Lecture 1

Monday, May 6, 2019 3:00 PM

Linear regression is one of many possible approaches to statistical modeling. The goal is to establish a relationship between several variables.

target, dependent

→ Response variable: this is the variable we care most about (y)

covariates, predictors, features, independent

→ Explanatory variables: these are the variables that the response depends on. ($x_1, x_2, \dots, x_p$)

There exists some true relationship:

$$y = f(x_1, x_2, \dots, x_p) \qquad [1]$$

but it's unknown. So, we use a linear regression model to approximate it.

The relationship the linear regression imposes is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \qquad [2]$$

where:

- y is the value of the response variable
- x_j is the value of explanatory variable $j = 1, 2, \dots, p$
- β_j is a parameter quantifying the infuence of x_j on y (β_0 is an intercept of the expected response)
- ε is a random error term which acknowledges that [2] without it does not equal [1]

  * we assume that $\varepsilon \sim N(\underset{E[\varepsilon]}{0}, \underset{Var[\varepsilon]}{\sigma^2})$

* we assume the explanatory variables are fixed and not random.

Given these assumptions equation [2] can be thought of as being composed of two parts:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}_{\text{"}\mu\text{" = deterministic}} + \underbrace{\varepsilon}_{\text{random}}$$

$$= \mu + \varepsilon$$

Consequently y is assumed to be a normal r.v. with mean $E[y] = \mu = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ and variance $Var[y] = \sigma^2$

Some of these assumptions might be unrealistic, so its important that we check them and modify our approach if they don't seem valid in light of the observed data.

A note on terminology: our regression equation is referred to as "linear" because it's linear in the parameters. It does not have to be linear in the x's. To check, observe whether partial derivatives of μ depend on the β's.

Example:

(i) $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ ✓ $\rightarrow \frac{\partial \mu}{\partial \beta_j} = x_j,\ j = 1, 2 \quad \frac{\partial \mu}{\partial \beta_0} = 0$

(ii) $\mu = \beta_0 + \beta_1 x_1 + \beta_2 \sin(x_2) + \beta_3 \ln(x_3)$ ✓ $\rightarrow \frac{\partial \mu}{\partial \beta_j}$ doesn't depend on βs.

(iii) $\mu = \beta_0 + \beta_1 e^{\beta_2 x}$ ✗ $\rightarrow \frac{\partial \mu}{\partial \beta_1}$ and $\frac{\partial \mu}{\partial \beta_2}$ depend on βs

As we'll see, there may be a variety of different purposes for using a linear regression model:

(1) Explanation
→ to understand the relationship

(2) Prediction
→ exploit the relationship

However, no matter the goal, our approach to fitting linear regression models to data (i.e., using observed data to estimate unknown regression parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p, \sigma$) stays the same.