MSSE Data C200, Spring 2021 — Wednesday, March 3rd

Discussion #8

Name:

Dummy Variables/One-hot Encoding

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 dummy variables that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 dummy variables for this dataset are x_A, x_B, and x_C, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{k,i} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

1. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

2. Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

   Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

3. Show that

$$\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$$

   where i is an element in group A, B, or C.

4. Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned} \hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$