Lesson 6. What is a Statistical Model?

1 Overview

- A **statistical model** is a mathematical representation of the relationships among random variables
- Some purposes of statistical modeling:
  1. Making predictions
     - e.g., predicting the price of a car based on its age, mileage, and model
  2. Understanding relationships
     - e.g., after taking mileage into account, how is the age of a car related to is price?
  3. Testing differences
     - e.g., is the rate of headache relief for migraine sufferers who take a new medicine sufficiently higher than those in the control group?

2 Basic terminology

	Definition
observational study	The people, objects, or cases on which data are recorded.
variables	The characteristics measured or recorded about each observational unit.
quantitative variable	Variable that records numbers (suitable for arithmetic) about the observational units.
categorical variable	Variable that records a category designation about the observational units.
response variable	Variable that measures the outcome of interest. Also known as the **dependent variable**.
explanatory variables	Variables whose relationship to the response variable is being studied. Also known as **predictors, predictor variables, independent variables**.
population	The group we want to make a statement about. The entire pool from which the sample is drawn.
parameter	A characteristic about the population.
sample	The collected data, gathered from a subset of the population.
statistic	A characteristic of the sample.

Example 1. You are interested in whether a midshipman's political inclination and GPA help predict his or her major. So you collect a sample of 50 mids, record each one's political inclination, GPA, and major, and analyze the data.

a. What is the population of interest?
b. Identify the response variable and the explanatory variables, and for each one indicate whether it is categorical or quantitative.
c. If you find that in your sample of 50 mids, the average GPA is 2.8, is 2.8 a parameter or a statistic?

3 Statistical models, formally

- **Population-level model** – the "true" but unknown model

$$Y = f(X_1, \dots, X_k) + \varepsilon$$

  - The **error** ε is the part of the response variable Y that remains unexplained after considering the predictors $X_1, \dots, X_k$

- **Fitted model** – the model estimated from sample data

$$\hat{Y} = \hat{f}(X_1, \dots, X_k)$$

4 The nature of statistical models

- Statistical models are simplifications of reality
  - Statistical models are not deterministic – their predictions are not expected to be perfectly accurate

    e.g., relationship between degrees F and degrees C is deterministic
    relationship between height and weight is statistical
  - Statistical models aim to explain as much variability as possible, given the data at hand
- Even though there's randomness and uncertainty, we can still get meaningful results
  - We will quantify how confident we are in those results
  - "All models are wrong, but some are useful." —George Box, statistician