Statistical Modeling

Purpose

Statistical modeling is a mathematical technique used to verify and quantify associations between one or more quantitative and/or qualitative *predictor* variables (x_1, x_2, …), and a single quantitative or qualitative *response* variable (y), or multiple multivariate normal response variables (y_1, y_2, …).

E.g., the association between income (x_1), whether or not someone at home cooks (x_2), and the number of dinners in the last k eaten outside the home (y).

Components

- Probability Model: $f(y, \theta)$

  Discrete: Bernoulli, Binomial, Poisson, Multinomial

  Continuous: Normal, Weibull, Multivariate Normal

Probability Models

Suppose there is a 6 week experiment with 15 animals in treatment group A and 15 animals in treatment group B. Consider the following measurements on each animal:

- Whether or not there were malignant tumors.
- The number of tumors that were malignant.
- The number of tumors.
- The average size of the tumors.
- The time to the first tumor.
- The number of tumors that were malignant, benign, or other.
- The average size and average weight of the tumors.

The corresponding probability models are Bernoulli, Binomial, Poisson, Normal, Weibull, Multinomial, and Multivariate Normal.

Bernoulli Trials

- The basis for the probability models we will examine in this chapter is the Bernoulli trial.

- We have Bernoulli trials if:
  - there are two possible outcomes (success and failure).
  - the probability of success, p, is constant.
  - the trials are independent.

The Geometric Probability Model

- A Geometric probability model tells us the probability for a random variable that counts the number of Bernoulli trials until the first success.

- You may not know in advance the number of trials needed.

- Geometric models are completely specified by one parameter, p, the probability of success, and are denoted Geom(p).

The Geometric Probability Model

Geometric model for Bernoulli trials: Geom(p)

p = probability of success

q = 1 – p = probability of failure

X = number of trials until the first success occurs

$$P(X = x) = q^{x-1}p$$

$$E(X) = \mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{q}{p^2}}$$

Independence

- When we don't have an infinite population, the trials may not be independent. But, there is a rule that allows us to pretend we have independent trials:
  - The 10% condition: Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.

The Binomial Model

- A Binomial model tells us the probability for a random variable that counts the number of successes in a fixed number of Bernoulli trials.

- Two parameters define the Binomial model: n, the number of trials; and, p, the probability of success. We denote this Binom(n, p).

The Binomial Model (cont.)

- In n trials, there are

$$_nC_k = \frac{n!}{k!(n-k)!}$$

ways to have k successes.

  - Read $_nC_k$ as "n choose k."

- Note: $n! = n \times (n-1) \times \ldots \times 2 \times 1$, and $n!$ is read as "n factorial."

The Binomial Model (cont.)

Binomial model for Bernoulli trials: Binom(n,p)

n = number of trials

p = probability of success

q = 1 – p = probability of failure

X = number of successes in n trials

$$P(X = x) = \binom{n}{x} p^x q^{n-x} \text{ where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$$\mu = np \qquad \sigma = \sqrt{npq}$$

The Normal Model

- Success/failure condition: A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures:
    - $np \geq 10$ and $nq \geq 10$.
- As long as the Success/Failure Condition holds, we can use the Normal model to approximate Binomial probabilities.
- The parameters of the normal model are
    - $\mu = np$ and $\sigma = \text{sqrt}(npq)$

Continuous Random Variables

- When we use the Normal model to approximate the Binomial model, we are using a continuous random variable to approximate a discrete random variable.
- So, when we use the Normal model, we no longer calculate the probability that the random variable equals a particular value, but only that it lies between two values.

The Poisson Model

- The Poisson probability model approximates the Binomial model when the probability of success, p, is very small and the number of trials, n, is very large.

- The parameter for the Poisson model is λ. To approximate a Binomial model with a Poisson model, just make their means match: $\lambda = np$.

The Poisson Model (cont.)

Poisson probability model for successes: Poisson(λ)

λ = mean number of successes

X = number of successes

e is approximately 2.71828

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$E(X) = \lambda \qquad SD(X) = \sqrt{\lambda}$$

What Can Go Wrong?

- Be sure you have Bernoulli trials.
  - You need two outcomes per trial, a constant probability of success, and independence.
  - Remember that the 10% Condition provides a reasonable substitute for independence.
- Don't confuse Geometric and Binomial models.
- Don't use the Normal approximation with small n.
  - You need at least 10 successes and 10 failures to use the Normal approximation.

What have we learned?

- Bernoulli trials show up in lots of places.
- Depending on the random variable of interest, we might be dealing with a
  - Geometric model
  - Binomial model
  - Normal model
  - Poisson model

What have we learned? (cont.)

- Geometric model
  - When we’re interested in the number of Bernoulli trials until the next success.
- Binomial model
  - When we’re interested in the number of successes in a certain number of Bernoulli trials.
- Normal model
  - To approximate a Binomial model when we expect at least 10 successes and 10 failures.
- Poisson model
  - To approximate a Binomial model when the probability of success, p, is very small and the number of trials, n, is very large.