IDS 702

Linear Regression - 5 (Interaction terms, multicollinearity)

September 15, 2022

Andrea Lane, PhD

Agenda

1. Pre-class reading questions
2. Interaction terms
3. Multicollinearity
4. In class analysis

Learning Objectives

By the end of this class, you should be able to:

- Interpret an interaction term in a regression model
- Identify and address multicollinearity issues in a regression model

1. Pre-class reading questions

Pre-class reading questions

- Including an interaction term in the regression model allows us to assess a difference in <u>(intercept/slope)</u> for different values of an independent variable
- Multicollinearity <u>(increases/decreases)</u> the certainty of the coefficient estimates, which means the standard error <u>(increases/decreases)</u> and the t-statistic <u>(increases/decreases)</u>, thereby <u>(increasing/decreasing)</u> the statistical power of the model

2. Interaction terms

Interaction terms

- Sometimes we may be interested in how the relationship between a predictor and Y changes based on another (typically categorical) predictor
- Multiply two predictors together: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- Example: We want to know the relationship between a certain drug dosage and anxiety level for those <65 yrs vs. $\geq$65 yrs

3 scenarios

Interaction terms

- If significant, the effect of one predictor on the outcome depends on the value of another predictor
- General practice is to include **main effects** (each variable without interaction, e.g., X_1 and X_2) when including interactions: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- However, interpreting main effects can be difficult when interaction is significant
- Can have higher order interactions ($X_1 X_2 X_3$) or continuous variable interactions but these are difficult to interpret

When should I include an interaction term?

- If the research question/domain calls for it
- If you see a difference during EDA

3. (Multi)collinearity

Multicollinearity: the problem

- You cannot include two variables with a perfect linear association as predictors in regression
- In real data, when predictors are collinear, we see standard errors inflate (which is bad)
- When might we get close:
  - Very high correlations ($|\rho| > 0.9$) among two (or more) predictors
  - When one or more variables are nearly a linear combination of others

Multicollinearity: how to identify

- Think about it during EDA
- Look at a correlation matrix of all predictors (including categorical predictors)
- If you are suspicious of a linear combination, run a regression for the suspected predictors and see if R^2 is near 1
- Look at Variance Inflation Factor (VIF): measures how much the multicollinearity between a variable and other variables inflates the variance of the regression coefficient for that variable

Variance Inflation Factor (VIF)

- $VIF_j = \dfrac{1}{1 - R^2_{X_j|X_{-j}}}$
- VIF will always be $\geq$1 (Why?)
- Generally, VIF =
    - 1 $\Longrightarrow$not correlated (why?)
    - Between 1 and 5 $\Longrightarrow$moderately correlated
    - Greater than 5 $\Longrightarrow$ highly correlated
    - Greater than 10 $\Longrightarrow$ HIGHLY correlated and we want to do something about it

Multicollinearity: what to do?

- Only a problem if you care about the coefficients for the correlated variables
  - Depends on the research question
  - Not so important if prediction is the main goal
- Can remove one of the predictors (which one? Depends on research question, or can look at largest T statistic) or combine
- Can scale your variables (may not always solve the problem)
- Multicollinearity tends to be unimportant in large samples

4. In class analysis

Wrap-up

- Data Analysis Assignment 1 due Fri, Sept 16 11:55 PM
- Statistical Reflection 2 due Fri, Sept 23 11:55 PM