Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 2: Looking at Data——Relationships

2.1 Relationships

Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

A **response variable** measures an outcome of a study.

An **explanatory variable** explains or causes changes in the response variable.

Certain characteristics of a data set are key to exploring the relationship between two variables. These should include the following:

- Cases
- Label
- Categorical or quantitative
- Values
- Explanatory or response

2.2 Scatterplots

The most useful graph for displaying the relationship between two **quantitative variables** is a **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.

How to Make a Scatterplot

- Decide which variable should go on each axis. If a distinction exists, plot the **explanatory variable** on the x axis and the **response variable** on the y axis.
- Label and scale your axes.
- Plot individual data values.

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice-versa.

2.3 Correlation

The **correlation** r measures the strength of the linear relationship between two quantitative variables. Using the notation explained on pp. 103–104 in the text:

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

Properties of Correlation:

- r is always a number between -1 and 1.
- $r > 0$ indicates a **positive** association.
- $r < 0$ indicates a **negative** association.
- Values of r near 0 indicate a very **weak linear relationship**.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.
- **Correlation** makes no distinction between explanatory and response variables.
- r has no units and does not change when we change the units of measurement of x, y, or both.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation r is always a number between -1 and 1.
- Correlation requires that **both variables be quantitative**.
- Correlation does **not describe curved relationships** between variables, no matter how strong the relationship is.
- The correlation r is **not resistant**; it can be strongly affected by a few outlying observations.
- Correlation is **not a complete summary** of two-variable data.

2.4 Least-Squares Regression

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We can use a regression line to **predict** the value of y for a given value of x.

Regression equation:

$$\hat{y} = b_0 + b_1 x$$

- x is the value of the **explanatory variable**.
- $\hat{y}$ is the **predicted value** of the response variable for a given value of x.
- b_1 is the **slope**, the amount by which y changes for each oneunit increase in x.
- b_0 is the **intercept**, the value of y when $x = 0$.

Least-Squares Regression Line (LSRL):
The least-squares regression line of y on x is the line that minimizes the sum of the squares of the vertical distances of the data points from the line.

The **square of the correlation**, r^2, is the fraction of the variation in values of y that is explained by the least-squares regression of y on x.

- r^2 is called the **coefficient of determination**.

2.5 Cautions About Correlation and Regression

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted} y \\ &= y - \hat{y} \end{aligned}$$

Outliers in the x direction are often **influential** for the least-squares regression line, meaning that the removal of such points would **markedly change the equation of the line**.

Cautions About Correlation and Regression:

- Both describe **linear** relationships.
- Both are affected by **outliers**.
- Always plot the data **before interpreting**.
- Beware of **extrapolation**.
  - Use caution in predicting y when x is outside the range of observed x's.
- Beware of **lurking variables（潜在变量）**.
  - These have an **important effect** on the relationship among the variables in a study, but are not included in the study.
- **Correlation** does **not** imply **causation（因果关系）**!

2.7 The Question of Causation

Association, however **strong**, does **NOT** imply **causation**.

We can evaluate the association using the following **criteria**:

- The association is **strong**.
- The association is **consistent**.
- Higher **doses（剂量）** are associated with stronger responses.
- **Alleged cause（声称的原因）** precedes the effect.
- The alleged cause is **plausible（合理的）**.