IDS 702: Module 1.10

Bringing the MLR pieces together I (illustration)

Dr. Olanrewaju Michael Akande

Diamonds data

- A diamond's value is often determined using four factors known as the 4Cs: color, clarity, cut (certification) and carat weight.
  - Color: evaluation based on absence of color; how pure the diamond is. This is a categorical variable with 6 levels.
  - Clarity: evaluation based on absence of blemishes. This is a categorical variable with 5 levels.
  - Certification: how well the diamond is cut; how well a diamond's facets interacts with light. This is a categorical variable with 3 levels.
  - Carats: carat weight measuring how much the diamond weighs. This is a continuous variable.
- We will use some data to draw inference about how these factors affect a diamond's price (continuous).
- You can read more about the 4Cs here.

Multiple regression of diamonds data

- A good starting model is

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i; \;\; \epsilon_i \sim N(0, \sigma^2).$$

where y_i is the price for observation i, and $\boldsymbol{x}_i$ is the vector containing the corresponding values for Carats, Color, Clarity, and Certification.

- Alternatively, write

$$\begin{aligned}\text{Price}_i = \beta_0 + \beta_1 \text{Carats}_i + \sum_{j=2}^{6} \beta_{2j} 1[\text{Color}_i = j] + \sum_{j=2}^{5} \beta_{3j} 1[\text{Clarity}_i = j] \\ + \sum_{j=2}^{3} \beta_{4j} 1[\text{Certification}_i = j] + \epsilon_i; \;\; \epsilon_i \sim N(0, \sigma^2).\end{aligned}$$

- Can also write

$$\begin{aligned}\widehat{\text{Price}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Carats}_i + \sum_{j=2}^{6} \hat{\beta}_{2j} 1[\text{Color}_i = j] + \sum_{j=2}^{5} \hat{\beta}_{3j} 1[\text{Clarity}_i = j] \\ + \sum_{j=2}^{3} \hat{\beta}_{4j} 1[\text{Certification}_i = j].\end{aligned}$$

Multiple regression of diamonds data

- This is just a candidate model.
- We will go through the full (almost!) modeling process and we will see if this model makes sense or if we need to make changes to it.
- We will start by doing EDA, all the way down to model assessment, including investigating multicollinearity.
- We will explore transformations, polynomial forms, interactions, etc.
- The data is in the file `diamonds.csv` on Sakai.

Move to the R script here.

What's next?

Move on to the readings for the next module!