Lesson 20. Using Existing Predictors to Create New Predictors – Part 1

1 Overview

- Suppose we have three quantitative variables, Y, X_1, and X_2
- The multiple linear regression model below allows us to fit linear relationships between Y, X_1, and X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad \varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

  - Visualized in 3D: a flat surface (plane) through a cloud of observations
- But... what if that's not the pattern in the data?
- In Lesson 9, we learned that sometimes transforming variables can help get a better fit
- In this lesson, we will do something similar
- We will learn about using existing predictors to create new forms of predictors that can
  - make the model more flexible, and
  - address non-linear patterns (especially if the linearity conditions are violated)

2 Polynomial terms

- We can include new predictors that take a quantitative predictor variable and raise it to some power
- This can be done for one or more quantitative variables
- For example:

- **Quadratic terms** allow us to curve the surface we are fitting to the data
- For a single quantitative variable X, a **polynomial regression model of degree** k has the form

3 Interactions

- In some situations, the slope with respect to one predictor might change for different values of the second predictor
- This is called an **interaction between the two predictors**
- In Lessons 18 and 19, we saw an interaction between a quantitative variable and an indicator variable
- Now we will consider interactions between two quantitative variables
- The **regression model with interaction for predictors X_1 and X_2**:

- The interaction term allows the slope with respect to one predictor to change for values of the second predictor
    - Visually in 3D: twist the surface we are fitting to the data

4 Complete second-order model

- The **complete second-order model for predictors X_1 and X_2**:

- For two predictors, a complete second-order model includes
    - linear and quadratic terms for both predictors, along with
    - the interaction term
- This extends to more than two predictor variables by including all linear terms, all quadratic terms, and all pairwise interactions

5 Guidance on creating and including new predictors

- When should we should try to include some of these new terms?

    - How do we check for this?
- It is important to avoid **overfitting**
    - i.e., making the model too complicated so that it fits the sample well, but doesn't translate to the population
    - We want a **parsimonious** model: the simplest model that captures the structure in the data

- Two ways to guard against including unnecessary complexity:

- If a higher-order term (interaction, cubic, etc.) is significant, leave the associated lower-order terms in the model (even if they aren't significant)
    - On the other hand, if a higher-order term is not significant, consider dropping it
- If linearity is met, we can make good point predictions, and we also have a reasonable summary of the general relationships among the variables
    - However, unless the other conditions for multiple linear regression (e.g., normality, independence) are met as well, we should not do formal inference (hypothesis testing, intervals)
    - In this case, we will only use p-values as a rough guide