CP Statistics – Chapter 3 Notes: Modeling Linear Relationships

3.1: Scatterplots and Correlation

Explanatory and Response Variables
A **response variable** measures an outcome of a study. An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is sometimes referred to as the ***independent*** variable and is typically symbolized by the variable ***x***. The response variable is sometimes referred to as the ***dependent*** variable and is typically symbolized by the variable ***y***.

Scatterplot
A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of the explanatory variable appear on the horizontal axis, and the values of the response variable appear on the vertical axis. If there is no clear explanatory/response relationship between the two variables, then either variable can be placed on either axis. Each individual in the data set appears as a single point in the plot fixed by the values of both variables for that individual.

Examining a Scatterplot
In any graph of data, look for patterns and deviations from the pattern. Describe the overall **pattern** of a scatterplot by the **form**, **direction** and **strength** of the relationship.

- **Form** can be described as **linear** or **curved**.
- **Direction** can be described as **positive** or **negative** or **neither**.
- **Strength** can be described as **weak**, **moderate** or **strong**.

A **deviation** from the overall pattern of a scatterplot is called an **outlier**.

Association

- Two variables are **positively associated** if as one increases the other increases.
- Two variables are **negatively associated** if as one increases the other decreases.

Correlation
Correlation measures the strength and direction of the relationship between two quantitative variables. Correlation is usually represented by the letter ***r***.

Facts about Correlation

1. When calculating correlation, it makes no difference which variable is *x* and which is *y*.
2. Correlation is only calculated for quantitative variables, not categorical.
3. The value of *r* does not change if the units of *x* and/or *y* are changed.
4. Positive *r* indicates a positive association between *x* and *y*. Negative *r* indicates a negative association.
5. Correlation is always a number between -1 and +1. Values close to +1 or -1 indicate that the points lie close to a line. The extreme values of +1 and -1 are only achieved when the points are perfectly linear.
6. Correlation measures the strength of a linear relationship between two variables, not curved relationships.
7. Correlation, like the mean and standard deviation, is nonresistant. Recall that this means that it is greatly affected by outliers.

3.2: Least-Squares Regression

Regression Line
A **regression line** is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes. The line is often to predict values of y for given values of x. Regression, unlike correlation, requires an explanatory/response relationship. In other words, when *x* and *y* are reversed, the regression line changes. Recall that correlation is the same no matter which variable is *x* and which is *y*.

Least-Squares Regression Line
The **least-squares regression line** is the line that makes the sum of the squares of the vertical distances from the data points to the line as small as possible.

Equation of the Least-Squares Regression Line
To find the equation of the regression line in the form $y = a + bx$, where *a* is the *y*-intercept and *b* is the slope, use the following equations:

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Outliers and Influential Points
A point that lies outside the overall pattern of the other observations is considered an **outlier**. If the removal of such a point has a large effect on the correlation and/or regression, that point is considered an **influential point**.

Interpreting Slope and Intercept
The **slope** of the linear regression line represents the average amount of change in the *y*-variable that can be expected for each increase of one unit of the *x*-variable.

The **intercept** of the linear regression line represents the expected amount that the *y*-variable will be when the *x*-variable is at 0.

In general terms, the *y*-intercept represents a starting point (when $x = 0$) and the slope represents a rate of change (the change in *y* per unit of *x*).

For example, suppose the regression equation $y = 30.4 - 0.72x$, where *y* is the temperature reading on a heat sensor in degrees Celsius and *x* is the distance in centimeters from the sensor to a heat source.

The *y*-intercept (30.4) tells us that the temperature reading when the heat sensor is 0 cm away from the heat source is approximately 30.4 degrees Celsius.

The slope (-0.72) tells us that, on average, the temperature reading decreases0.72 degrees Celsius for every centimeter that the heat sensor is moved away from the heat source.

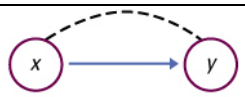
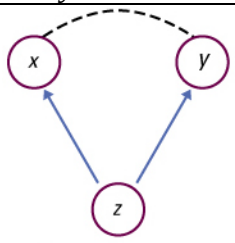
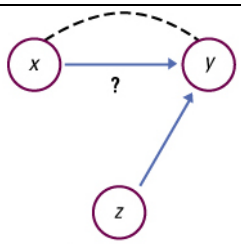
4.3: Establishing Causation

Lurking Variable

A ***lurking variable*** is a variable which is not among the variables of a study and yet may influence the interpretation of the relationships among those variables. For example, consider the statistical relationship between ice cream sales and drowning deaths. These two variables have a positive, and potentially statistically significant, correlation with each other. One might be tempted to conclude that more ice cream sales *cause* more drowning deaths to occur. The real cause of a corresponding increase in both of these variables is a lurking variable – *warm weather*. People eat more ice cream *and* go swimming more when it is warm.

The Question of Causation

An association between two variables doesn't necessarily mean that *x causes y*. Consider 3 possibilities in this situation and diagrams of the relationship between variables. In the diagrams below, *z* represents a *lurking variable*.

Causation – *x* does indeed cause *y*	**Common Response** – a 3rd variable, *z*, causes changes in both *x* and *y*	**Confounding** – both *x* and a 3rd variable, *z*, cause changes in *y*
 Causation	 Common response	 Confounding

Also consider that sometimes it is believed that *x causes y*, when it is actually the other way around (*y causes x*). Call this **Reverse Causation.**

Confounding – definition

Two variables are said to be confounded when their effects on a response variable cannot be distinguished from each other.

Examples

1. You observe that when more people wear coats there are also more people with colds. Does wearing a coat cause the common cold? No – weather is a ***common response*** lurking variable here. *Cold weather* causes more people to wear coats and get colds.
2. A study looks at the effects of after school tutoring and finds that students who get tutoring fail classes at a higher rate than those who do not go to tutoring. So does tutoring just cause students to do worse? No – this is ***reverse causation***. Students go to tutoring because they are doing poorly in class, not the other way around.
3. People who drink large amounts of alcohol tend to die earlier than those who do not. Does alcohol *cause* an early death? Certainly, but other factors are to blame as well. People who drink a lot probably also have other harmful behaviors such at poor diet, which can also lead to early death. So both behaviors have causal links to early death, but how much does each contribute? Hard to say, since these variables are ***confounded***.