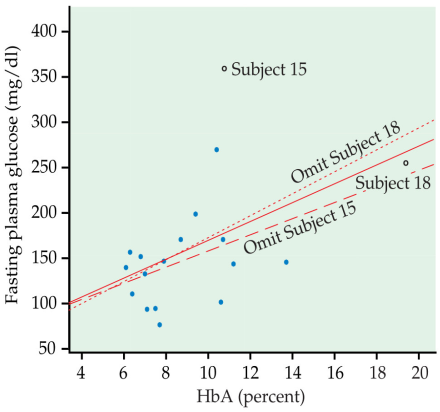

FIGURE 2.24 Three regression lines for predicting FPG from HbA, for Example 2.22. The solid line uses all 18 subjects. The dotted line leaves out Subject 18. The dashed line leaves out Subject 15. “Leaving one out” calculations are the surest way to assess influence.