Lecture 5

Wednesday, May 29, 2019 2:30 PM

Final Exam: Friday August 9

9:00 - 11:30 am

Location: TBD

Multiple Linear Regression (MLR)

MLR is required when we wish to fit a model relating a response variable y to several explanatory variables $(x_1, x_2, \ldots, x_p)$. That said, MLR is also required if you have one categorical explanatory variable.

As motivation, consider the Bike Share example in which x = "season" where

$x=1 \leftrightarrow$ Spring
$x=2 \leftrightarrow$ Summer
$x=3 \leftrightarrow$ Fall
$x=4 \leftrightarrow$ Winter

Let $x_1 = \begin{cases} 1 & \text{if season = Summer} \\ 0 & \text{otherwise} \end{cases}$

$x_2 = \begin{cases} 1 & \text{if season = Fall} \\ 0 & \text{o.w.} \end{cases}$

$x_3 = \begin{cases} 1 & \text{if season = Winter} \\ 0 & \text{o.w.} \end{cases}$

To model the relationship between y = # of bikes rented and x = season we fit the following MLR model:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\mu = E[y]} + \varepsilon$$

To interpret these coefficients we consider the expected response for different values of x_1, x_2, x_3:

- $E[y \mid x_1 = x_2 = x_3 = 0] = \beta_0$ = expected response in Spring
- $E[y \mid x_1 = 1] = \beta_0 + \beta_1$ = expected response in Summer
- $E[y \mid x_2 = 1] = \beta_0 + \beta_2$ = expected response in Fall
- $E[y \mid x_3 = 1] = \beta_0 + \beta_3$ = expected response in Winter

Thus a given β_j is interpreted as the expected difference in y in category j relative to the baseline category.

Example: β_2 is the difference in the expected number of bike rentals in the Fall relative to the Spring.

*In general, a categorical explanatory variable with m levels is decomposed into $m-1$ indicator variables where one of the levels is treated as a "baseline" against which every other level is compared.

We now turn to the general MLR setup:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

What we care about is

- Estimation
- Prediction
- Model Selection
- Inference
- Model Checking

First we adopt vector/matrix notation to make our lives easier. In this case we have n observations of y: $\{y_1, y_2, \ldots, y_n\}$ and n observations of $x_j = \{x_{1j}, x_{2j}, x_{3j}, \ldots, x_{nj}\}$ for $j = 1, 2, \ldots, p$. A single observation is

$$(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}).$$

The general model that relates these values is:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{\mu_i} + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. As before $\mu_i = E[y_i]$ is deterministic and the ε_i's are random error terms assumed to be iid $N(0, \sigma^2)$. Consequently $y_i \sim N(\mu_i, \sigma^2)$.

Thus we have n equations which can be written in vectors and matrices as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We can rewrite this as $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$n \times 1$ $\quad n \times (p+1)$ $\quad (p+1) \times 1$ $\quad n \times 1$

* Read Sections 3.1 / 3.2 in the textbook for a review of relevant linear algebra.