Factors and Interactions

Adam J Sullivan, PhD

1/29/2018

Factors

What are Factors?

- Factors are categorical data.
- Factors contain
  - Levels
  - Can be numerical or character data

Why do we use them?

- Factors allow us to group things by category.
- Factors create dummy variables or indicator variables in our regressions.

What is an indicator variable?

- Consider the scenario where we have 3 treatments: A, B, & C
- We could have two indicator variables:
  - I(Treat_A) is
    - 1 if patient is on treatment A
    - 0 if patient is not on treatment A
  - I(Treat_B) is
    - 1 if patient is on treatment B
    - 0 if patient is not on treatment B
  - Treatment C would be both:
    - I(Treat_A) $= 0$
    - I(Treat_B) $= 0$

What does this mean in regressions?

- Indicator variables change the regression:

$$Outcome = \beta_0 + \beta_1 Age + \beta_2 I(Treat_A) + \beta_3 I(Treat_B)$$

- For a person on Treatment A:

$$Outcome = (\beta_0 + \beta_2) + \beta_1 Age$$

- For a person on Treatment B:

$$Outcome = (\beta_0 + \beta_3) + \beta_1 Age$$

- For a person on Treatment C:

$$Outcome = \beta_0 + \beta_1 Age$$

What does this mean in Regression?

- We can see that a factor leads to multiple different regression lines.
- Each line then has a different intercept than the others.
- In this regression age has the same effect, just the baseline is different.

Are there different types of factors?

- We can have different types of factors
  - Nominal
  - Ordinal

Nominal Factors

- Nominal factors are factors that represent named categories.
- These are categories that do not have an intrinsic ordering.
- Examples:
  - Gender
  - Sex
  - Race/ethnicity
- We must treat these as indicator variables in models.

Ordinal Factors

- Ordinal factors are factors that represent some ordered categories.
- These factors have an intrinsic ordering.
- Examples:
  - Likert Scales (Poor, Neutral, Good)
  - BMI (Underweight, Normal, Overweight, Obese)
  - Age Groups (under 18, 18-25, 25-35, 35+)
- In regression models can be indicator variables or a trend.

Indicator Variables vs Trends

- We saw with indicator variables that we have multiple variables to represent the factor.
- Each category leads to a different regression.
- Consider this:

$$Outcome = \beta_0 + \beta_1 age + \beta_2 I(BMI = underweight) + \beta_3 I(BMI = Over$$

- We then have 3 different regressions:
  - 1 for normal BMI
  - 1 for underweight BMI
  - 1 for overweight+ BMI

Our 3 regressions

- Normal BMI

$$Outcome = \beta_0 + \beta_1 age$$

- Underweight BMI

$$Outcome = (\beta_0 + \beta_2) + \beta_1 age$$

- Overweight+ BMI

$$Outcome = (\beta_0 + \beta_3) + \beta_1 age$$

Indicator Variables vs Trends

- With a trend we allow the factor to have one slope.
- Instead of 1 category leading to a new regression, each category leads to a further increase.
- Our model

$$Outcome = \beta_0 + \beta_1 age + \beta_2 BMI$$

Our Regressions

- Normal BMI

$$Outcome = \beta_0 + \beta_1 age$$

- Underweight BMI

$$Outcome = (\beta_0 + \beta_2) + \beta_1 age$$

- Overweight+ BMI

$$Outcome = (\beta_0 + 2\beta_2) + \beta_1 age$$