Variable Interactions

1/23/2023

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

Quick Recap on MLR

Remember that a Multiple Linear Regression (MLR) model has the following assumption:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

$$car\ price \approx \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \cdots + \beta_p * engineSize$$

A potential issue with the MLR model we need to account for:

- This model assumes **constant association** between each predictor and car price.
- For example, the average increase in *car price* associated with a one-unit increase in *engineSize* is always β_p, regardless of the value of other predictors.
- This assumption may be incorrect.

Introducing Variable Interactions to a MLR Model

Let us add a new assumption to the MLR model:

- The average increase in *car price* associated with a unit increase in *engineSize* depends on the model *year* of the car.
- In other words, there is an interaction between *engineSize* and *year*.
- This interaction can be included as a new predictor, which is the product of *engineSize* and *year*.

Note:

- This assumption is not analytical.
- It is based on the knowledge on the field of application – car prices in this case.
- This is the “data engineering” part of a typical Data Science task.
- You need to know:
  - Your data
  - The field you are working on.

$$car\ price \approx \underbrace{\beta_0 + \beta_1 * mileage + \beta_2 * mpg + \beta_3 * year + \beta_4 * engineSize}_{\text{What we trained on Python last week}} + \underset{\text{A new coefficient}}{\boxed{\beta_5}} * \underbrace{year * engineSize}_{\text{The interaction term}}$$

Interpreting Variable Interactions of a MLR Model

$$car\ price \approx \ \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \beta_3 * year + \beta_4 * engineSize + \beta_5 * \ year * engineSize$$

Note that the MLR model with the interaction term can be re-arranged as:

$$car\ price \approx \ \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \beta_3 * year + (\beta_4 + \beta_5 * year)\ engineSize$$

This is to show that the average increase in *car price* with a unit increase in *engine size* depends on the model *year* of the car.

Interpreting Variable Interactions of a MLR Model

$$car\ price \approx \ \beta_0 + \beta_1 * mileage + \beta_2 * mpg + \beta_3 * year + \beta_4 * engineSize + \beta_5 * \ year * engineSize$$

Note that the MLR model with the interaction term can also be re-arranged as:

$$car\ price \approx \ \beta_0 + \beta_1 * mileage + \beta_2 * mpg + (\beta_3 + \beta_5 * engineSize) * year + \beta_4 * engineSize$$

This is to show that the average increase in *car price* with a unit increase in model *year* depends on the *engine size* of the car.

- If the interaction term is a product, its effect on the model goes both ways.
- The interaction term does not have to be a product – other interactions are possible.
- Both in the slides and in Python, we will go over a product term as an interaction – it is the most common type.
- Some additional ideas, such as **variable transformations** – breaking the linearity assumption – will also be covered in Python.

Reference

Source for slides: https://www.statlearning.com/resources-second-edition